



REVIEW

EDITORS	John C. Caldwell Gigi Santow	Health Transition Centre, ANU, Canberra Health Transition Centre, ANU, Canberra
ASSOCIATE EDITORS	Allan G. Hill Arthur C. Kleinman Stephen J. Kunitz Samuel H. Preston Jon E. Rohde	Harvard University, Boston Harvard University, Cambridge University of Rochester, Rochester University of Pennsylvania, Philadelphia UNICEF, New Delhi
EDITORIAL ADVISORY BOARD	Lawrence Adekun Moses Kweku Aikins Alaka Basu Caroline Bledsoe John H. Bryant Pierre Cantrelle Lincoln C. Chen John Cleland Monica Das Gupta Herman L. Delgado Bob Douglas Aleyya El Bindari Hammad Sally Findley Julio Frenk Godfrey Gunatilleke Carl Kendall Peter Kunstadter Rance P.L. Lee Shirley Lindenbaum David Mechanic Richard Meckel Marvelous Mhloyi Joyce Lewinger Moock W. Henry Mosley Cynthia Myntti Mark Nichter David Nyamwaya Patrick O. Ohadike I.O. Orubuloye Alberto Palloni Olikoye Ransome-Kuti P.H. Reddy Zeba Sathar Santhar Sermsri Richard Smith Andrew Tomkins Budi Utomo Susan Reynolds Whyte Huda Zurayk	Obafemi Awolowo University, Nigeria Medical Research Council, The Gambia Institute of Economic Growth, University Enclave, Delhi Northwestern University, Evanston Aga Khan University, Karachi ORSTOM, Paris Harvard University School of Public Health, Boston London School of Hygiene and Tropical Medicine, London National Council of Applied Economic Research, New Delhi Institute of Nutrition of Central America & Panama, Guatemala The Australian National University, Canberra WHO, Geneva Columbia University, New York Instituto Nacional de Salud Pública, Mexico Marga Institute, Sri Lanka FHI Washington Div., Va. University of California, San Francisco Chinese University of Hong Kong The City University of New York Rutgers University, New Brunswick Brown University, Providence University of Zimbabwe, Harare The Rockefeller Foundation, New York The Johns Hopkins University, Baltimore The Ford Foundation, Jakarta The University of Arizona, Tucson African Medical and Research Foundation, Nairobi United Nations Regional Institute for Population Studies, Accra Ondo State University, Nigeria University of Wisconsin, Madison Federal Ministry of Health, Lagos Population Centre, Bangalore Pakistan Institute of Development Economics, Islamabad Mahidol University, Thailand All Souls College, Oxford Institute of Child, University of London, London University of Indonesia, Jakarta University of Copenhagen, Copenhagen The Population Council, Cairo
PRODUCTION	Jennifer Braid Wendy Cosford	Health Transition Centre, ANU, Canberra Health Transition Centre, ANU, Canberra

ISSN 1036-4005

Published by Health Transition Centre, National Centre for Epidemiology and Population Health, The Australian National University, GPO Box 4, Canberra ACT 2601, Australia.

Email address: mgs868@nceph.anu.edu.au or jxb868@nceph.anu.edu.au

Annual subscription: *Health Transition Review*, Health Transition Centre, Fax 61-6-249-0740

Publication frequency: April and October.

Printing by the Printery, The Australian National University

Cover design by Graphic Design Unit, The Australian National University

Copyright © 1992 Health Transition Centre. All rights reserved. The copyright owner's consent does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Written permission must be obtained from the publisher.

Preface



The assumptions governing mortality studies in both contemporary and historical demography have recently undergone considerable change. In contemporary demography work was long based on the assumption of relatively mechanical relationships between modernization and mortality decline, and the main disputes concerned the proportionate contribution of different components of modernization etc.: particularly those of economic modernization, working through the agency of rising living standards, *versus* technological/scientific modernization through the agency of curative or preventative medicine. In both cases it was assumed that infectious disease patterns could be considered as essentially passive, dependent, variables, and the main questions to be resolved were the relative importance of known relationships and how these worked themselves out in given instances.

Recently, however, a more complex picture has emerged. The extent of slow-downs in Third-World mortality decline generally has been much debated but in certain regions, particularly in Africa, political and military disruption has damaged basic health infrastructures and threatened to reverse the trend. At the same time, infectious disease patterns themselves have proved to be less passive than previously thought. The appearance of drug resistant strains of malaria provided an early, ominous, instance of this, but a more dramatic development has been the rise of 'new diseases', especially AIDS. Such developments, in turn, raise more basic issues about the nature of modernization itself and its relationship to mortality decline. Urbanization, improved transport, and the growth of long-range economic linkages have evidently provided new opportunities for infectious micro-organisms to thrive, whilst recent events have posed questions concerning ethnicity, political modernization and the post-colonial state in both the northern and southern hemispheres.

At the same time the conceptual underpinnings of historical mortality studies have undergone substantial revision. Here too, the classical approach saw mortality as a dependent variable responding to changes in living standards; a view both reinforced and narrowed in its focus by McKeown's nutritional determinism. Despite the latter's wide influence, however, empirical evidence for a strong nutritional role in secular mortality change before the late nineteenth century has been scarce whilst a number of studies have yielded results hard to reconcile with the hypothesis (Kunitz and Engerman this volume). Such findings fostered the view that, historically, secular mortality had responded mainly to climatic or biological factors lying outside the realm of economy and society and led in turn to a relative marginalization of mortality in historical studies since, throughout the 1980s, historical demographers sought to build closer links with their colleagues in cognate areas of social and economic history.

The conceptualization of mortality change as 'autonomous' created considerable heuristic problems, past climatic and, above all, micro-biological changes being recalcitrant to historical observation, but it also imposed a radical discontinuity on mortality studies generally as between past and contemporary populations. Among the latter, mortality was clearly related to the state of economy and society, as well as to medical developments, whereas in the former it appeared to be an entirely independent variable. There thus seemed little common ground between the two domains which were apparently unable to address each other's concerns.

Recent work with historical populations has attempted to overcome these problems by shifting attention from mortality, considered as a single output variable, to the dynamics of infectious disease

processes as such. Central to these studies in 'historical epidemiology' is an analysis of the way in which spatial structure: embracing such phenomena as population density, migration flows and economic linkages as well as the characteristics of the environment itself, affect levels of exposure to infectious agents. The British Society for Population Studies' 1991 Conference, *Historical Epidemiology and the Health Transition*, assembled demographers and related professionals working with both historical and contemporary populations in order to develop a common framework for understanding the dynamics of infectious disease mortality in both these contexts.

The resulting papers, a selection of which form the present volume, demonstrate the importance of exposure to infection as a determinant of mortality variations in both time and space. Historically, it seems clear that the popularity of the 'autonomous mortality' concept reflects an excessive concentration on factors affecting resistance to infection and on the wage/price ratio as a proxy for the world of human affairs in its totality. It is evident that levels of exposure to infection responded to a much wider range of factors, not least of which, and implicated repeatedly in the papers read to the conference, was the large-scale organization of space by state structures and global economic systems: neither in historical Europe nor the contemporary Third World can the dynamics of exposure to infection be understood without reference to the progressive incorporation of populations within systems of this kind.

A second, and less expected, theme to recur was the relationship between power and what can be termed 'situational knowledge' of a given milieu, its disease hazards and the mechanisms of disease transmission operating within it. Such knowledge, obtainable by means including both empirical experience and the application of abstract conceptual frameworks, such as germ theory, to given instances, enables exposure to infection to be reduced by avoiding hazardous environments and activities. Often, however, individuals were unable to act on such knowledge, being constrained to move to less healthy environments, against what Dobson (this volume) aptly terms the 'contours of death', as a result of economic powerlessness, as with migrants to the cities of early-modern Europe, or, as in the mines of tropical Africa (Fetter, this volume) direct politico-military coercion.

Where contemporary health problems, such as malaria and HIV are concerned, the issue emerges again in a different form: the need to transform health professionals' general conceptual knowledge of disease transmission into lay knowledge of the consequences of specific actions in specific situations. At one level this raises the kind of technical questions considered by Bradley (this volume) in his discussion of malaria: how are disease risks encountered in the course of specific daily activities in specific environments and how can they be reduced? But at another it raises issues of cultural translation akin to those involved between the statements 'smoking increases the risk of lung cancer' and 'if I smoke I am likely to die of cancer'.

The issue of power relations recurs at both these levels. Unlike the older 'top down' style of technology-intensive program, the formation of situational knowledge requires the active participation and collaboration of the populations concerned, and thus that they be capable of making their voices heard by health professionals. Beyond this, however, success is unlikely where professionals of any and all kinds are identified with a hostile official culture and an unresponsive, or repressive, state apparatus. As in Europe, so in the Third World, the struggle for public health raises questions going to the heart of the relationship between state and society.

The papers contained in this Supplement are based on presentations originally made to the annual conference of the British Society for Population Studies, *Historical Epidemiology and the Health Transition*, which was held at the University of Southampton on September 11–13, 1991.

The Supplement was produced in Canberra at the Health Transition Centre, The Australian National University by Jennifer Braid and Wendy Cosford.

John Landers
All Souls College
Oxford

Introduction



John Landers

All Souls College, Oxford OX1 4AL, UK

The experience of the 1980s, both in historical research and in the current demography of developed and developing countries, has revealed inadequacies in the approach traditionally adopted by demographers to the problem of secular change in infectious-disease patterns and their associated mortality levels. Classical population theory assumed that pre-transitional mortality change was a function of changing living standards, with the mortality transition itself arising from some combination of economic development and techno-medical changes, and heralding an epoch in which infectious disease would be of marginal demographic significance.

Progress towards this goal, it was assumed, would be closely related to that in the sphere of economic development, with the latter bringing mortality benefits both directly, through the effects of improved living standards, and through the provision of resources for the implementation of health technologies and programs. Research has shown, however, that pre-transitional mortality change was poorly correlated with real-wage levels in Europe, whilst several poor countries have attained lower mortality by the late twentieth century than have others experiencing significantly greater economic development, and the processes of social and economic modernization have created new opportunities for the spread of killing infections.

The failure of simple 'standard of living' determinism in historical-mortality studies has given rise to a new approach, termed 'historical epidemiology', which focuses attention on the immediate determinants of differences in exposure and resistance to infection. The conference 'Historical Epidemiology and the Health Transition', held by the British Society for Population Studies at Southampton in September 1991, brought together demographers and epidemiologists working on both historical and present-day populations so as to explore these developments and move towards a new framework for the analysis of secular-mortality change; a framework combining the insights of each of the relevant subdisciplines. Central to this was a perceived need to move beyond restrictive economic determinisms, and narrowly conceived unilinear models of change, in order to recognize both the historically-specific character of individual episodes of mortality change and their underlying structural regularities.

The past and the present: history and transitions

The relationship between contemporary and historical demography is complicated by the latter's curious dual affiliation; at once a form of historiography, the broader manifold of historical enquiry, and as a subdiscipline of demography, a component of the social sciences. It exists in an uneasy state of intellectual cohabitation. Historians are fascinated by variation, the specificity of time and place, and the need to encounter the past on its own terms. Above all they are apt to become uneasy when asked to draw the 'lessons of history'.

The social sciences, by contrast, have pursued a love affair with generalization, with typology, and the incorporation of diverse times and places within common frames of reference.

These divergent orientations have led to conflict and misunderstanding. Historians accuse social scientists of a distorting 'present mindedness' and a cavalier neglect of historical detail, whilst to social scientists, historians can resemble anti-conceptual Philistines whose perverse obsession with trees blinds them to the existence of forests. But this is not a necessary state of affairs, for there can equally well be a creative tension, even a synergy, between the two. Testing the social scientist's general model requires access to the empirical variations uncovered by the historian, whilst the specificity of an historical instance can be determined only by comparison which requires a generalizing frame of reference capable of placing diverse instances in a common set of relationships.

If this is to be achieved in the present context then we must have a general conceptual model of mortality change, and to avoid the pitfalls it is essential that the model's contours be clearly specified. This is especially true in three respects:

Closure

We must first distinguish between 'closed' and 'open' models. In closed models the possible trajectories of the system are specified in advance by the structure of the model itself which thus incorporates a predetermined set of pathways along which development can occur. Open models, by contrast, make no such *a priori* specification, and the range of possible trajectories is thus unlimited.¹

Continuity

Long-term change in one or more output variables can be represented by either 'continuity' or 'discontinuity' models, the distinction lying in the structure of the imputed causality. In continuity models this remains constant regardless of the system's overall trajectory and any changes in the output variables, whereas in discontinuity models the causal structure itself may vary as the system evolves. The distinction is most often encountered in the case of closed models of transition between distinct stages where, in Tilly's (1978) words, discontinuity models require us

to formulate three subtheories: one concerning the dynamics of the 'pre-industrial' or 'traditional' demographic system, another concerning the dynamics of the 'modern' or 'industrial' system, and a third concerning the transition between them (Tilly 1978:50).

In principle, however, the distinction applies to all models of change over time.

Levels of causality

A well-founded model must specify not only the output variable's ultimate determinants but also the causal pathways through which these exercise their effects. In practice this involves the construction of sets of proximate determinants, or intermediate variables, which respond to changes in the ultimate determinants and act directly on the output variable.

¹ Demographers have considered this problem generally in the form of a question such as 'health transition or transitions?' (Frenk et al. 1991). The conceptualization advanced here is based on Foucault's distinction between closed and open systems of classification (Foucault 1970; Bowler 1989).

The Notesteinian tradition

It is a truism that present-day demographic conditions, for a substantial majority of the world's population, differ greatly from those of the lengthy period from the first agricultural revolution to the very recent past, and that populations in different countries have travelled different distances from their previous condition. It is likewise at least arguable that, demographically, those countries that have travelled the shortest distances have more in common with the condition of the presently developed countries two centuries ago than they do with their more advanced contemporaries.

These observations underlie an approach dominating studies of long-term demographic change since Notestein first formulated the theory of demographic transition (Notestein 1945). This 'Notesteinian tradition' in population studies envisages a single evolutionary sequence with socioeconomic and cultural changes pushing societies along a 'tramline' of development in processes which might occur at varying speeds, but whose internal structure and external implications are recognizably the same throughout their various manifestations in time and space.

For Notestein it was the agricultural, industrial, and sanitary revolutions which boosted living standards, ameliorated the consequences of urbanization, and endowed human populations with a hitherto unprecedented degree of control over their collective and personal environments: not least through medical progress with commensurate reductions in mortality, a process assumed to have begun first in England and then spread to Europe and the world at large. Such a model allowed a negligible role at best for prior cultural differences between populations; rather, cultures were assumed to be either 'traditional' or 'modern': the former governed by a religious passivity in the face of nature, and bereft of effective knowledge as to its manipulation; the latter characterized by an individualistic instrumental rationality and equipped, through the scientific method, with an ever-expanding knowledge base and all the technologies that this allowed.

Notesteinian models are thus both closed and discontinuous. Closed because there is only one possible trajectory of development: movement occurs at different speeds, or not at all, but is restricted to one single track leading to the terminus of demographic modernity. Discontinuous because the determinants of traditional and modern demographic regimes are fundamentally distinct: the former being dominated by the vagaries of nature, in the case of mortality, and the constraints of kinship systems, religion, and collective moral codes in that of fertility; the latter by a scientifically-informed technology, an industrial economy, and a calculus of individual rationality.

The basic observations underpinning this approach to demographic change may be truisms, but they are not thereby necessarily true, or rather their truth may turn out to have a slippery and unhelpful character. Unilineal 'stages' models of social and economic development achieved a wide currency in the social sciences during the immediate post-war decades, but in the longer term they have proved less successful; not least because they often attribute to the industrialization process some characteristics of Western industrial societies which were actually rooted in their pre-existing culture.²

² The most important of these was of course Rostow's (1960) model of economic growth whose indirect influence may have served to promote the popularity of the Notesteinian approach, although the broader background to the formulation of post-war demographic theory has been relatively little studied (see Hodgson 1988). The confusion between features stemming from industrialization and those based on pre-existing culture is most apparent in the case of family structure (Laslett 1977). For some

The epidemiological transition

The Notesteinian approach to mortality analysis is most fully elaborated in Omran's model of 'epidemiological transition' (Omran 1971, 1983) recognizing as it does that mortality transition involves more than simple quantitative reductions in mortality levels and their short-term variability.³ Like Notestein, Omran postulated a developmental sequence of three stages (or 'ages'), but distinguished these in terms of mortality's age and cause-of-death structure rather than by their overall demographic regimes. Empirically this schema, with its ages of 'pestilence and famine', 'receding pandemics', and 'chronic and degenerative diseases', relied on the imputed historical experience of Western populations, but variations in the context of transition from late eighteenth-century England to the contemporary Third World required the recognition of variant forms.

For Omran there were three dimensions on which such variations could occur: the pace of mortality reduction, the causal weight of endogenous socioeconomic change relative to exogenous technologies, and the resulting degree of demographic disequilibrium. Within this framework he distinguished three scenarios of change ('models' in his terms) of which the Western or 'classical' scenario served as the point of reference. Alongside this was a special case of accelerated transition, largely representing Japanese experience, and a third 'contemporary' scenario which effectively functioned as a residual.

The classical epidemiological transition

This was the 'base case' intended to reflect the demographic history of Western Europe from the mid-eighteenth century, where the motive force had been endogenous socioeconomic development subsequently reinforced by the urban-sanitary revolution and scientific medicine. The pace of change, being internally generated and relatively slow, meant that disequilibrating consequences of rapid population growth were minimized.

The accelerated model

The accelerated model referred to instances of unusually rapid transition of which the type case was Japan. Here mortality decline owed more to technological factors than it had in the West but occurred in the context of a prior 'slow process of modernization'. The relevant cultural milieu had been such as to form 'national and individual aspirations favouring a controlled rate of population growth' and provide motivations for fertility control.

The contemporary model

In the contemporary case the driving force had been exogenous public health and medical technology which had been imported without appreciable socioeconomic development. This 'exogeneity' operated on different levels; the 'hardware' had been physically developed outside the countries concerned, but also manifested an institutional

recent views on relationships between modernization, the origins of capitalism and prior cultural differences see Baechler, Hall and Mann (1988).

³ For the purposes of this discussion we shall treat the mortality, epidemiological, and health transitions as a set of hierarchically-nested concepts. The mortality transition refers to the secular decline in mortality and its associated changes in mortality patterns of age and cause of death. The epidemiological transition includes the former but embraces a broader set of changes in morbidity and the disease environment. The term 'health transition' is used in Caldwell's sense and 'includes the social and behavioural changes that parallel the epidemiological transition and may do much to propel it' (Caldwell 1990:xi).

and cultural complex alien to their society and culture. Hence the transition was initiated without the socioeconomic and cultural changes which had blunted its disruptive impact in the classical scenario; 'death control' far outstripped birth control, and the stage was set for explosive population growth.

Epidemiological transition and historical evidence

Omran's model has been very influential but has shortcomings in terms of both empirical adequacy and explanatory sufficiency. We must look at both of these issues in turn.

Empirical adequacy

The first question to be asked of any closed model incorporating historical data is how far the trajectories it specifies correspond to the historical experience of actual populations. Here, this arises chiefly in respect of Omran's 'base case' which represented both the point of reference endowing the model with its analytical coherence and the most fully documented instance of long-term demographic change currently available. Three main points emerge:

The primacy of mortality

Those Western European countries for which we currently have the best data turn out to have had remarkably different experiences in respect of Omran's Proposition I; the primacy of mortality in determining population growth (Wrigley and Schofield 1981; Wrigley 1983). Sweden in the century after 1750 shows the 'expected' pattern of falling mortality and fluctuating birth rates. In England, however, mortality decline was of secondary importance and a rise in the birth rate due to changing nuptiality was responsible for most of the country's natural increase. France provides yet a third variant; mortality fell sharply in the late eighteenth century, but a countervailing drop in both nuptiality and marital fertility kept population growth within narrow bounds.

The stabilization of mortality

A second element of Omran's model to fall victim to empirical research is the assumption that death rates declined initially because of the attenuation of mortality crises, the violent short-term upswings due to war, famine or large-scale epidemics. Such a 'stabilization of mortality' was widely accepted in the demographic literature of the 1950s and 1960s which assumed that crisis mortality and the 'background' levels of so-called 'normal' years were clearly distinct, the latter varying little in time or space (Flinn 1974, 1981). Hence differences in cohort life expectation were principally determined by variations in the incidence and severity of crises.

On this view the earlier and later stages of mortality decline were qualitatively distinct. At first the stabilization of mortality, due to administrative measures such as quarantines, improved military organization, and better distribution of food in times of scarcity, produced a secular amelioration whilst leaving background levels little changed. Substantial improvements in the latter had to await scientific medicine and the kind of growth in incomes which only an industrial economy could effect.

This picture appeared plausible given Europe's history of wars and subsistence crises, and above all its experience of bubonic plague from the later Middle Ages to the end of the seventeenth century. In fact urban plague outbreaks provided model examples of mortality crises, being both extremely destructive and sharply focused in space and time, and the eventual disappearance of the disease from Western Europe has been attributed

to successful administrative action (Flinn 1979; Appleby 1980). Similarly, if less dramatically, warfare impinged less on civilian life in the course of the eighteenth century than earlier (Gutmann 1980), whilst the scourge of subsistence crises was clearly blunted and had effectively disappeared on a continental scale by the 1820s (Post 1977; 1985).

More detailed evidence, however, has cast doubt on the stabilization thesis (Schofield 1984; Perrenoud 1991). The English data, for instance, have shown no simple relationship between the secular level of mortality and its short-term instability. The latter declined appreciably with the disappearance of plague, but life expectation also fell as mortality stabilized at a new and higher level reflecting endemic, and smaller-scale epidemic, causes of death (Wrigley and Schofield 1981; Dobson 1989a, b). Background mortality levels could thus vary significantly over time, and evidence from elsewhere in Europe has shown mortality crises persisting to the very eve of the secular decline, if not beyond.

Finland's worst crisis occurred in the 1860s, after more than a century of slow and irregular improvement in mortality, and was followed immediately by the onset of sustained decline (Vallin 1991). Similarly, in Russia a substantial decline in 'normal' mortality occurred in the mid-1920s, hard on the heels of the worst national crisis Europe had seen since the Middle Ages and after some two decades of moderate improvement (Lorimer 1946). At the same time, from a theoretical perspective, Watkins and Menken (1985) have demonstrated that there is only a very limited scope for famines, or by implication any other form of crisis mortality, to act as long-term regulators of population growth.

The chronology of decline

A further difficulty concerns the use of Western Europe's mortality decline as a 'base case' in itself, or rather the interpretation of the region's experience after 1750 as a unitary transition. Such an interpretation implies that declines were both continuous over time and qualitatively different from earlier large-scale movements, being somehow bound-up with an unprecedented and irreversible process of modernization. The latter point bears on the explanatory issues which we shall consider below, but a closer examination of mortality trends themselves is enough to cast doubt on any unitary interpretation.

Whilst it is true that the Nordic countries did display a single continuing fall in mortality, this was not the case elsewhere in the region. In England an initial phase of decline seems to have run its course by the 1830s, followed by a lapse of a generation before the onset of further sustained improvement. We should bear in mind that national life expectation in 1870 differed very little from that of the late sixteenth century. Similarly, in France the rapid reduction of mortality from 1750 had ground to a halt by the 1840s and was not resumed until the last twenty years of the century. Data from the Netherlands likewise show no sustained improvement in mortality over the central decades of the nineteenth century, and this seems to have been true in Germany as well (Anderson 1988; Vallin 1991).

Summary

These findings reveal that mortality's role in determining population growth was more variable, and endemic mortality more important in determining the overall level, than the transition model implies. The dating of the transition itself is much less certain than the Notesteinian tradition would have us believe. Indeed, the latter probably places the transition a century too early, and the declines of the late eighteenth century are better seen in the context of the cyclical movements which had characterized European mortality

in earlier centuries; an issue which moves us from the level of empirical adequacy to that of explanatory sufficiency.

Explanatory sufficiency

In an earlier section I outlined some general issues relating to models of mortality change and stressed the importance of distinguishing levels of causality. Recent demographic literature has displayed a growing interest in the proximate determinants of population variables,⁴ but Omran's model pays little attention to these in their own right, preferring to cluster them around sets of ultimate determinants, amongst which socioeconomic development and medical progress took pride of place. He distinguished three scenarios within this framework, and we shall now consider how far the causality postulated in each case has been supported by subsequent research.

The classical model

The conceptualization of Western Europe's mortality decline as a 'classical epidemiological transition' is, as we have seen, open to empirical doubt, but how far can its explanation in terms of socioeconomic modernization be sustained? The original Notesteinian transition model incorporated an explanation first formulated in the inter-war period and pointing to general improvements in living conditions and, in particular, to medical advances as the factors responsible for reductions in mortality as early as the eighteenth century.⁵

The medical explanation was comprehensively assaulted by McKeown and Brown (1955) and has since found only a very limited defence.⁶ McKeown in turn substituted a nutritional explanation, postulating increases in agricultural output. He relied on an argument by elimination, his celebrated 'Sherlock Holmes' procedure, but applied this in different ways to the eighteenth and nineteenth centuries, reflecting differences in the data. In the eighteenth century he rejected parish registers as valid evidence and fell back on indirect inference, arguing that fertility increase was *a priori* unacceptable as an explanation of population growth, that no social reforms capable of reducing exposure to infection had occurred and that, having eliminated medical advances, only nutritional remained (McKeown and Record 1962; McKeown 1976).

As we have seen, however, historical demographers do in fact regard the birth rate as the prime cause of England's eighteenth century population growth, and McKeown's explanation of the contemporary mortality decline has likewise failed to find favour. The root of the problem is the lack of any convincing evidence for a substantial improvement in popular nutrition; indeed economic historians have seriously considered the possibility of a deterioration at this time (Floud, Wachter and Gregory 1990). Furthermore mortality improved dramatically among the aristocracy (Kunitz and Engerman, this volume), and it is hard to believe that this group had previously been underfed. Elsewhere in eighteenth-

⁴ A rigorous proximate-determinants framework was first developed in the context of fertility analysis (for example, Bongaarts 1978; Hill 1990), but a number of attempts have been made to apply this approach to mortality (for example, Mosley and Chen 1984). The model developed below differs from these in being aimed at inter-population, rather than inter-individual or household, differences.

⁵ See Landers (1993), Chapter 1, for a review of conflicting interpretations of English mortality declines.

⁶ These have chiefly concerned the effects of inoculation, and subsequently vaccination, on smallpox mortality (see Razzell 1974; Mercer 1985).

century Europe the case for a nutritional explanation seems even weaker. Many regions remained unaffected by economic modernization, but mortality decline appears to have been widespread (Lee 1979), and historical demographers generally view this as an 'autonomous' phenomenon arising from climatic or microbiological changes (Perrenoud 1984; Landers 1987).

For the period after 1837, when official government statistics become available, McKeown relied on cause-specific death rates grouped by mode of transmission. Sanitary reform (or 'municipal hygiene' as he termed it) would, he argued, only affect water- and food-borne causes of death so that movements in these would provide an indicator of such reforms' contribution to the reduction of general mortality. Nutritional factors, by contrast, would be the major determinant of mortality in the airborne group, and so his conclusion that these, particularly respiratory tuberculosis, had dominated England's mortality decline led him to reject government action in favour of nutritional improvement as the primary explanation.⁷

The independent evidence for such improvements, brought on by steam transport and the expansion of European-style agriculture overseas, is incomparably stronger for this period than for the eighteenth century. McKeown's argument thus commanded widespread acceptance on this point, exercising a general influence on mortality studies which it would be difficult to exaggerate. Here too, however, dissenting voices have recently been raised, and Szreter in particular has shown that McKeown's own data do not support his conclusions to anything like the extent which he himself claimed (Szreter 1988).

Thus, an analysis of different contributions to overall mortality decline before 1900 shows that water- and food-borne causes actually account for over a third of the total. Among the airborne group, the decline in neither smallpox nor scarlet fever (amounting jointly to a further 18 per cent) can reasonably be attributed to nutritional factors, and McKeown also failed to confront the problem, which he acknowledged, of distinguishing respiratory tuberculosis from bronchitis for much of the period. Since deaths from the latter rose as the former declined, there is more than a suspicion that some of the fall in recorded tuberculosis mortality was due to better diagnosis. If the two causes are grouped together, they account for only a fifth of the pre-1900 reduction in overall death rates, and Szreter argues further that such decline as occurred in tuberculosis mortality was partly an effect, rather than, as McKeown had argued, a cause of reductions elsewhere. There are other objections to a purely nutritional explanation of tuberculosis mortality; geographical variations especially are easier to explain by differences in exposure to infection, reflecting workplace and housing conditions. This can be seen dramatically in the case of Glasgow, where severe tuberculosis mortality persisted to the middle of the twentieth century and can plausibly be explained by the city's large number of one-roomed dwellings (McFarlane 1989).

This 'classical' transition cannot then be explained simply in terms of dietary improvement. Other dimensions of living standards such as housing, which featured non-market allocation, also made significant contributions, and direct political intervention

⁷ In this discussion we have followed McKeown's usage of the term 'nutrition' to refer to food intake. In fact, 'diet' would be a more satisfactory term since nutritional status is jointly determined by food intake, the organism's ability to extract nutrients from the latter, which may be impaired by previous morbidity, and the energy demands on the organism from whatever source.

evidently played a major role, perhaps the major role, in furthering mortality decline. Hence we cannot, as McKeown's argument implies, attribute the process to socioeconomic development alone; Omran's more balanced formulation is preferable, although it is possible that he underestimates the contribution made by the sanitary revolution.

The accelerated transition

Re-dating the Western mortality transition makes it difficult to treat the Japanese experience as an 'accelerated' version, since England's sustained mortality decline now appears to have predated the Japanese by a generation at most, rather than by a century and a half. Indeed Johannson and Mosk's (1987) recent analysis of Japanese experience in the first half of the twentieth century has shown how this can most fruitfully be seen against the background of developments in contemporary Europe, rather than those of the late-eighteenth and early-nineteenth centuries.

Omran paid little attention to the determinants of mortality decline in the accelerated scenario, referring simply to 'sanitary and medical advances as well as ... general social improvements' (Omran 1971:535), but these take centre stage in Johannson and Mosk's (1987) analysis, which gives their work a much broader significance. The authors begin by establishing a general two-dimensional model of mortality levels, in which these depend on the relative weight of exposure and resistance to infectious agents. These variables are themselves determined by a variety of socioeconomic, cultural, biological, and political factors.

England, Italy, and Japan around 1900 had closely comparable levels of mortality, with life expectation in the mid-forties, stemming from quite different configurations of exposure and resistance variables. England, then the world's second richest country, displayed high levels of nutritionally-based resistance offset by high exposure consequent on urbanization and a poorly developed system of public health.⁸ Japan, at the opposite extreme, was then a low-income country with a poorly nourished population, but enjoying 'a particular combination of natural and investment-related advantages which gave its comparatively poor population an unusually high level of *protection from exposure* [authors' italics]' (Johannson and Mosk 1987:219).

Government action contributed significantly to the latter through effective public-health interventions, but in later decades, as Japan became both wealthier and 'modernized', exposure and resistance each increased, effectively blocking further movements in mortality. Increased exposure reflected both the penalties of urbanization and a diminution in public-health investment as a growing share of national resources was allocated to imperial expansion. The latter's catastrophic denouement at the end of the Pacific war saw the devastation of Japan's urban infrastructure and a temporary collapse in life expectation to less than 25 years. Further government investment in health, the availability of new health technologies, and renewed urbanization subsequently combined to effect a dramatic improvement in both exposure and resistance variables and a corresponding decline in mortality.

⁸ This judgement of course may now require reassessment in the light of Szreter's (1988) subsequent analysis.

The contemporary model

Omran saw contemporary health transitions as driven by the external provision of 'death control' technologies to Third World countries. This view, current in the preceding decades, was particularly influenced by the spectacular achievements thought to have been wrought by anti-malaria programs in countries such as Sri Lanka. Such mortality declines were entirely 'exogenous' in origin and bore no relationship to the socioeconomic level of the countries concerned which had acted simply as passive recipients of technologies developed in the course of the Western epidemiological transition. This explanation was challenged by another based on McKeown's arguments, crediting improved living standards with the major contribution to mortality decline (Kunitz 1987). Subsequent research, however, has undermined any straightforward dichotomy between exogenous and endogenous explanations.

First, Preston, in a regression analysis remarkable for its elegance and economy, showed that in neither developed countries nor the Third World did mid-twentieth century mortality declines derive primarily from increasing *per capita* national income. Around half of the latter's life-expectancy gains between the quinquennia 1935-1939 and 1965-1969 arose from structural changes in the relationship between national income and mortality, whilst these accounted for 80 per cent of such gains worldwide (Preston 1976, 1980).

Preston's analysis thus lent only equivocal support to the 'exogenous' interpretation of Third World mortality declines, for the difference between these two percentages implied that developed countries actually gained more from non-economic factors than did the Third World. He further concluded that, whilst progress had been especially marked in previously malarial regions, malaria probably acted as a 'proxy for a host of potentially eliminable conditions' rather than providing the key to mortality decline. Direct contributions from developed countries to Third World health expenditures had been very modest, amounting to less than three per cent of the total, and the former's principal contribution had been 'the development of low-cost health measures exploitable on a massive scale, demonstration of their effectiveness in relatively small areas [and] training and provision of personnel' (Preston 1980).

Such a conclusion leads on to the question of how far host-country conditions influenced the successful implementation of such programs and the effective use of personnel. This question was pursued by Caldwell in a paper whose methodology effectively reversed that of Preston's (Caldwell 1986). Instead of measuring the average contribution of income growth to mortality change, Caldwell examined the residuals, 'deviant' instances of populations which had much higher or lower mortality than that predicted by their economic status. Focusing particularly on Sri Lanka, Costa Rica, and Kerala, he demonstrated the critical importance for health breakthroughs of government interventions in health and nutrition, female autonomy, and the presence of an 'educated, capable and demanding public'.

The presence of such a public reflected investment in schooling, and the existence of a participatory, plural, political system with a tradition of egalitarian radicalism reinforced by pre-existing cultural attitudes to health and hygiene. Thus, whilst the contribution of economic development narrowly conceived as income growth may have been equivocal, it is evident that sociopolitical and cultural conditions in Third World countries play an important part in promoting or retarding mortality declines; such countries cannot be seen merely as passive recipients of exogenous technologies.

The lessons of history

Omran, working in the Notesteinian tradition, put forward a closed, discontinuity model depicting a transition between two demographic regimes whose mortality patterns were determined in radically distinct ways. The proximate determinants of mortality were not considered in any detail in the original epidemiological transition model, but they were central to the work of McKeown, whose arguments centred on the relative importance of exposure and resistance to infection and on their determining factors. He opened up new possibilities analytically, even while shutting them down empirically through his insistence on virtually monocausal explanations of mortality change.

McKeown's was thus a continuity model in a double sense: first, because the framework of exposure and resistance⁹ could be applied generally across demographic regimes, and secondly, in a much narrower sense, because the same variable, nutrition, dominated mortality's determinants in successive demographic regimes. Such nutritional determinism cannot, as we have seen, be supported by evidence, but the framework of exposure and resistance variables has been developed fruitfully by Johansson and Mosk (1987) in an open model which allows a potentially unlimited set of possible trajectories.

Empirically, the authors delineate three distinct pathways to low mortality, pursued by England, Italy and Japan. These differ both in their proximate causality, the relative importance of changes in exposure and resistance, and in their ultimate determinants: economic growth, in the English case, contrasting with the more 'politically driven' experience of Japan and Italy which led to a correspondingly uneven progress. Moreover this analysis implicitly makes difficult the relationship between mortality decline and modernization, as both Fascist Italy and pre-war Japan were dominated by self-consciously modernizing, and in the latter case Westernizing, ideologies. This was defined in terms of a strong state bolstered by militarism and external aggrandizement; policies which resulted, not in mortality decline, but in stagnation and eventual catastrophic increase.

Caldwell's (1986) analysis also points to the existence of distinct pathways to low mortality. Whereas the First World's mortality decline was at least 'economically facilitated' by rising real incomes, the experience of Costa Rica, Kerala, and Sri Lanka suggests a 'socially willed' process with governments of varying persuasions responding to popular demands for better health. Alongside this stands the alternative 'politically imposed' path taken by Communist regimes where the Party uses its monopoly of political power to direct investment towards health provision. Caldwell's primary concern is with the ultimate, societal, factors promoting or impeding the implementation of health measures based on Western scientific medicine. But his variant pathways imply differing configurations of proximate determinants of mortality, and he concedes an important role to indigenous, or 'pre-scientific' conceptions of hygiene and health-promoting behaviour in countries.

Such conceptions have also been emphasized in the Japanese case, whilst the relationship between scientific medicine and the development of effective health measures in Western Europe also proves to have been more complex than might first appear. Here,

⁹ Here, as throughout this paper, we are concerned with mortality from infectious causes and thus with exposure to infectious micro-organisms. The approach adopted can, however, easily be generalized to mortality as a whole by broadening the concept of exposure to include exposure to environmental hazards and associated risk factors.

military and naval authorities began to develop effective preventive techniques a century before the formulation of germ theory, and micro-biologically efficacious forms of water filtration also preceded the latter by some decades (Mathias 1979; Curtin 1989). Such procedures were based on culturally informed 'commonsense' judgements as to cleanliness and health, and on essentially pre-scientific theories of disease causation. An understanding of such factors, and their relationship to large-scale variations in the proximate determinants of mortality, requires a long-term perspective of the kind provided by historical epidemiology.

Historical epidemiology

The epidemiological transition model in its original form thus has serious shortcomings. The pre-transitional demographic regime displayed substantial variations in mortality in both time and space and seems to have endured for very much longer than had been thought. Nor does it appear realistic to think in terms of the disappearance of infectious disease mortality as necessarily bound-up with a unitary process of 'modernization'. Such mortality in the past evidently possessed its own dynamic, raising questions as to which variables and which order of variables, whether social, economic, or biological, were responsible, and what relationships these bore to the ones governing both transitional and present-day patterns.

The investigation of these questions requires an open continuity model capable of encompassing diverse demographic regimes and a potentially unlimited range of time paths. I have argued that a framework of proximate determinants governing levels of exposure and resistance can furnish the elements of such a model, and this approach has been adopted by a number of recent studies. The first question suggested by this approach is the relative long-term importance of the exposure and resistance variables, a problem intimately related to that of the real-wage determinism considered in the paper by Kunitz and Engerman.

The failure of such determinism, and with it of narrowly resistance-based interpretations, implies that exposure variables were of greater importance in determining secular mortality change. Such variables fall into two categories: those stemming from 'structural' characteristics of populations, whether social, economic or spatial, and the so-called 'autonomous' factors linked to climatic variations or microbiological change in disease organisms. The latter are now widely seen as the only plausible explanation, but this conclusion can be avoided by a more rigorous specification of the determinants of exposure and a fuller incorporation of spatial structure into models of demographic regimes.

Landers's contribution considers these points in the context of pre-transitional metropolitan mortality, using London as an example. It is complemented by Dobson's paper on mortality elsewhere in south-east England, a paper which takes the argument one step further in looking at contemporary perceptions of spatial differences in mortality. These were affected by the physical and human geography of the region, but a key role was played by the latter's incorporation in a global economic and epidemiological system. Langford and Storey pursue the consequences of such incorporation beyond the boundaries of Europe, and into the twentieth century, showing how Sri Lanka's specialization as a primary producer opened it up to the world economic system and to the influenza pandemic of 1918-19.

Real wages, resistance and mortality

In the twentieth century socioeconomic mortality differentials, both within and, the low mortality of some poor countries notwithstanding, between countries, have been pronounced and enduring, whilst the secular decline in developed-country mortality has, broadly speaking, paralleled unprecedented long-term increases in *per capita* national income. Neither in the contemporary Third World, nor in transitional Europe, can this large-scale association be translated into narrow, mechanical, relationships between economic welfare and mortality decline. But did such relationships hold in pre-transitional demographic regimes?

The English case has received the greatest attention in this respect, both because of its substantive importance to the Notesteinian tradition and because it is relatively well documented. Much of this has focused on the relationship between mortality levels and so-called 'real wages', a measure of money wages' buying power. This relationship in turn is closely bound up with the nutritional hypothesis. This is because staple foods consumed a very large share of popular budgets, and recorded real-wage changes largely reflected changes in their price. Hence, if nutrition determined mortality levels, we would expect a strong correlation between these and real wages. Strictly speaking, such a correlation would itself not be unshakeable evidence for nutritional determinism—other components of expenditure might in principle be responsible—but in practice such an inference would be hard to avoid.

In fact, as Kunitz and Engerman point out, no such long-term correlation does emerge from the English data; Wrigley and Schofield's (1981) reconstruction of English population history 1541-1871 shows no positive association between life expectation and real wages over the period, and there is even some suggestion of a negative relationship at times. At least as surprising is the failure of the expected cross-sectional relationship. Mortality differentials are very hard to determine before the statistical era, but a comparison of data from the British peerage with the Wrigley-Schofield series reveals that the peerage had no systematic advantage over the general population before the latter part of the eighteenth century.

The absence of a real-wage correlation may in principle be explained in one of three ways: the real-wage data may provide an inadequate measure of economic well-being and thus of nutritional levels; the latter may have determined mortality through relationships too complex to be detected in simple comparisons of the kind described; or mortality variations may have been determined primarily by non-nutritional factors.

The measurement of welfare

Economic historians have generally used real-wage data for want of anything better,¹⁰ but such data are open to a number of objections. Changes in working hours, levels of unemployment, or payments in kind, may seriously distort the relationship between wages and earnings, and they may fail altogether to capture the experience of those, such as subsistence farmers, who were outside the market sector. Furthermore, as feminist writers have pointed out, individual welfare reflects both household income and its distribution within households, something on which real-wage data are entirely silent.

¹⁰ One alternative source, which has received increasing attention recently, is information on adult stature. For an important study, together with a valuable discussion of the general problem of measuring economic well-being in the past, see Floud et al. (1990).

Such imperfections must be taken seriously, but they are unlikely to obscure a 'true' underlying economic determination of English mortality. England's economy was sufficiently market-oriented for the welfare of wage earners to be an adequate indicator of conditions generally for most of the relevant period (Lee 1978). Similarly, discrepancies in the relationship between wages and earnings make short- and even medium-term comparisons problematic. But it is scarcely plausible that they could offset the long-term movements, of a generation or so, in food prices which underlie real-wage changes on this time scale. The failure of the expected real-wage relationship is thus unlikely to be a 'false negative', and even if it were so, the social elite's relatively high mortality would remain as an obstacle to a simplistic economic, nutritional determinism.

The form of the economic effect on mortality

The expected relationship might also be obscured by 'non-linearities' in the association between welfare and mortality. One example would be a temporally-lagged relationship due to cohort effects if childhood nutrition or morbidity has a substantial impact on mortality in later life. An extreme instance of this, alluded to by Kunitz and Engerman (this volume), is the recently postulated foetal determination of some adult mortality risks. Such effects could in principle mask those of secular changes in economic welfare and might also distort cross-sectional comparisons if social mobility is at all common. Cohort effects have been adduced to explain a number of phenomena including recent increases in Russian adult mortality (Anderson and Silver 1989), but it seems unlikely that they are responsible in the present case; the weight of childhood mortality in determining life expectancy will have muffled the impact of any long-term lagged relationship of wages on mortality, whilst mobility into the social elite was too restricted to have distorted cross-sectional comparisons (Stone 1984).

A second type of non-linearity could arise from threshold effects in the relationship between mortality and welfare causing the observed association to vary with shifts in overall national income. As Kunitz and Engerman (this volume) point out:

unless the risk of premature death is lower among the super rich than among the very rich and the risk among the very rich is lower than among the merely rich, then there must be threshold effects (p.40).

Again, however, it was bread prices that chiefly determined real-wage changes in the past, and since their principal effect was to vary the population's command over basic necessities it seems unlikely that any thresholds of this kind would have been crossed.

Non-economic factors

These two options having been excluded, we are left with the likelihood that mortality variations had non-economic causes, and Kunitz and Engerman consider two groups of possibilities.

Cultural: Probably the most widely canvassed of the non-economic variables, if also the most difficult to specify analytically, are cultural factors of the kind whose role in health transitions has been stressed by Caldwell. In a longer perspective, the relatively low mortality of minorities such as Jews is, as the authors point out, a remarkable and consistent feature of historical demographic research, but such research usually encounters the 'culture' variable in connection with infant-feeding practices and their influence on fertility and infant mortality. These can be substantial, but in the English case there do not seem to

have been changes in infant feeding on a sufficient scale to explain the observed mortality fluctuations, which in any case extended beyond infancy. England also seems to have been culturally more homogeneous than European nations such as France and Germany before the nineteenth century.

Political: The second group of factors are broadly political or administrative. Some relate to government spending and are thus strictly 'non-market' rather than non-economic; they include the redistribution of income through taxation and welfare payments, as well as the direct provision of entitlements to goods and services. Such entitlements assumed great significance as effective preventive, and subsequently curative, health measures developed from the later nineteenth century; Caldwell has demonstrated their current importance in mediating the relationship between national income and mortality. McKeown, as we have seen, seriously underestimated the role of government intervention in English mortality decline, and Kunitz and Engerman link this to a particularly Anglo-Saxon preoccupation with longitudinal associations between national income and mortality, a preoccupation which, they argue,

appears to reflect the same political culture that is associated with the developmental policies that the English-speaking (British) world has generally followed; that is, an emphasis on the growth of individual income as the most effective way to improve the quantity and quality of life (p. 44).

The same political culture may also have been responsible for obscuring systematically the important role of local government in redistributing wealth in pre-transitional England.¹¹ Here the 'Poor Law' system of social welfare apparently functioned as a remarkably efficient means of transferring wealth from the healthy to the sick and between households at different stages of the family-life cycle. The effects of this 'communal-risk sharing' in reducing both mortality and fertility are only now beginning to be understood but may well have contributed significantly to England's 'low pressure' demographic regime (Smith 1981; Slack 1990).

On a larger scale, however, the technological limitations of a pre-industrial economy, and the very modest amounts of wealth available for redistribution, meant that the impact of such governmental action on mortality was necessarily limited. Kunitz and Engerman attach primary importance to administrative action in the face of crises such as harvest failure, epidemics, and warfare. Here European states developed a series of responses such as quarantines, the provisioning of food-deficit regions and improvements in military organization which reduced mortality by impeding the diffusion of pathogens and thus lowering overall levels of exposure to infection.

Exposure and mortality

The failure of both secular and cross-sectional correlations between pre-transitional levels of mortality and economic well-being implies that these were not determined primarily by variations in nutritionally determined resistance to infection, and thus that exposure was their key proximate determinant. Kunitz and Engerman, as we have seen, echo the arguments of Omran in suggesting that administrative action in the course of the

¹¹ We should note here that Malthus's own writings played an important part in this, through his neglect of the important contribution made by such transfers to the maintenance of the nuclear family-household system with its associated marriage pattern (Laslett 1985).

seventeenth and eighteenth centuries successfully reduced the incidence of epidemics, and more generally, of infections with outcomes dominated by the conditions of exposure; thus lowering mortality. This is an attractive notion in many ways; it is a strong candidate to explain plague's disappearance from Western Europe, and such effects may also have contributed to the decline of eighteenth-century mortality. But it cannot be the whole answer, for it does not easily account for shifts in non-crisis mortality such as the late seventeenth century increase in England. This limitation, combined with the failure of the nutritional explanation, has fostered the widespread acceptance of the 'autonomous death rate' hypothesis, a contention that secular mortality changes reflected those in infectious micro-organisms, or in climate, and were thus altogether independent of economy and society (Chambers 1972; Perrenoud 1984).

Such a claim has important implications, and where historical demography is concerned these are strongly counter-heuristic; by locating the determinants of mortality change outside the human world the concept of the autonomous death rate effectively removes them from observation and analysis. From the viewpoint of health-transition theory it implies a radical-discontinuity model of mortality; pre-transitional mortality becomes a biological phenomenon, whilst transitional levels are socioeconomically determined, and medical technology becomes increasingly important at later stages.

Recent studies in historical epidemiology have sought to avoid this by taking a broader view of the determinants of exposure to infection and examining consequences of spatial structure for mortality patterns. Landers attempts to draw out their implications with a conceptual model of the proximate determinants of both exposure and resistance to infection. In this model, populations are conceived as networks whose nodes are the constituent individuals, whilst the links are pathways which allow pathogens to move between the latter and any biotic or other reservoirs of infection present in the environment.

At any given time the level of exposure depends on the density of this matrix (termed the degree of conduction present), the number and variety of pathogens (the population's pathogenic load) and the interaction between the two. On the resistance side, the main determinants are the immunological and nutritional status of members of the population. Changes in the pathogenic load over time reflect the introduction of new pathogens and the ability of the existing strains to persist in the population. The latter depends on an additional proximate determinant, the degree of retention, but the former is determined in a more complex manner since 'introductions' can occur in two distinct ways.

First, there are the truly 'autonomous' processes of microbial mutation and recombination through which new strains arise from the existing pathogens, and secondly there are physical introductions of new strains from the outside world. These depend both on the volume of physical movement of goods or people, and thus on a proximate determinant termed bounding, and on the pathogenic load of the outside world populations concerned. The latter are exogenous to the population in question, as are any microbial changes, whereas levels of conduction, bounding, and retention represent structural features of a population's epidemiological regime which jointly determine its 'exposure potential', just as its 'resistance potential' depends on diet.

Different combinations of proximate determinants give rise to different forms of epidemiological regime, with differences in both mortality levels and other features such as their short-term stability. In his paper, Landers uses this frame of reference to investigate the 'high potential' regimes thought to have characterized metropolitan centres in pre-

transitional Europe. Here the crowding, poor sanitation, and other environmental deficiencies combined to establish high levels of conduction and retention, whilst incorporation into national and global networks of migration and trade implied weak bounding, leading to a high exposure potential and the development of endemic pools of infection.

Under these circumstances mortality rates would be high, especially in the early years of life, but relatively stable in the short term as the high levels of immunity acquired by those surviving childhood infections would limit the scope for major epidemics. Landers shows how the aggregate vital data from London conform to the outlines of the high potential model but also how specific cause-of-death patterns varied over time in response both to autonomous factors and to structural features of the contemporary economy and society.

Medical topography

London was incorporated into a global epidemiological system by virtue of its role as an entrepôt and trading centre but also functioned as a 'primate' centre dominating England's urban system on both national and regional scales. Dobson's paper complements that of Landers in focusing on the epidemiological regimes of south-eastern England, the region surrounding the capital. Here again there is a striking pattern of spatial variation, which is not explicable by economic well-being or nutrition, but can be accounted for in terms of differences in exposure to infection arising from the physical and human geography of the region.

The basic contrast is between the high-mortality regimes of low-lying marshy and estuarine districts and the much lower levels characterizing dry upland districts, despite the fact that the latter experienced substantial economic stress for much of the period. The key to the marshlands' high mortality was their infestation by the malarial vector *a. atroparvus* and consequent endemic malaria. Elsewhere the more favourable conditions in the uplands were linked to their very lack of economic development. As Dobson remarks: 'The very fact that these regions were poor, backward, impenetrable, relatively isolated and inaccessible gave them a distinct advantage over their more frequented and busy counterparts' (p.89).

This was because the development of increasingly sophisticated economic networks reduced the bounding of local epidemiological regimes, promoted the diffusion of pathogens, and thus raised the level of exposure to infection among the populations concerned. The worst epidemiological effects of such economic modernization occurred in the Medway and Thames dockyard towns which suffered repeated epidemics of diseases such as bubonic plague and typhus. The pattern described by Dobson is thus one of local variation in exposure to infection within a system dominated by its proximity to the endemic reservoir of London, and incorporation within a global economic network, as well as its physiographic structure. The latter affected exposure potentials in some areas, but the realization of these depended on historically-specific changes in pathogenic load, for the marshland's excess mortality arose from plasmodia which had themselves been introduced

in an earlier episode of 'modernization' by Dutch engineers draining the Fens in the sixteenth century ¹² (Dobson 1980).

This raises the second of the two issues with which the author is concerned, for just as the pattern of malarial mortality was not a timeless, physiographically imposed, feature of the region's demography, so it was eliminated prior to the establishment of any 'scientifically correct' understanding of the underlying disease processes. Dobson demonstrates just how aware contemporary observers were of south-east England's 'contours of death', and the health consequences of the marshland's 'bad air' (or 'malaria'), even though they attributed these to 'morbific particles suspended in the atmosphere or deadly vapours emanating from telluric effluvia and sluggish waters' (p. 84).

The eventual successful assault on malaria in England proceeded through environmental manipulation from a basis of empirical observation and culturally informed intuitive judgements on dirt and health, rather than one of scientific knowledge (Dobson 1980). On another level, however, Dobson points to a further paradox concerning knowledge and behaviour, for it was precisely the healthy-upland areas which were emptying out in the course of the period, whilst the most unhealthy urban areas attracted immigrants. The forces of modernization, and the emergent economic world system, thus constrained individuals to migrate against the 'contours of death' despite their knowledge of the likely negative consequences.

Influenza in Sri Lanka

Studies in historical epidemiology have revealed the consequences flowing from pathogens' physical movement as regions become incorporated into global systems, or hitherto distinct systems merge into each other, whilst the Notesteinian tradition emphasizes economic modernization, through the spread of wage labour and the commoditization of the economy, as the route to lower mortality. The epidemiological consequences of these processes conflict since economic modernization implies an ever larger and more differentiated global system in which hitherto 'traditional' Third World populations have become incorporated as primary producers. This phenomenon, with its associated improvements in transport and communications, has reduced greatly the bounding of regional epidemiological regimes, thus raising their exposure potential even as its benefits have, at least in principle, fostered increased resistance, through better nutrition, and ameliorated other components of exposure.

Hitherto, the most dramatic realization of this global increase in exposure potential has been the influenza pandemic of 1918-19, and Langford and Storey's paper considers its impact on the population of Sri Lanka. Sri Lanka was incorporated into the global system through its specialization in the export of tea and rubber, a substantial proportion of the country's food requirement being imported, and correspondingly, the epidemic's entry points appear to have been Colombo, the island's main port, and Talaimanna in the north which connected the railway systems of Sri Lanka and the Indian subcontinent.

The impact of the epidemic on the island itself reflects the effects of economic factors on levels of conduction, as well as those of physical geography, and the interaction of elements within the pathogenic load. The surplus deaths, at some 50,000, represented

¹² Or possibly re-introduced, as malaria may have been present in Roman times and subsequently disappeared. It should be noted that sixteenth-century drainage, and related activity, also provided increased opportunities for mosquitoes to breed (M.J. Dobson, personal communication).

about 1.1 per cent of the island's population, but there was appreciable variation between districts; the dry zone generally fared worse than the wet, and estate districts were harder hit than the others. Some of these differences reflect interactions between influenza and malaria, but the likelihood that estate living increased exposure to infection is underlined by the higher mortality among Indian Tamils (generally employed as estate workers) compared to other ethnic groups on the island.

The origins of the pandemic itself remain tantalizingly obscure. Its proximity in time to the closing months of the First World War suggests that these two events were related, but the havoc wreaked among populations far from the theatres of war, and displaying very different social and economic characteristics, suggests that any causal relationship was both complex and indirect. In the Sri Lankan case wartime conditions had disrupted both export markets and transport, but the main economic hardship stemmed from the epidemic's own effects on food production, first in India and then on the island itself. These brought shortages, and although their potential impact on mortality cannot be gauged, Langford and Storey point to the importance of the Sri Lankan social response in mitigating their effects and in coping generally with the disruption caused by the crisis.

Health transitions

Research in historical epidemiology has shown how structural characteristics of populations affect the proximate determinants of mortality, raising or lowering exposure potentials, and how such potentials are realized in practice. The second group of papers in this volume is concerned with policy interventions aimed at manipulating exposure variables in the process of health transition and with the social, political, and cultural factors which have influenced their success. Fetter relates the concepts of historical epidemiology directly to the experience of the African mining camps, showing how ecological factors and power differentials interacted in the course of the region's incorporation into the global economic system. Mortality first increased and then fell, reflecting ecological changes in exposure to infection, but the speed and extent of both the decline and the ensuing health transition reflected the sociopolitical structure of the areas concerned.

Aaby and Bradley are both concerned with responses to specific diseases. Aaby asks how far medical intervention, in the form of vaccination, can reduce measles mortality without substantial nutritional improvements, a question which involves weighing the relative contributions of exposure and resistance variables to prior mortality. Malaria, the subject of Bradley's paper, was the model for the technology-intensive 'top-down' strategy of disease control which was so influential in the 1950s and 1960s. Bradley demonstrates, however, that these very characteristics, and the accompanying neglect of the proximate determinants of exposure, were implicated in the subsequent widespread failures, and shows how the immediate prospects for future success depend both on greater sensitivity to disease ecology and the enlistment of popular participation.

Landers's model of the proximate determinants of mortality emphasizes the distinction between potential and realized exposure to infection. We have seen how modernization can increase the former, through weaker bounding, even as realized exposure falls with reduced conduction and retention, leading in turn to a lower pathogenic load. The inter-continental diffusion of HIV has realized this increased potential, and is aggravated by the lack of specific curative or preventive medical techniques. Hence cultural and behavioural changes are needed to slow the spread of the virus, requiring in turn substantial changes in

the perception of risk. Cleland et al.'s contribution focuses on current perceptions among African populations together with the likely social bases for their modification.

The mines of Africa

The demographic implications of Europe's colonial expansion have been considered by a number of writers, but these have concentrated on its role in promoting the global spread of infection or on the mortality experience of Europeans in the tropics (for example, McNeill 1977; Ladurie 1981; Curtin 1989). Fetter's paper, by contrast, considers the historical epidemiology of African workers in the mines of Central and Southern Africa and the contribution this made to regional health transitions. Exploring both the utility and the limitations of an ecological approach, he begins by defining the mines as 'a set of discontinuous micro-environments defined by the presence of minerals which could profitably be exploited through western technology' (p.128).

Avoided by the indigenous inhabitants, these environments were afflicted by a pathogenic load imposing harsh mortality penalties on aggregated settlement. Like movement from the uplands to the dockside towns in Dobson's south-east England, migration to the mining camps thus ran against the contours of mortality and stemmed from the region's incorporation into a global economic system. Here, however, the 'pull' and 'push' of economic forces were mediated by the exercise of politico-military power and the Europeans' 'ability to coerce the earlier inhabitants into a variety of activities which they would not have otherwise chosen' (p.129).

The first mining camps experienced higher mortality than the villages from which their labour force derived, but in time this differential was reversed through a two-stage process. The need for productive mine workers forced the introduction of preventive techniques acquired in Europe's epidemiological transition together with others developed for tropical environments during colonial expansion. The second stage, however, was implemented more unevenly owing to differential power relations and clashes of interest within and between the white-minority regimes.

The high incidence of migratory labour was associated with excess pulmonary mortality. In Katanga, and later Northern Rhodesia, the solution of 'stabilizing' labour, by allowing mine workers long-term residence with their families, effectively freed the population to make its own adaptations to the new epidemiological regime, with the result that Katangan mortality fell by three-quarters between the quinquennia 1921-25 and 1931-35, to a level half that prevailing in contemporary Northern Rhodesia. The rejection of such policies south of the Zambezi reflected the political power of the settlers, especially that of white semi-skilled labour, relative to colonial and company bureaucracies. That black workers were denied urban tenure was, Fetter argues, to have important implications for the health transition, since the necessary changes in private hygiene required both physical facilities, and a predictability in daily life, denied to migrant labourers. Hence the epidemiological experience of these populations reflected both ecologically-conditioned proximate determinants of mortality and power relations between social and racial groups arising from the region's incorporation into a global-economic system.

The problem of malaria

The most persuasive argument for the exogenous 'death-control' model of Third World epidemiological transition was provided by the example of malaria, where DDT-backed control, and later eradication, programs seemed to yield major mortality gains in the post-

war decades. Yet, with certain exceptions, it is just these gains which have proved the most tenuous in the longer term and there is now a growing prospect that the mortality declines will be halted, or even reversed, by the spread of drug-resistant pathogens. The post-war programs were, as Bradley points out, technology-intensive and organized on the lines of military operations. In retrospect this contributed to their undoing, and in poorer countries 'the remarkable fact is not that eradication failed but that it came so close to success' (p. 152). One contributing factor was the failure to grasp variations in the epidemiological regimes of malarial regions; biological differences in the anopheline vectors, and thus the basic-case reproduction rate (BCRR), affected the regimes' degree of pathogenic retention and the consequent level of exposure.

Two general patterns emerge. In high retention environments the pathogenic load was sufficiently intense for malaria to become 'holoendemic' with mortality that was severe, stable and concentrated at early ages, in a pattern akin to those of the metropolitan centres described by Landers. Elsewhere, lower retention gave rise to a pattern of recurrent epidemics with little age-specificity. In holoendemic regimes, as in much of sub-Saharan Africa, a failure to recognize the configuration of proximate determinants led to health services being blamed for a problem that was, in Bradley's words, 'biologically intractable'.

The author distinguishes two styles of disease control: 'public health' and 'individual protection'. The former, aiming in this case to modify the environment and prevent vector breeding, had earlier been practised with some success in tropical environments of the kind described by Fetter, where investment resources, commercial opportunities, and administrative coercion made this form of 'top-down' manipulation feasible. The programs of the 1940s and 1950s represented the triumph of such an approach with a corresponding loss of interest in the individual patient as such.

Ironically, as such programs have come to grief, only the latter's self-medication has stopped mortality increasing in line with the prevalence of the disease. Hence, what Bradley describes as a 'long, disorderly sorting out process' has seen a revival of the individual-protection approach, allied to the emergent concept of primary health care. This, however, emphasizes trained personnel rather than capital investment. Furthermore, in the absence of a vaccine, effective prevention involves methods, such as insecticide-impregnated bednets, which require the perception of environmental risk by the population at large, their adoption of appropriate risk-related behaviour, and active collaboration with health professionals. Hence the social and cultural context of program implementation have become crucial to success.

Measles: resistance and exposure

The scope for direct program intervention as an independent variable in mortality decline is considered further by Aaby, whose argument is based on the relative contribution of exposure and resistance variables to mortality levels. The discussion is centred on the case of measles but has implications for other causes of death and, as his title implies, for historical health transitions in developed countries. Where no effective drug treatment exists direct medical intervention is confined to prevention, which in this instance can be effected through vaccination. But vaccination's impact on general mortality cannot be inferred simply from the number of measles deaths prevented; if these fall primarily among the weak and undernourished, then the lives gained may be lost disproportionately to other causes of death, with little overall gain.

Hence medicine's contribution to mortality decline in such instances depends very much on the anterior significance of resistance and exposure in determining the outcome of infection. Resistance to measles, of course, is determined partly by immune status, and thus the history of previous exposure, but nutritional factors have also been credited with a strong contribution. Aaby's paper, by contrast, emphasizes the importance of exposure variables, particularly the severity of the infective dose, in determining outcomes. For Aaby it is the locus of infection, the home or elsewhere, and the severity of the infecting case which determine the severity of the dose and thus the fate of the infected individual. He shows how these effects account for a variety of phenomena: ranging from birth-order effects to inter-country variations in lethality, which are inexplicable on the nutritional hypothesis.

The probability of infection in the home depends on the level of conduction, but the argument also implies an important short-term interaction between the latter and changes in pathogenic load. This phenomenon, which the author terms 'amplification', is a form of positive feedback stemming from the way the severity of the infecting case determines that of the infective dose an individual receives, and thus the severity of his or her own infection. The larger the proportion of severe cases, the greater the likelihood of being infected by such a case, and thus the larger the number of 'second generation' severe cases. For Aaby it is this potentially exponential rise in the number of severe cases which explains why institutional and virgin-soil epidemics often show increasing lethality over time.

Such exposure-based mechanisms can thus explain the characteristically high mortality of virgin-soil epidemics without recourse to the effects of either genetic selection or social breakdown on levels of resistance. For the understanding of health transitions in Western populations, the 'lessons for the past' include the likely role of changes in fertility, school provision, and child care, in affecting levels of exposure. Above all, however, the argument demonstrates conclusively the dual role of vaccination in both boosting individual immunological resistance and lowering the risks of exposure in a way which does not depend on improved nutrition.

AIDS: perceptions of risk

The role of modernization in raising the exposure potential of epidemiological regimes has recurred at a number of points in our discussion. Developments in long-distance transport, from steam ships to jet airliners, have substantially reduced regional bounding, and their effects have been reinforced by those in land transport from steam railways to diesel lorries. At the same time urbanization and, in some areas at least, behavioural change have partially offset the lowering of conduction effected by improved housing, sanitation and hygiene. The most sinister realization of this increased exposure potential, in recent years, has been the spread of HIV and AIDS, phenomena constituting a major challenge for medicine's contribution to health transitions. Life-saving drug therapies are unavailable; effective vaccines remain an uncertain prospect; and the epidemiology of AIDS itself appears to be more complex than that of many conditions encountered hitherto, varying between populations in ways that have yet to be fully understood.

Under these circumstances the only effective counter seems to lie in behavioural changes capable of reducing conduction. The anxieties of developed countries notwithstanding, it is in sub-Saharan Africa that mortality impacts are likely to be most severe, a prospect made all the more serious by the relatively modest progress of health transitions in much of the region. To the historical epidemiologist AIDS appears as a

disease of state formation, propagated through warfare and the growth of inter-regional trade, analogous to plague, typhus and other infections in early modern Europe. In Africa, however, the process has been more rapid and destabilizing partly because of imported technologies—in this case the assault rifle and the diesel lorry—and the offsetting gains made by Europe through the amelioration of subsistence crises have been less in evidence.

Africa cannot afford to wait for the administrative and economic changes which stabilized mortality in Europe, and the characteristics of the pathogen mean that the solutions to AIDS, though a collective problem, lie more in the domain of the individual. Here, as Cleland et al. emphasize, it is the perception of risk and the adoption of appropriate behaviour which is crucial; more precisely it is the formation of appropriate, individual-level, knowledge concerning the consequences of behaviour. This in turn requires the collaboration of the population with health professionals and a perception of the latter as both authoritative and legitimate. Here again, an understanding of cultural and social-structural variables is of the greatest importance. For individuals do not encounter or conceptualize risks as isolated, atomized, individuals, and groups 'based on a common activity ... on a common religious or political belief, and on gender may prove to be the key agents of change' (p.190).

Health transitions past and present

The proximate-determinants framework lets us relate past and present mortality transitions to each other and to the dynamics of pre-transitional mortality. A recurring feature of the papers in this collection has been the importance of exposure to infection, and its determinants, in different epidemiological regimes. Factors such as settlement density, communications, personal hygiene and certain features of the physical environment have been, and will in all probability continue to be, major influences on mortality among populations with very different socioeconomic and cultural characteristics.

Nonetheless, the present differs from all but, historically speaking, the most recent past in that all societies now have access to a scientific understanding of disease processes and the resulting curative and preventive techniques. It is this factor which Caldwell emphasizes in his comparison of contemporary and historical health transitions. Discrepancies between the two in such respects as the extent of mortality differentials by educational and socioeconomic status can, he suggests, be explained in terms of differences in the development of, and degree of access to, the medico-scientific knowledge base.

In the course of his argument Caldwell makes two key distinctions. The first is between a general scientific knowledge of disease processes and a knowledge of how, and in what context, specific disease risks are incurred: 'knowing that infection was caused by germs was not the same thing as knowing that there were germs in milk, that they existed there in sufficient densities to endanger a child's health, that they could be destroyed—at least temporarily—by boiling...' (p.211). The acquisition of this more specific, empirical, knowledge was an uneven, and sometimes hazardous, process; as when green vegetables were over-boiled to ensure their sterility, or tuberculin therapy employed in the treatment of tuberculosis.¹³ Alongside this is the distinction between science as specific knowledge, such as germ theory, and a culturally conditioned 'scientific outlook', which includes

¹³ I am grateful to Dr Charles Webster for drawing my attention to these examples.

procedures for acquiring and organizing such knowledge but extends to broader issues such as a respect for science and scientific authority. Caldwell suggests that the latter was widespread already in the West before the formation of germ theory but has to be acquired as part of the educational process in Third World countries, giving rise to important infant and child mortality differentials with respect to maternal education.

The findings of historical epidemiology suggest a further general distinction between conceptual knowledge, embodied in a framework of generalizing abstractions of which scientific knowledge is but one example, and the kind of situational, empirical, knowledge open to anyone capable of systematic observation and able to connect causes with effects. This distinction should not be pushed too far: empirical knowledge requires generalizing concepts if it is to be intelligible, but it is useful in that the same empirical knowledge, such as the health risks attaching to a given locality or behaviour, may be expressed in very different conceptual frameworks, so that one may do entirely the right thing for completely the wrong reasons. Dobson's work on medical topography illustrates this point well, as do the successful hygienic and sanitary measures developed in the century before the formation of germ theory. What is interesting about the Western, particularly, perhaps, the English, experience¹⁴ is the extent to which scientific discoveries underwrote previous empirical knowledge of relationships between health and such things as cleanliness and ventilation, thereby legitimating a range of pre-existing attitudes and behaviour.

That such benign consequences were not inevitable can be seen in the limited range of counter-examples. We saw above that a correct general understanding of disease processes sometimes gave rise to specific measures that were quite inappropriate in the circumstances. More importantly, scientific knowledge may sometimes come into conflict with prior empirical knowledge and thereby undermine entirely appropriate health-related behaviour. Thus in nineteenth-century London, where the health risks of heavily-polluted Thames water were widely recognized, industrial chemists refuted popular views by showing it to be 'unpolluted' (that is, free of chemical toxins) and thereby helped delay government action until the new microbiology re-affirmed the legitimacy of popular empirical knowledge (Luckin 1986).

Alongside the transformation of mortality and epidemiology, the concept of health transition involves changes in behaviour and the development of a 'health culture'. This involves the formation, among populations, of empirical knowledge concerning health risks and appropriate health-related behaviour: empirical knowledge which requires the articulation of scientific medicine's general conceptual knowledge with the concrete circumstances of given environments and cultural milieux, and thus a certain mutual regard between lay populations and health professionals. Professional indifference to popular perceptions and prior empirical knowledge risks the development of inappropriate program interventions and behavioural prescriptions, whilst popular acceptance of such prescriptions requires an acceptance of professional authority and legitimacy. Such a politics of knowledge in the health field, with its associated 'educated, capable and demanding public', cannot be separated from broader political questions but is of the

¹⁴ This judgement is based on what seems to have been a marked increase in concern with personal hygiene in late eighteenth and early nineteenth-century England, and a corresponding disparagement of continental, especially French, standards. This apparently represented a break with the earlier English association of 'excessive' personal cleanliness with sexual decadence, something the English also identified with France (Razzell 1974; Stone 1977; Corbin 1986).

utmost importance. Historical epidemiology has demonstrated the ambiguous and equivocal character of economic modernization's relationship to mortality decline in the past, and there is nothing in the experience of sub-Saharan Africa to suggest that this character has changed. Health transition is not an inevitable process; it is a goal for which to strive.

References

- Anderson, B. and B.D. Silver. 1989. Patterns of cohort mortality in the Soviet population, *Population and Development Review* 15:471-501.
- Anderson, M. 1988. *Population Change in North-Western Europe, 1750-1850*. London: MacMillan.
- Appleby, A.B. 1980. The disappearance of the plague: a continuing puzzle. *Economic History Review* (2nd ser.) 33:161-173.
- Baechler, J., J.A. Hall and M. Mann eds. 1988. *Europe and the Rise of Capitalism*. Oxford: Basil Blackwell.
- Bongaarts, J. 1978. A framework for analyzing the proximate determinants of fertility. *Population and Development Review* 4,1:105-32.
- Bowler P.J. 1989. *Evolution: The History of an Idea*. Berkeley: University of California Press.
- Caldwell, J.C. 1986. Routes to low mortality in poor countries. *Population and Development Review* 12:171-220.
- Caldwell, J.C. 1990. Introductory thoughts on health transition. Pp. xi-xiii in *What We Know about Health Transition: the Cultural, Social and Behavioural Determinants of Health*. eds. J. Caldwell, S. Findley, P. Caldwell, G. Santow, W. Cosford, J. Braid and D. Broers-Freeman. Canberra: Health Transition Centre, The Australian National University.
- Chambers J.D. 1972. *Population, Economy and Society in Pre-Industrial England*. Oxford: Oxford University Press.
- Corbin, A. 1986. *The Foul and the Fragrant: Odor and the French Social Imagination*. Leamington Spa: Berg.
- Curtin, P.D. 1989. *Death by Migration*. Cambridge: Cambridge University Press.
- Dobson, M. 1980. 'Marsh fever': a geography of England. *Journal of Historical Geography* 6:357-389.
- Dobson, M. 1989a. The last hiccup of the old demographic regime: population stagnation and decline in late seventeenth and early eighteenth-century south-east England. *Continuity and Change* 4,3:395-428.
- Dobson, M. 1989b. Mortality gradients and disease exchanges: comparisons from old England and colonial America. *Social History of Medicine* 2,3:259-298.
- Flinn, M.W. 1974. The stabilisation of mortality in pre-industrial western Europe. *Journal of European Economic History* 3:285-318.
- Flinn, M.W. 1979. Plague in Europe and the Mediterranean countries. *Journal of European Economic History* 8:139-46.
- Flinn, M.W. 1981. *The European Demographic System, 1500-1820*. Brighton: Harvester.
- Floud, R., K. Wachter and A. Gregory. 1990. *Height, Health and History: Nutritional Status in the United Kingdom, 1750-1980*. Cambridge: Cambridge University Press.
- Foucault, M. 1970. *The Order of Things: The Archaeology of the Human Sciences*. New York: Pantheon Books.
- Frenk, J., J.L. Bobadilla, C. Stern, T. Frejka and R. Lozano. 1991. Elements for a theory of the health transition. *Health Transition Review* 1:21-38.

- Gutmann, M.P. 1980. *War and Rural Life in the Early Modern Low Countries*. Princeton: Princeton University Press.
- Hill, A.G. 1990. Understanding recent fertility trends in the Third World. Pp. 146-163 in *Fertility and Resources*, eds. J. Landers and V. Reynolds. Cambridge: Cambridge University Press.
- Hodgson, D. 1988. Orthodoxy and revisionism in modern American demography. *Population and Development Review* 14:541-69.
- Johansson, S.R. and C.M. Mosk. 1987. Exposure, resistance and life expectancy: disease and death during the economic development of Japan, 1900-1960. *Population Studies* 41:207-236.
- Kunitz, S.J. 1987. Explanations and ideologies of mortality patterns. *Population and Development Review* 13:379-408.
- Ladurie, E. Le Roy 1981. A concept: the unification of the world by disease (fourteenth to seventeenth centuries). Pp. 28-83 in *The Mind and Method of the Historian*, ed. E. Le Roy Ladurie. Brighton: Harvester.
- Landers, J.M. 1987. Mortality and metropolis: the case of London 1675-1825. *Population Studies* 41:59-76.
- Landers, J.M. 1993. *Death and the Metropolis: Studies in the Demographic History of London 1670-1830*. Cambridge: Cambridge University Press.
- Laslett, P. 1977. Characteristics of the Western family considered over time. Pp. 12-49 in *Family Life and Illicit Love in Earlier Generations*, ed. P. Laslett. Cambridge: Cambridge University Press.
- Laslett, P. 1985. Gregory King, Robert Malthus and the origins of English social realism. *Population Studies* 49:351-62.
- Lee, R.D. 1978. Models of preindustrial dynamics with applications to England. Pp. 155-208 in *Historical Studies of Changing Fertility*, ed. C. Tilly. Princeton: Princeton University Press.
- Lee, W.R. ed. 1979. *European Demography and Economic Growth*. London: Croom Helm.
- Lorimer, F. 1946. *The Population of the Soviet Union: History and Prospects*. Geneva: League of Nations.
- Luckin, B. 1986. *Pollution and Control: A Social History of the Thames in the Nineteenth Century*. Bristol: Adam Hilger.
- Mathias, P. 1979. Swords and ploughshares: the armed forces, medicine, and public health in the late eighteenth century. Pp. 265-285 in *The Transformation of England*, ed. P. Mathias. London: Methuen.
- McFarlane, N. 1989. Hospitals, housing and tuberculosis in Glasgow, 1911-51. *Social History of Medicine* 2:59-86.
- McKeown, T.R. 1976. *The Modern Rise of Population*. London: Edward Arnold.
- McKeown, T.R. and R.G. Brown. 1955. Medical evidence relating to English population changes in the eighteenth century. *Population Studies* 9:115-141.
- McKeown, T.R. and R.G. Record. 1962. Reasons for the decline of mortality in England and Wales during the nineteenth century. *Population Studies* 16:94-122.
- McNeill W.H. 1977. *Plagues and Peoples*. Oxford: Basil Blackwell.
- McNeill W.H. 1980. Migration patterns and infections in traditional societies. Pp. 27-36 in *Changing Disease Patterns and Human Infections*, eds. N.F. Stanley and R.A. Joske. London: Academic Press.
- Mercer, A.J. 1985. Smallpox and epidemiological-demographic change in Europe: the role of vaccination. *Population Studies* 39:287-308.
- Mosley, W.H. and L.C. Chen. 1984. An analytical framework for the study of child survival in developing countries. Pp. 25-48 in *Child Survival: Strategies for Research*, eds. W.H. Mosley and L.C. Chen. Supplement to *Population and Development Review* Vol. 10.

- Notestein, F.W. 1945. Population: the long view. Pp. 36-57 in *Food for the World*, ed. T.W. Schulz. Chicago: University of Chicago Press.
- Omran, A.R. 1971. The epidemiologic transition: a theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* 49:509-538
- Omran, A.R. 1983. The epidemiologic transition theory: a preliminary update. *Journal of Tropical Pediatrics* 29:305-316
- Perrenoud, A. 1984. Mortality decline in its secular setting. Pp 41-69 in *Pre-Industrial Population Change*, eds. T. Bengtsson, G. Fridlitzius and R. Ohlsson. Stockholm: Almqvist and Wiksell.
- Perrenoud, A. 1991. The attenuation of mortality crises and the decline of mortality. Pp. 18-37 in *The Decline of Mortality in Europe*, eds. R.S. Schofield, D. Reher and A. Bideau, Oxford: Oxford University Press.
- Post, J.D. 1977. *The Last Great Subsistence Crisis in the Western World*. Baltimore: Johns Hopkins University Press.
- Post, J.D. 1985. *Food Shortage, Climatic Variability and Epidemic Disease in Pre-Industrial Europe: The Mortality Peak of the Early 1740s*. Ithaca: Cornell University Press.
- Preston, S.H. 1976. *Mortality Patterns in National Populations*. New York: Academic Press.
- Preston, S.H. 1980. Causes and consequences of mortality declines in less developed countries during the twentieth century. Pp. 289-359 in *Population and Economic Change in Developing Countries*, ed. R. Easterlin. Chicago: University of Chicago Press.
- Razzell, P. 1974. 'An interpretation of the modern rise of population in Europe' - a critique. *Population Studies* 28:5-17.
- Razzell, P.E. 1977. *The Conquest of Smallpox*. Fittle, Sussex: Caliban.
- Rostow, W. 1960. *The Stages of Economic Growth*. Cambridge: Cambridge University Press.
- Schofield, R.S. 1984. Population growth in the century after 1750: the role of mortality decline. Pp. 17-39 in *Pre-Industrial Population Change*, eds. T. Bengtsson, G. Fridlitzius and R. Ohlsson. Stockholm: Almqvist and Wiksell.
- Slack, P. 1990. *The English Poor Law 1531-1782*. London: MacMillan.
- Smith, R. M. 1981. Fertility, economy and household formation in England over three centuries. *Population and Development Review* 7:595-622.
- Stone, L. 1977. *The Family, Sex and Marriage in England 1500-1800*. London: Weidenfeld and Nicholson.
- Stone, L. 1984. *An Open Elite? England 1540-1880*. Oxford: Oxford University Press.
- Szreter, S. 1988. The importance of social intervention in Britain's mortality decline c1850-1914: a re-interpretation of the role of public health. *Social History of Medicine* 1:1-38.
- Tilly, C. 1978. Introduction. Pp. 3-56 in *Historical Studies of Changing Fertility*, ed. C. Tilly. Princeton: Princeton University Press.
- Vallin, J. 1991. Mortality in Europe from 1720 to 1914: long-term trends and changes in patterns by age and sex. Pp. 38-67 in *The Decline of Mortality in Europe*, eds. R.S. Schofield, D. Reher and A. Bideau, Oxford: Oxford University Press.
- Watkins, S.C. and J. Menken. 1985. Famines in historical perspective. *Population and Development Review* 11:165-170.
- Wrigley, E.A. 1983. The growth of population in eighteenth century England: a conundrum resolved. *Past and Present* 98:121-150.
- Wrigley, E.A. and R.S. Schofield. 1981. *The Population History of England 1541-1871: A Reconstruction*. London: Edward Arnold.

The ranks of death: secular trends in income and mortality*



Stephen J. Kunitz¹ and Stanley L. Engerman²

¹Department of Community and Preventive Medicine, University of Rochester Medical Center, 601 Elmwood Avenue, Rochester, NY 14642, USA

²Department of Economics, Hark 203, University of Rochester, Rochester, NY 14627, USA

I've been rich, and I've been poor, and rich is better - Pearl Bailey

Abstract

This paper is about the reasons why there is so rarely a strong association between the secular trend in real wages and mortality. We classify the reasons in several different categories: epidemiological, sociological, geographical, demographic, technologic, and economic. In addition, we offer some speculations on why there has been an expectation that such an association would exist. We suggest that this expectation is found primarily in British work (including Britain overseas) and has to do with the history of the British experience of industrial and economic growth and its pervasive interpretations.

Introduction

That the secular trends in *per capita* income or real wages do not invariably predict the secular trend in mortality is on the face of it surprising, for everyday experience suggests to us, as it did to our ancestors, that it is better to be rich than poor, and that the life chances of the rich are generally better than those of the poor. Nonetheless, as Wrigley and Schofield (1981:310–332), have shown using several centuries' worth of British data, the association between real wages and mortality is only evident in a few years of significant economic decline, and many years of crisis mortality were not preceded by declines in wages.¹ Livi-Bacci (1990:99-107), has summarized comparable data from other European populations, and Pope (1992), as well as others, has demonstrated a comparable

*Our initial assignment from John Landers, the conference organizer, was to discuss why the usual expectation of a secular relationship between changing real wages and changing mortality was not always realized. We more broadly interpreted the question to relate to class and status ('rank') differences in mortality. It should be clear that the 'ranks of death' we refer to differ from those suggested in the pioneering work of P.M. Ashburn as well as its use in the familiar poem by Thomas Moore:

*The Minstrel Boy to the war has gone
In the ranks of death you'll find him.*

¹ It should be noted that while the expected relationships would be between income and mortality, the longer availability of estimates for real wages than for incomes means that many historical studies use the former as a proxy for the latter. In addition, there are rather complex issues relating to the expected form of the relationship between income and mortality which will influence the nature of the procedures used for analytical purposes. These reflect not only the specifics of the epidemiological relationship between income-related variables and mortality but also the statistical procedures used in examining the long-term time series in which there are attempts to distinguish secular trends from business cycles.

phenomenon in the United States in the nineteenth century: in the antebellum period, at a time of rapid economic expansion, life expectancy of native-born white Americans was actually deteriorating.

In this paper we shall describe secular patterns of change in income and mortality; suggest some of the reasons why commonsense observations of the association between income and life chances are not reflected in the secular association between them; explore some of the reasons why, however, such a relation has commonly been assumed to be the case; and discuss some ways of thinking about the causes of disease and death that may be applicable to considerations of the impact of real wages and income on mortality.

Secular changes in income and mortality

One of the major characteristics of the modern world has been a substantial secular decline in death rates at all ages as well as a marked diminution in fluctuations in mortality from one year to the next. Starting most dramatically with eighteenth-century Britain, the declines spread to most of the developed nations of Europe and its overseas extensions in the nineteenth century, and then in the twentieth century to most of the rest of the world (see Table 1). For the developed nations, of the 25-point decline in crude death rates between 1750 and 1970 20 per cent occurred before 1900 and over one-half after the First World War. For the less-developed regions, of the estimated 20-point decline in death rates, all occurred in the twentieth century, more than one-half coming in the post-World War II era. The decline in fluctuations in death rates, a precondition to any sustained mortality decline, occurred first in the developed nations by the early nineteenth century, whereas in less-developed parts of the world such dampening occurred much later in the nineteenth century.²

With the well-known divergence between developed and less-developed nations in the growth and levels of *per capita* income over this period, there was a sharp change in the cross-national relation between income and mortality. The initial differences in mortality before the widespread onset of modern economic growth appeared relatively minor, at least in comparison with those which emerged in the epoch between the middle of the nineteenth century and World War II. After World War II, the accelerated mortality decline in the less-developed nations meant a narrowing of the worldwide gap, and a movement toward the narrower differentials of the more distant past. Such narrowing has continued, at least into the 1980s, and, even while aggregate differences in mortality remain which appear correlated with income levels, these appear relatively small by historical standards.³

Table 1

Estimated and conjectured average annual crude birth rates, crude death rates, and rates of natural increase for currently more developed and less developed regions, from 1750 to 1970 (rates per 1,000 per year)

Period	More-developed regions				Less-developed regions		
	Birth rate	Death rate	Natural increase		Birth rate	Death rate	Natural increase
1750–1800	38	34	4		41	37	4

² Sources: Stolnitz (1955); Kuznets (1966:Chap. 2); United Nations (1973: Vol I. Chap. 5); Flinn (1981); World Bank (1990:230–231).

³ See Gwatkin (1980), and Stolnitz (1955). There may be some reversal of the decline in the less-developed nations, but mortality rates still will be lower than they had been earlier. Nevertheless, the crude death rates in the lowest-income region in sub-Saharan Africa in the 1980s are below the rates in the developed nations before World War I.

1800–1850	39	32	7	41	36	5
1850–1900	38	29	9	40	38	2
1900–1910	34	21	13	41	34	7
1910–1920	26	23	3	40	37	3
1920–1930	28	16	12	41	31	10
1930–1940	22	14	8	41	29	12
1940–1950	20	15	5	40	28	12
1950–1960	22	10	12	43	22	21
1960–1970	20	9	11	41	17	24

Source: United Nations (1971:Table 3)

Clearly these mortality changes seem related to what some would call modern economic growth or, used descriptively, modernization. The concept of modernization often includes, among other things: substantial growth in income; an eventual narrowing of the distribution of income among individuals and families within a nation; an increased role for the government in the provision of goods and services within the economy; a broadened application of the results of modern science and technology; a shift in the location of population from rural to urban areas; and a reduction in the share of production in the agricultural sector and increase in the share from industry and services.⁴ Such changes are not inevitable, for it is clear that not all nations have experienced all or even some of these changes. It does appear that as economic growth, or the desire to achieve growth, has spread, many of these features have emerged in less-developed nations, even though it is possible for major social and economic changes to occur without the accompanying full range of modern development, and for some countries or parts of countries to remain backward even though the population may wish for economic growth and modernization. Thus, for example, the diffusion of modern technology may permit changes in a country that are seemingly not related to the level of its national income.

Britain, the first industrial nation, had a pattern of mortality decline basically consistent with that of the developed nations described above. Indeed, both because it paradigmatically typifies the expectations of modernization and also because of the great depth of historical research on the issue, Britain is no doubt the nation whose mortality decline has been most studied and which has therefore done most to set the terms of the ensuing debate. According to Fogel's (1986) calculations of standardized death rates, from 1700 to 1980 the British death rate fell to a quarter of its earlier level, from 28 per 1,000 to seven per 1,000, with 80 per cent of this decline occurring after 1850, and about a half of the overall decline occurring in the twentieth-century. According to Crafts's (1985:44-47) estimates, British economic growth averaged 0.3 per cent per annum in the years 1700 to 1830. Matthews, Feinstein and Odling-Smee (1982:22), indicate that the British growth of Gross Domestic Product per man-year averaged about 1.2 per cent per annum from 1856 to 1973. Clearly, the accelerated increase in the growth of *per capita* income is associated with an accelerated decline in mortality, as well as with dramatic shifts in the location of the population from rural to urban areas, and with significant changes in the contribution of various diseases to mortality; a shift from epidemics to endemic infectious diseases and, finally, to a phase characterized by chronic diseases.⁵

The early date of British growth, Britain's unique patterns of household organization, the great role played by moneyed-wage labour, and the importance of Marxian and other debates concerning modern

⁴ See, among many sources, Kuznets (1966). Obviously not all these changes have occurred in all places and it may be that in the future the patterns will not be the same as in the past. For example, in many countries tourism, rather than industry, is seen as an engine of growth.

⁵ With this has come a shift in the age-structure of mortality, with an increasing proportion of elderly among decedents. See Omran (1971).

economic growth, all make the centrality of Britain to debates on the relation of income and mortality understandable. But precisely these and other characteristics may limit the applicability of the British discussion to other nations. For example, the more limited role of state intervention in the process of economic change in Britain than elsewhere in Western Europe may mean a greater reliance on individual-level behaviour and market mechanisms than upon state-directed measures. The early onset of economic growth, which predated many scientific, technological, and medical measures, meant that there were fewer personal and public-health interventions available to Britain at any level of national income than were accessible to follower nations. And the economic growth patterns of follower nations may not resemble fully that of the leader nation, if only because that niche in the world economy had already been filled.

Lastly, the greater salience of wage labour in Britain, and the long time period covered by the historical record, have made possible the use of an independent variable—real wages—that is sometimes limiting and can be misleading. Given the lack of a fully articulated concept of national income before the 1930s, and, thus, the inability to relate income and mortality for much of the past, the widespread use of measures of real wages (nominal wage rates divided by a relevant consumer price index) is understandable. Real wage data are more generally available in the past; they go back for longer periods than incomes can be constructed; and they permit details concerning intertemporal changes that most income estimates do not. As a result, they provide incomplete measures of economic well-being for societies in which farmers represent a large proportion of the population (more typical outside of Britain). Moreover, because wages are paid to individuals, they may not provide the best possible measure of family well-being.

Clearly, then, worldwide there has been an association between economic growth and reduced mortality. Moreover, because of its leading role in the economic expansion of the modern era, arguments that may have been adequate to explain the association in Britain have often been extrapolated to explain the patterns of mortality change elsewhere. However adequate or inadequate such explanations have been in respect of the British situation, it is inappropriate to extrapolate them to other populations. For there has been a great deal of diversity among peoples, much of which contributes to changes and differences in mortality in ways that may obscure or make irrelevant the significance of the level of and changes in wages that have so dominated debates about the British pattern.

Some problems in assessing the association between income and mortality

There are various categories of reasons which explain why it might be difficult to see, at the national level, a relation between wages and mortality. We shall discuss these briefly, using both longitudinal and cross-sectional studies for our examples.

Epidemiologic

In an early study of the relationship between social status and mortality, Antonovsky (1967) argued that life expectancy of rich and poor tended to be the same in periods when epidemics and pandemics dominated the mortality regime, and that differences began to emerge only when epidemics abated. Life-expectancy patterns of the British peerage and the general population generally support this view (Hollingsworth 1964; Wrigley and Schofield 1981). For it was only in the mid-eighteenth century that the experience of the two groups diverged. This seems to have been at a time when epidemics were receding and when nutritional status, domestic arrangements, and individual hygiene began to have a relatively greater impact upon health than was the case previously (Kunitz 1983, 1987).

Another example, presented by Curtin (1989), is taken from the late nineteenth and early twentieth century experience of the British Army in India. As Table 2 indicates, officers died at higher rates than

other ranks.⁶ Unfortunately, the data are not age adjusted, and it is possible that the officers were older than the others and thus would have had higher mortality on that basis. Nonetheless, it seems plausible to argue that exposure to a new disease environment would have adversely affected the experience of all ranks about equally.

Table 2
Death rates of officers and other ranks of the British army in India, 1899–1913

Years	Average annual rate per 1,000	
	Officers	Other ranks
1899–1908	11.97	11.63
1909–1913	6.27	4.74

Source: Curtin (1989:99)

A third example is taken from a series of studies by Barker and his colleagues (Barker and Osmond 1987; Barker 1989; Barker et al. 1989), in which birth weight, weight at one year of age, and early environmental conditions have been associated with the risk of death from cerebrovascular and ischaemic heart disease, as well as other conditions during adulthood. The causal mechanisms are not at all clear and presumably involve interactions with adult diet and other factors. Whatever the mechanism, the increase in relative risk of death from ischaemic heart disease as birth weight and weight at one year decline is very impressive. If there is an association between prenatal and early childhood environment on the one hand and risk of premature death in adulthood on the other, then the temporal association between income and mortality, even when lagged by one or several years, would not necessarily be significant. This would be particularly true in open societies with a great deal of upward and downward social mobility between generations, and in eras when non-infectious diseases dominated the epidemiologic regime.

The point of this example, then, is that cohort effects of which we have very little understanding as yet may well transmit increased or decreased risks of death across generations in such a way as to weaken contemporary associations between income and mortality.

Sociologic

The social organization and culture of groups may well have an impact upon health status and the risk of death that is relatively independent of income. Our first example is taken from a contemporary prospective cohort study of the associations among social support, socioeconomic status, and mortality (Berkman and Breslow 1983:Chap. 4). The measure of social support included marital status, contacts with friends and relatives, and involvement in church and non-church groups. The data are displayed in Table 3.

Table 3
Age-adjusted mortality rates from all causes (per 1000): social network index and socioeconomic status, men and women ages 30–69, 1965–1974

⁶ Curtin (1989) has pointed to the contemporary argument that this might be explained as the result of different degrees of discipline applied to the groups.

Social network index	Socioeconomic status				
	Lower	Lower middle	Middle	Upper middle	Upper
Men					
Least connections	11.8	12.7	16.6	11.7	7.3
Most connections	8.0	11.1	6.3	8.1	4.2
Total	9.9	11.7	9.5	9.1	5.2
Women					
Least connections	9.9	7.3	8.2	6.8	5.7
Most connections	7.3	1.1	4.9	4.1	5.2
Total	8.8	4.6	6.6	5.2	5.4

Source: Berkman and Breslow (1983:137)

Several results are noteworthy. First, in general, people of high socioeconomic status had larger networks than people of low status (data not shown). Secondly, within each socioeconomic stratum, people who reported a lot of contact with others had better survival over the nine-year follow-up period than people with little contact; even controlling for health status at the outset of the study. Thirdly, there were major differences in mortality across status groups associated with the social network index. For example, upper-middle status men with low network scores had higher mortality (11.7 per 1000) than lower status men with high network scores (8.1 per 1000). Thus, depending upon the socioeconomic composition of the population and the variations in social organization within and among strata, there might or might not be an association between income and mortality.

Our second example is taken from Woodbury's (1926) classic study of infant mortality in American cities from 1911 to 1915. It was observed that while income had the expected relationship to infant mortality, there were major differences among ethnic groups that could not be explained either by income (measured as earning of the father) or by feeding practices. The data are displayed in Table 4 and indicate the following: when level of income is controlled, Jewish infants had about half the mortality rate of white native-born infants and substantially lower rates than infants of other nationalities and races as well; Italians had half the rate of Portuguese; and Germans had about the same rate as native-born whites.

Table 4
Relative mortality by nationality of mother, when influences of type of feeding and earnings of father are eliminated, 1911–1915

Colour and nationality of mother	Ratio of actual to expected deaths %			
	On basis of average rate	Eliminating feeding	Eliminating earnings	Eliminating both
Total	100.0	100.0	100.0	100.0
White	97.4	96.9	99.2	98.3
Native	84.4	80.9	95.3	88.5
Foreign-born	114.2	119.4	103.4	109.3
Italian	93.4	109.9	76.6	86.8
Jewish	48.1	55.6	48.1	49.3
French-Can.	154.0	127.9	154.9	127.6
German	92.7	94.2	92.7	97.0
Polish	141.4	169.5	118.1	149.5

Portuguese	180.1	166.5	146.9	141.1
All other	116.5	119.9	106.2	106.5
Coloured	138.8	149.3	108.4	118.4

Source: Woodbury (1926:114)

These results for Jews are not unique. Similar findings have been reported from Europe as far back as the late eighteenth century (Schmelz 1971). A variety of explanations have been offered, ranging from the attention lavished on Jewish infants by their mothers, to birth spacing, to the observation of dietary laws, and to the bactericidal effects of chicken soup. It is not clear even now what the best explanation is, but all are agreed that it is a cultural phenomenon that has been largely independent of income.

Our third example is drawn from Caldwell's (1986) study of several anomalous countries which have both low income and low mortality and thus violate the more commonly observed association between low income and high mortality. The two nations (Costa Rica and Sri Lanka) and one state (Kerala, in India) he scrutinized particularly closely, had several things in common which resulted in this remarkable accomplishment: female autonomy; considerable investment in both education and health services on Western lines, with females achieving educational levels similar to those of males; efficient, responsive health services accessible to all regardless of income; provision of either a nutritional floor or food distribution in an egalitarian fashion; and universal immunizations and intensive prenatal and postnatal care. In each case these measures derive from a history and political culture which has led to 'a broad social consensus as to the value of educational and health goals, and as to their cost, for successive governments to accept most of their adversaries' innovations instead of nullifying them' (Caldwell 1986:210).

These three examples have been meant to illustrate the fact that at the individual, ethnic-group, and national levels, cultural and social-organizational factors operating independently of income have a profound influence on mortality at all ages. Thus we see once again that non-economic factors may be of enormous significance and that, depending upon the composition and culture of the population being observed, income may have greater or lesser effects on death rates.

Geographic

Under this rubric we include the well-known differences between rural and urban mortality that prevailed until early in the present century. Table 5 displays age-adjusted death rates by occupational class for males 15 years of age and older in England and Wales in 1910-1912 at a time when the differences were still very much in evidence. Within the urban occupations there is a clear gradient of mortality by occupational status. The differences are small or non-existent between farmers and farm labourers. For our purposes, however, what is especially significant is that farm labourers have lower mortality from virtually every cause, including those not shown here, than much more highly-paid urban professional and salaried workers.

Table 5

Age-adjusted mortality rates (per 100,000) for men 15 years of age and above in England and Wales by occupational class, 1910-1912

Occupational group	Phthisis (tuberculosis)	Non- pulmonary tuberculosis	Bronchitis	Pneumonia	Diabetes	Cancer
Professional and salaried	145	14	52	71	25	126

Skilled labour	180	14	110	81	15	131
Unskilled labour	244	19	172	149	11	159
All wage earners	172	15	122	97	12	129
Farmers	61	10	55	60	20	101
Farm labourers	85	14	67	55	7	93

Source: Collins (1927:23–24)

A similar result emerges from an analysis of infant and child mortality rates in urban and rural Holland in 1877-1881 (see Table 6). The rates of death were lower among the children of the rural poor than they were among the children of the urban well-to-do.

As in our previous examples, these too suggest that the composition of the population, in this case rural-urban mix, can have a profound influence on the association between income and mortality.

Table 6
Mortality rates per 1,000 live births for children in Holland by economic status, 1877–1881

Economic status	Rates per 1,000 live births			
	Urban ^a		Rural ^b	
	< 1 year	2–5 years	< 1 year	2–5 years
Rich	93	52	110	27
Well-to-do	139	94	91	52
Less well-to-do	157	103	107	64
Poor	166	115	112	70

^aRotterdam and Dordrecht

^b40 rural communities

Source: Collins (1927:45)

Demographic

It is plausible that with advancing age, considerations other than income become increasingly important in determining the risk of death. One example of this is the so-called mortality cross-over in which the old in poor populations have lower rates of death than the old in relatively rich populations. A second example comes from a prospective cohort study in Zutphen, The Netherlands, in which men who were between 40 and 59 in 1960 were interviewed, examined, and then followed for 25 years (Duijkers et al. 1989). The men were grouped in four categories: (1) professionals, managers, and teachers; (2) small business owners; (3) non-manual workers; and (4) manual workers. Table 7 displays the all-cause mortality-risk ratios for each category using professionals, managers, and teachers as the reference group.

Table 7
Risk ratios of socioeconomic status on 25-year mortality stratified by age in a sample of men (N=856); Zutphen follow-up study

Socioeconomic status	Birth period			
	1900–1909		1910–1919	
	Risk ratio ^a	95% C.I. ^b	Risk ratio	95% C.I.
Small business owners	0.91	0.63–1.32	2.25	1.27–3.97
Non-manual workers	0.90	0.61–1.32	1.63	0.91–2.89
Manual workers	0.92	0.66–1.28	2.07	1.17–3.66

^aReference population is men in the highest social class: professional, managers and teachers.

^b95 per cent confidence interval.

Source: Duijkers et al. (1989) Table 2.

For the older men, those in their 50s in 1960, there was no difference in the risk of death from one occupational group to another. For the men who were in their 40s in 1960, there was increased risk among the owners of small businesses and among manual workers. The authors speculated

... that age is an increasingly important factor as one becomes older, even to such an extent that other factors like SES (socio-economic status) are no longer important. It is also

possible that the harmful effects related to SES play a more important role at a relatively young age. As a result the different socioeconomic groups become more and more alike as time goes by (Duijkers et al. 1989:661)

The point of this example is to suggest that the aging process itself may have consequences that over time become relatively more important than socioeconomic status. This is simply another case in which socioeconomic status is but one of several factors, and perhaps not the most significant, associated with increased or decreased risk of death.

Technologic

The association between income and mortality may be unstable because the technology of public and personal health has changed. In a cross-national study of gross national product *per capita* and life expectancy, Preston (1976) has shown that higher income is indeed correlated with greater expectation of life at birth in the 1930s just as in the 1960s, but that the shape of the relationship has changed.

The later curve appears to be steeper at incomes less than \$400 and flatter at incomes over \$600. In other words, for low income countries, a given increment in income tends to be associated with a larger gain in life expectancy in the 1960s than in the 1930s⁷ (Preston 1976:76).

And, he continues several pages later,

The relations ... appear to shift systematically during the twentieth century. In general, in order to attain a certain life expectancy between 40 and 60, a nation required an income level almost three times greater in the 1930s than in the 1960s ... The magnitude of the shifts, combined with regional income data, suggests that some 75-90 percent of the growth in life expectancy for the world as a whole over these three decades is attributable to factors exogenous to a nation's contemporary level of income (Preston 1976:83).

These exogenous factors seem to be a variety of easily exported public- and personal-health technologies.

Preston's data may be understood in two ways. First, over time at the same *per capita* income increasing levels of life expectancy have been achieved. And secondly, over time the same life expectancy can be achieved at lower *per capita* income. A similar result is observed if one considers life expectancy in Britain, Japan, and Germany in the late nineteenth and early twentieth centuries. Despite having per capita incomes considerably below those of Britain, Germany and Japan had earlier achieved comparable life expectancies. The mechanisms for this were, in Germany, public-health measures⁸, and in Japan, both public-health measures and differences in personal behaviour.⁹

These examples are meant to illustrate two points. One, that income may be less crucially related to life expectancy at some times than at others because new factors have emerged as significant in the interim. Among them have been new techniques for postponing death. Two, implicit in the comparison

⁷ See also Preston and Haines (1991).

⁸ See Evans (1987). Hamburg, being less willing to introduce public-health measures than other German cities, had a major cholera epidemic in 1892, when no other European city did. See also Brown (1988, 1989).

⁹ Hot tea, with boiling water, was safer than non-boiled water, the eating utensils differed for each individual, and there was a market for the collection of human waste for fertilizer. See Mosk and Johansson (1986); Hanley (1987); Jannetta (1987); and Johansson and Mosk (1987).

among Britain, Japan, and Germany is the trade-off between public spending and private income. Perhaps it was the greater British belief in *laissez-faire* individualism, contrasted with the elevated role of the state on the continent, that has led to a greater emphasis on individual-level factors such as real wages and *per capita* income in analyses of the changing pattern of mortality.

Economic

So far we have treated the notion of real wages as unproblematic. It is, however, highly problematic for a number of reasons. First, real wages can only be relevant in a society in which there are wages to be paid. Subsistence societies and those in which there are non-monetary exchanges may be status-ridden and divided by wealth of various sorts, but wages play no role. Moreover, status group membership may be unassociated, or inversely associated, with mortality. For example, in an as yet unpublished analysis of infant and child mortality among the matrilineal Hopi Indians of the village of Oraibi during the late nineteenth century, Levy (1990) has shown that women of prime and alternate lineages of high-status clans had both higher fertility and higher rates of child loss than other women. Prime lineages controlled most of the clan lands; alternate lineages stood ready to inherit should the prime lineage die out; other lineages, even of high-status clans, were marginal. High-status clans controlled the best lands. By the wealth criteria relevant to Hopi Indians of a century ago, the richest women had the highest rate of child loss. Levy has proposed that this is the result of attempts to assure the birth and survival of sufficient daughters to inherit the high quality land these high-status lineages controlled. The paradoxical result of high fertility was shortened birth intervals and crowding, both of which increased the risk of death.

Not only must there be wages for wages to have an effect, but the distribution of wages must be taken into account. For example, if there is a threshold effect for wages such that, once passed, very little additional mortality reduction can be expected, then if income increases throughout a society there could well be a changing relation between income and mortality. If everyone rises above the threshold, an income effect on mortality would no longer be observable. There is indeed some evidence for a threshold effect. Reviewing a study of infant mortality in Baltimore in 1915, Collins (1927) observed:

The infant mortality rate ... shows a consistent fall as income increases even above ... \$1,250. ...Data in the original report show infant mortality according to income in three classes above \$1,850, up to \$2,850 and over. But there seems to be little tendency to decrease after \$1,850. Additional income beyond a certain minimum does not seem to be associated with infant mortality (Collins 1927:50).

Put another way, unless the risk of premature death is lower among the super rich than the very rich, and the risk among the very rich is less than among the merely rich, then there must be threshold effects. No such gradient among the rich has ever been demonstrated of which we are aware.

So far we have discussed wages rather than income, for real wages have been the most frequently used measure in discussions of the standard of living and its relationship to mortality, and we have ignored the difficulties involved in estimating the real wage.¹⁰ The measurement of real wages requires estimates of nominal wages for some specific composite or subgroup of workers in some specific location, and some measure of their cost of living based on an appropriate set of prices for the particular basket of goods consumed at that particular location. Given different regional patterns of wages and of prices; the relation of age, experience, and life-cycle status to wage rates; and the existence of wage differentials by occupation and by skill level, it is not surprising that many different measures of

¹⁰ Much of the following material on the measurement of real wages is from Engerman (1990).

changes in real wages over time have appeared in the literature. Thus von Tunzelmann (1979) has provided a range of measured real-wage changes of between zero and 150 per cent for the period between 1750 and 1850.

There is also the vexing question of how real wages are to be related to even material standards of living. The nominal wage figure is for an employed individual, and could be hourly, daily, weekly, or monthly, depending upon the specific case, or it could be based on the application of a piece-wage system, directly related to output. To go from this wage rate to an estimate of annual earnings, and to family earnings, and then to a welfare equivalent requires, as has frequently been noted, a number of major adjustments. These include allowances for the extent of unemployment, for the labour effort or intensity demanded, for any disutility or utility of urban living, and for the disutility of working conditions, whether health-related or not. To go from real hourly wages to real annual earnings requires information on hours worked per day; the number of days worked over the course of the year; the valuation of any in-kind payments (particularly board and lodging); the role of other income sources, such as gifts, relief, and Poor Law payments; and aid from neighbours and kin. It is also important to know the extent to which workers held more than one job during the year, including time spent in cottage-type industries. And, to go from principal wage-earner to family incomes, it is necessary to understand the determinants of labour-force participation rates by age and gender, and the contributions to total family income of women and children.

Several additional significant problems of measurement must be considered, problems which could serve to obscure any relation between income and mortality that might exist. They affect attempts at verbal description as well as statistical analysis, since the usual analysis presumes a virtually simultaneous impact (within a short period of years) and a set of relations of a rather specific type. While alternative specifications and descriptions might be suggested, it remains uncertain as to the sensitivity of using aggregate data to capture these relations.

In some cases of disease and medical impacts, consequences occur many years after their initiating causes. For example, those survivors from cohorts that were aged 0-4 when an epidemic struck in the early nineteenth century in central Massachusetts had, many years later, lower mortality rates than did both older and younger non-stressed cohorts (Meindl and Swedlund 1977; Meindl 1982). Similarly, the study cited previously by Barker et al. (1989), indicated that the best predictor of adult heart attacks was birthweight (and, thus, maternal health). Thus the impact of the contemporary level of individual income might be offset by the importance of some influences, including incomes, much earlier in life. In addition, there may be a threshold level of income needed before which there will be an impact upon mortality or, as in the Preston studies and in the study of infant mortality described above, some level of income after which it will appear that there is no relation of income and mortality. Thus some attempts to find a continuous relation between income and mortality will be misleading; the appropriate set of measurements must allow for the existence of the particular threshold level.

There are also the difficulties posed when the full impact of any cause must allow for a cumulation of effects. No one year, by itself, may leave a dramatic effect on mortality. What may be required is the accumulation of causal factors over a period of years. Thus, for example, in a population of high enough income, one year's recession may not have any effects, whereas several years of decline will have. And, in earlier periods, the effect of a shortfall in any one year's harvest will depend upon the harvest conditions in prior years.

While statistical procedures can be devised to handle these problems, as well as such other problems as allowing for the magnitude of the secular trend, there remains some question as to whether these can fully capture the appropriate relation. In regard to the detrending of time-series data, the precise explanation for the trend behaviour (and thus for the determination of the most appropriate

statistical form) and the nature of the linking between the trend and cyclical patterns are important considerations.¹¹

Two problems raised by the foregoing discussion of real wages merit somewhat more elaboration: aggregation from local to national levels, and the role of the state not only in providing services but in redistributing both income and food in times of want. We shall deal with each briefly.

Aggregation. There is a long tradition of studies of the relationship between income, wages, and social class on the one hand and mortality and health status on the other, several of which we have already cited. Historically most of these have been cross-sectional and at the local level. The temptation, however, is often to extrapolate from these locally or temporally restricted data to the national level, with results that may well be misleading. For example, Hobsbawm (1957), in his now classic contribution to the standard-of-living debate, asserted that mortality fell from the 1780s to the 1820s and then rose until the 1840s. This assertion seems to have been based on local studies, for aggregate data of the national level, which were not available to Hobsbawm in 1957, do not bear it out.¹² This suggests that even if any relationship existed between wages and mortality among the working class featured in the standard-of-living debate, such a trend may not be generalized to the entire population. Local conditions were sufficiently diverse that one could not adequately extrapolate to the national level without very broad coverage indeed. In the absence of such coverage of both income and mortality, generalizations would be either impossible or questionable.

Government intervention. In his paper on the changing relationship between social class and life expectancy, Antonovsky (1967) suggested that the convergence of the life expectancy of the rich and poor in this century has been due largely to the emergence of welfare states. We suggest that the very emergence of the nation-state itself in the eighteenth century had the effect of dampening fluctuation in mortality from both epidemics and famine. This was the result of actions which only a reasonably effective bureaucracy could engage in successfully: quarantine, the control of armies on their territory, and the relief of local famines by shifting food from regions of surplus to those of want. Indeed, writing at the end of the eighteenth century, Malthus was one of the first to observe that mortality rates were declining, and to relate the decline to changes in personal and public behaviour.

I think it appears that in modern Europe the positive checks to population prevail less and the preventive checks more than in past times, and in the more uncivilized parts of the world.

War, the predominant check to the population of savage nations, has certainly abated even including the late unhappy revolutionary contests; and since the prevalence of a greater degree of personal cleanliness, of better modes of clearing and building towns, and of a more equable distribution of the products of the soil from improving knowledge in political economy, plagues, violent diseases, and famines have been certainly mitigated, and have become less frequent (Malthus 1960:Vol. 1:315).

He was not alone. Contemporary continental mercantilist writers such as J. P. Frank and successors such as Chadwick in Britain, Villerme in France and Virchow in Germany recognized that mortality was not exogenous but the result of social arrangements which could be modified by personal and public behaviour.¹³ Such behaviour must have diminished the sensitivity of mortality to

¹¹ See the debate discussed in Brenner (1977) and Kasl (1979).

¹² See Wrigley and Schofield (1981:414) and Deane and Cole (1962:127).

¹³ See, for example, Rosen (1958); Stangeland (1966); and Coleman (1982).

fluctuation in wages and income, though it is difficult to provide adequate historical data in support of such an assertion.

India at the start of the present century, however, provides an example of the impact of changing public policy on mortality. The development of a railroad network and the enhanced desire of the British to provide an effective relief system, including the use of the transport system to move grain into regions of potential famine, led to a dramatic decline in famine-related mortality, as well as declines in average death rates after the start of the twentieth century. Indeed, McAlpin (1983:218), argues that the one major famine in India after the first decade of the twentieth century, in 1943-44, can be attributed to the wartime destruction of the transport system and the shift in government concerns with World War II.¹⁴

Clearly, then, the ability and willingness of governments to intervene in what would otherwise have been calamitous conditions has attenuated the responsiveness of mortality to fluctuations in income. This has been true, not only in welfare states in the twentieth century, but in earlier centuries as well.

The point we have wanted to make in this section is that, broadly speaking, there are two classes of reasons why there is no secular association between mortality and real wages. The first has to do with the fact that mortality rates are the product of many different forces, not only wages, and it is a mistake to attempt to capture them all under the rubric of economic status. The second has to do with the fact that, even accepting the very real importance of wages, there are profound difficulties and heroic assumptions that must be made when reconstructing a wage series to relate to trends in mortality. This leads us to our final topic, which has to do with ways of thinking about wages and income as causes of death.

Thinking about the association between income and mortality

Considering this litany of problems, what is surprising is not that there is no very impressive temporal association between real wages and mortality, but that anyone ever expected there would be. Mortality rates and life expectancy are the final common pathways of numerous factors, including but not limited to wages and income. The fact that within any particular population at any particular moment the rich are likely to do better than the poor is not to be wondered at. The rich, after all, are better able than the poor to clothe, feed, and house themselves (although, also, whether out of perversity or of ignorance, sometimes better able to indulge in behaviour damaging to their health). This ability to better clothe, feed, and house themselves is why, within the appropriately chosen groups, income is generally but not always inversely related to risk of death, and why, all things considered it is better to be rich than poor.

For example, Table 6 shows a tendency for infant and child mortality rates to increase with decreasing economic status within the urban and rural populations (with an anomalous observation among rich rural residents), even though between urban and rural residents of the same classes mortality rates differ substantially. That the association between socioeconomic status and mortality is not invariably monotonic, however, is illustrated in Table 3. The highest class invariably has lower mortality than the lowest class, regardless of social network score. In the middle ranks, however, the pattern is far from clear, even when holding constant the social network score.

Given the facts that economic status is not always monotonically associated with mortality risk, and that numerous factors other than wages and income influence mortality, are there ways of usefully conceiving of the relationship between wealth and mortality? The expectation of a monotonic relationship between the two is analogous to the sometimes accurate but generally oversimplified

¹⁴ See also Sen (1981).

association between specific micro-organisms and the onset of specific diseases. Just as for some preventative and treatment purposes it is reasonable to consider only the micro-organism and its host, that is, when immunizing children or when treating certain infections with antibiotics, so for some purposes it is useful to consider only the association between wealth and mortality. On the other hand, just as truly adequate understanding of particular diseases requires an understanding of the full ecological setting, including sociocultural, microbiological, and pathophysiological factors, so does the full understanding of mortality more broadly involve consideration of income or wealth as but one of many factors which influence the risk of death in any particular setting.

Inevitably this kind of understanding involves local knowledge of a highly detailed and complex nature which is often at odds with a strong desire that many of us also feel to achieve a more universal, elegantly simple understanding at a very general level. Usually it is the latter kind of knowledge which is considered science. The former is thought of as too particularistic to deserve such a label. Without getting into the debate about what is and what is not science, it should be clear that both kinds of knowledge are of value. It is important and useful to be able to make generalizations across societies. It is equally important to be able to point to the many deviant cases that force us to question the validity of the generalizations we sometimes make with such facility, and often so facilely.

These observations suggest that the association between wages and mortality is not entirely similar to that between an essential nutrient and a deficiency disease. Which is to say, death is not caused by the lowering of wages in the same way that pellagra is caused by the absence of nicotinic acid. Stated so baldly, the assertion seems obvious. It is not. Implicitly that is the assumption underlying the question with which we began.

Conclusion

It seems general that, over the past several centuries, both within and among societies, the rich, on average, have had lower mortality rates than the poor. It appears that at some stage the differentials widened, but in the twentieth century there has tended to be some reduction without, however, yet achieving complete equality. Within this general pattern, however, there remain deviant cases. In some instances, a trade-off between the effects of individual income and social measures, such as public health, makes income a less than perfect predictor of mortality; in others individual differences in cultural and other aspects of behaviour mean that different mortality rates are experienced at similar levels of individual income. In addition the changing importance of different diseases—the shift from infectious to non-infectious diseases—would confound any relationship.

Clearly, for most nations sometime in the past three centuries, there have been increases in *per capita* income and reductions in the level of mortality. At a gross level, therefore, these changes are obviously associated. Nevertheless, the fluctuations in each do not follow a systematic pattern nor, for reasons discussed above, should we have expected that to have been the case. Since mortality has always been influenced by so many factors, including income, it would not be expected that any simple one-to-one relation within a limited time span could be found. Moreover, given the dramatic changes in the distribution of incomes, occupations, educational levels, and so on, that occur with the process of economic growth, any study at the level of national aggregates would mean combining groups with possibly quite divergent experiences.

Whatever the relation of income and mortality posited within and among nations cross-sectionally, the temporal association of national income and mortality rates has had many fewer adherents; and most of these seem to have been in the English-speaking world. More speculatively, it might be asked why this assumed relationship has been of more salience in Britain and its overseas extensions than it seems to have been elsewhere. Focusing on real wages and *per capita* incomes in the analysis of mortality reductions appears to reflect the same political culture that is associated with the developmental policies

that the English-speaking world has generally followed; that is, an emphasis on the growth of individual income as the most effective way to improve the quantity and quality of life. Historically, as suggested above, other societies have followed other paths to a similar demographic outcome.

References

- Antonovsky, A. 1967. Social class, life expectancy and overall mortality. *Milbank Memorial Fund Quarterly* 45:31–73.
- Barker, D.J.P. 1989. Rise and fall of western diseases. *Nature* 338:371–372.
- Barker, D.J.P. and C. Osmond. 1987. Inequalities in health in Britain: specific explanations in three Lancashire towns. *British Medical Journal* 294:749–752.
- Barker, D.J.P., C. Osmond, P.D. Winter, B. Margetts and S.J. Simmonds. 1989. Weight in infancy and death from ischaemic heart disease. *The Lancet* September 9:577–580.
- Berkman, L.F. and L. Breslow. 1983. *Health and Ways of Living: The Alameda County Study*. New York: Oxford University Press.
- Brenner, M.H. 1977. Health costs and benefits of economic policy. *International Journal of Health Services* 7:581–623.
- Brown, J.C. 1988. Public reform for private gain? Public health and sanitary infrastructure in German cities, 1877–1910. Paper presented at the annual meeting of the Social Science History Association, Chicago.
- Brown, J.C. 1989. Public health reform and the decline of urban mortality: the case of Germany, 1876–1912. Paper presented at the annual meeting of the Population Association of America, Baltimore.
- Caldwell, J.C. 1986. Routes to low mortality in poor countries. *Population and Development Review* 12:171–220.
- Coleman, W. 1982. *Death is a Social Disease: Public Health and Political Economy in Early Industrial France*. Madison: University of Wisconsin Press.
- Collins, S.D. 1927. *Economic Status and Health: A Review and Study of the Relevant Morbidity and Mortality Data*, Public Health Bulletin No. 165, September, 1926, Treasury Department, US Public Health Service. Washington DC: US Government Printing Office.
- Crafts, N.F.R. 1985. *British Economic Growth During the Industrial Revolution*. New York: Oxford University Press.
- Curtin, P.D. 1989. *Death by Migration: Europe's Encounter with the Tropical World in the Nineteenth Century*. Cambridge: Cambridge University Press.
- Deane, P. and W.A. Cole. 1962. *British Economic Growth, 1688-1959*. New York: Cambridge University Press.
- Duijkers, T.J., D. Kromhout, I.P. Spruit and G. Doornbos. 1989. Inter-mediating risk factors in the relation between socioeconomic status and 25-year mortality (The Zutphen Study). *International Journal of Epidemiology* 18:658–662.
- Engerman, S.L. 1990. Reflections on the 'standard of living debate': new arguments and evidence. Paper presented at the Conference on Capitalism and Social Progress (The Hartwell Symposium), University of Virginia, Charlottesville.
- Evans, R.J. 1987. *Death in Hamburg*. New York: Oxford University Press.
- Flinn, M. 1981. *The European Demographic System, 1500-1820*. Brighton: Harvester.
- Fogel, R.W. 1986. Nutrition and the decline in mortality since 1700: some preliminary findings. Pp. 439–527 in *Long-Term Factors in American Economic Growth*, eds. S.L. Engerman and R.E. Gallman. Chicago: University of Chicago Press.
- Gwatkin, D.R. 1980. Indications of change in developing country mortality trends: the end of an era? *Population and Development Review* 6:615–644.
- Hanley, S.B. 1987. Urban sanitation in preindustrial Japan. *Journal of Interdisciplinary History* 18:1–16.
- Hobsbawm, E.J. 1957. The British standard of living, 1790-1850. *Economic History Review* 10:45–68.
- Hollingsworth, T.H. 1964. The demography of the British peerage. *Population Studies* 18:supplement.
- Jannetta, A.B. 1987. *Epidemics and Mortality in Early Japan*. Princeton: Princeton University Press.

- Johansson, S.R. and C. Mosk. 1987. Exposure, resistance and life expectancy: disease and death during the economic development of Japan, 1900-1960. *Population Studies* 41:207-235.
- Kasl, S.V. 1979. Mortality and the business cycle: some questions about research strategies when utilizing macro-social and ecological data. *American Journal of Public Health* 69:784-788.
- Kunitz, S.J. 1983. Speculations on the European mortality decline. *Economic History Review* 36:349-364.
- Kunitz, S.J. 1987. Making a long story short: a note on men's height and mortality in Britain from the first through the nineteenth centuries. *Medical History* 31:269-280.
- Kuznets, S. 1966. *Modern Economic Growth: Rate, Structure, and Spread*. New Haven: Yale University Press.
- Levy, J.E. 1990. The demographic consequences of social stratification in an 'egalitarian' society. Paper presented at the annual meeting of the American Anthropological Association, New Orleans.
- Livi-Bacci, M. 1990. *Population and Nutrition: An Essay on European Demographic History*. New York: Cambridge University Press.
- Malthus, T.R. 1960 (first published 1803). *Essay on the Principle of Population*, 2nd ed., 2 vols. London: J.M. Dent.
- Matthews, R.C.O., C.H. Feinstein and J.C. Odling-Smee. 1982. *British Economic Growth, 1856-1973*. Stanford: Stanford University Press.
- McAlpin, M.B. 1983. *Subject to Famine: Food Crises and Economic Change in Western India, 1860-1920*. Princeton: Princeton University Press.
- Meindl, R.S. 1982. Components of longevity: developmental and genetic responses to differential childhood mortality. *Social Science and Medicine* 16:165-174.
- Meindl, R.S. and A.C. Swedlund. 1977. Secular trends in mortality in the Connecticut Valley, 1700-1850. *Human Biology* 49:389-414.
- Mosk, C. and S.R. Johansson. 1986. Income and mortality: evidence from modern Japan. *Population and Development Review* 12:415-440.
- Omran, A.R. 1971. The epidemiologic transition: a theory of the epidemiology of population change. *Milbank Memorial Fund Quarterly* 49:509-538.
- Pope, C. 1992. Adult mortality in America before 1900: a view from family histories. Pp. 267-296 in *Strategic Factors in Nineteenth-Century American Economic History*, eds. C. Goldin and H. Rockoff. Chicago: University of Chicago Press.
- Preston, S.H. 1976. *Mortality Patterns in National Populations: With Special Reference to Recorded Causes of Death*. New York: Academic Press.
- Preston, S.H. and M.R. Haines. 1991. *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Princeton: Princeton University Press.
- Rosen, G. 1958. *A History of Public Health*. New York: MD Publications.
- Schmelz, U.O. 1971. *Infant and Early Childhood Mortality among the Jews of the Diaspora*. Jerusalem: The Institute of Contemporary Jewry, The Hebrew University of Jerusalem.
- Sen, A. 1981. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford: Clarendon Press.
- Stangeland, C.E. 1966 (first published 1904). *Pre-Malthusian Doctrines of Population: A Study in the History of Economic Theory*. New York: Augustus Kelley.
- Stolnitz, G. 1955. A century of international mortality trends: I. *Population Studies* 9:24-55.
- United Nations, Department of Economic and Social Affairs. 1971. *The World Population Situation in 1970*. ST/SOA/SER.A/49. New York.
- United Nations, Department of Economic and Social Affairs. 1973. *The Determinants and Consequences of Population Trends* 2 vols. ST/SOA/SER.A/50. New York.
- Von Tunzelmann, G.N. 1979. Trends in real wages, 1750-1850, revisited. *Economic History Review* 32:33-49.
- Woodbury, R.M. 1926. *Infant Mortality and its Causes*. Baltimore: The Williams and Wilkins Company.
- World Bank 1990. *World Development Report, 1990*. New York: Oxford University Press.

Wrigley, E.A. and R.S. Schofield, 1981. *The Population History of Britain 1541-1871: A Reconstruction*. London: Edward Arnold.

Historical epidemiology and the structural analysis of mortality



John Landers

All Souls College, Oxford OX1 4AL, UK

Abstract

Attempts to explain long-term variations in pre-transitional Western European mortality in terms of changing living standards have met with little success, and this has led to the view that such variations were biologically, or climatically determined. This conclusion can, however, be avoided by a fuller specification of the determinants of exposure to infection that incorporates the dimensions of spatial structure. This paper advances a model of the proximate determinants of exposure and resistance to infection, and derives predictions for the mortality patterns of pre-transitional metropolitan centres that are tested against data from London c1670–1830. The latter generally bear out the predictions of the model whilst also demonstrating the importance of certain features of England's political economy over this period.

Demographers' understanding of long-term and much cross-sectional mortality variation has long been dominated by the concept of 'transition' with its implication that the present and future states of high-mortality populations can be fruitfully understood in terms of the past experience of populations currently experiencing much lower levels. Much of the theorizing previously carried out in this mould may now strike us as unacceptably simplistic, unilinear, or Eurocentric, but it remains true that the range of possible relationships between demographic variables and their determinants is not unlimited and that it can be usefully elucidated by historical studies.

Thus the past can further our understanding of the present and future by defining the limits of the possible, and the contours of the probable, rather than by defining necessary evolutionary sequences, and it may indeed be the very particularity of contemporary Third World experience which emerges most clearly in the light of European demographic history. This result in itself, however, would be a healthy corrective to the view that the latter is recapitulated necessarily in the former and can thus serve as a source of policies and prescriptions: a view owing more to ahistoric social-scientific preoccupations with developmental 'stages' than to truly historical concerns with the specificity of time and place.

The secular decline in mortality, the checks and interruptions that this has suffered, and the persistence of large-scale differentials are all bound up with movements in infectious disease mortality¹ (Arriaga 1989; Horiuchi 1989). Moreover, earlier expectations that mortality transition would yield regular monotonic declines in the latter have proved unfounded, and developed-world populations may even face an increase in infectious-disease deaths in parallel with further reductions in those from non-infectious causes (Coleman and Salt 1992:246–253). If these processes are to be elucidated it will first be necessary to understand the long-term dynamics of infectious-disease mortality itself.

In this context, data from pre-transitional Europe have an essential contribution to make, offering as they do a much greater time depth than that afforded by material from contemporary high-mortality

¹ This has, of course, ceased to be the case largely where differentials within the developed world are concerned, but it remains substantially true for those within Third World countries and between high- and low-mortality populations.

populations. Nonetheless, until recently historical demography's contribution has been hampered by conceptual inadequacies in the frame of reference adopted for the analysis of pre-transitional mortality. New developments in the field, particularly the rise of the approach termed 'historical epidemiology', promise, however, to resolve many of these. In this paper I shall review briefly the theoretical problems attending earlier analyses of pre-transitional mortality change and the conceptual framework which has emerged from recent work. I shall then outline one particular model generated by the latter and put it to an empirical test.

European historical demography and epidemiology

The dynamics of infectious-disease mortality² in pre-transitional Europe provide the historical demographic context of secular mortality decline. Only in such a context can we hope to isolate mortality changes stemming in fact from 'modernization' from those which reflect the intrinsic variability and instability of 'traditional' demographic regimes, and thus to distinguish historical contingency from the necessary interrelationships governing long-term social and economic change. In spite of this, however, most studies of mortality change in the relatively recent past have placed them in a greatly foreshortened historical context. Whilst it may be acknowledged that twentieth-century patterns are the outcome of historical processes, recognition of the latter is generally confined to the changes which have occurred since the eve of the mortality transition.

In the case of England, whose historical experience has been particularly influential in this respect, mortality patterns prevailing in the nineteenth century's central decades have been taken as a benchmark for the pre-transitional mortality regime and the point of origin for the transition itself, despite their historically specific character. In fact, the early Victorian pattern differed in important respects from that characterizing earlier periods—for instance in the structure of geographical and socioeconomic mortality differentials (Kunitz 1983, 1987; Landers 1986)—and appears to have been of relatively recent origin.³

This truncation of perspective should come as no great surprise, nor as any reproach to the demographers concerned. Studies of earlier periods have been pursued generally in a separate compartment from those of the immediate pre-transitional decades, let alone the transition itself; and the two have pursued quite distinct intellectual agendas with a large measure of incommensurability. This partly reflects varying methodologies, imposed by differences in the sources available before and after 1837, but it has been underpinned by fundamental interpretative and theoretical divergences.

The last century-and-a-half of mortality change in the developed world has been the subject of many competing accounts, but there is little disagreement that such change must be explained in terms of structural characteristics, social, economic or political,⁴ of the populations in question and of the

² Strictly speaking, we cannot separate the effects of infectious from those of non-infectious causes of death in historical populations. The former, however, were certainly responsible for the great majority of deaths before the present century, and it seems safe to assume that differences in overall mortality substantially reflected variations in the infectious component.

³ The influence of McKeown has, of course, been fundamental in this respect. By rejecting the evidential status of material dating from before the establishment of Civil Registration (in 1837), McKeown was constrained to take the mid-nineteenth century as the 'before' point for his series of 'before and after' comparisons (McKeown and Record 1962; McKeown 1976). Among other things this led him, as Szreter (1988) has pointed out, to understate greatly the contribution of smallpox to long-term mortality decline.

⁴ In this context, and throughout, we use the term 'structure' to refer to enduring sets of relationships characterizing populations, their economic, social and political arrangements, as well as their cultural life, as opposed to contingent or adventitious elements that may be present at any given time. This usage is familiar to social historians, and is

global system within which they are embedded. When we move back to the 'parish register' period of European historical demography, however, the situation is very different. Attempts to explain mortality change in terms of living standards, or real wages, have been generally unsuccessful, and interpretations based on the incidence and severity of mortality crises, themselves determined partly by economic and political variables, have also proved unconvincing (Schofield 1985; Lee 1986).

Under these circumstances the dominant 'neo-classical' theory of pre-industrial population dynamics looks to fertility, and in particular to nuptiality, as the variable articulating demographic and economic structures in Western Europe.⁵ Mortality change is now seen as an 'exogenous' variable, determined not by the level of real wages, but by some other, and more elusive, factor or factors. In practice, these have often been sought outside the realm of economy and society altogether, in microbiological or climatic changes (Perrenoud 1984, 1991), and the notion of exogenous mortality has thus been assimilated to the older concept of the 'autonomous death rate'; a variable whose determinants lie altogether outside the world of human agency (Chambers 1972).

Neo-classical theory has provided a remarkably successful account of social, economic and demographic interactions in pre-industrial populations, and it has fostered closer links between historical demography and cognate areas of economic and social history, but it has had problematic consequences for the study of mortality (Landers 1986). The failure of attempts to explain mortality variations in terms of real wages, and the widespread acceptance of the autonomous death rate, have left the determinants of mortality undefined and effectively removed from empirical enquiry.

This is unfortunate in itself but doubly so if we wish to place the transition in its deeper historical context. For as we progress through the early to the central decades of the nineteenth century we move abruptly from an interpretative framework in which mortality is biologically or climatically determined to one in which it is inextricably interwoven with the structures of economy and society. These and other difficulties with the concept of autonomous death rates have led a number of recent studies to examine a broader range of factors capable of influencing levels of infectious-disease mortality. These studies, which take as their goal the development of an historical epidemiology, have analysed possible structural determinants of mortality 'beyond the real wage', and particularly the effects of spatial structure on levels of exposure to infectious agents (Dobson 1989a, b; Landers 1991).

The structural model

The relationships postulated by studies of this kind can be elucidated usefully through a set of intermediate variables, or proximate determinants, intended to depict the pathways through which the ultimate determinants of mortality variation, be they economic, social, or biological, exercise their effects.⁶ We shall start by looking at the determinants of mortality at a given time t and then consider the question of change over time.

associated particularly with the work of Braudel (1980). It is however, broader than that usually employed by demographers and social scientists since it incorporates phenomena such as systems of belief as well as social and economic structures narrowly defined.

⁵ For a definition of neo-classical population theory, in the sense in which the term is used here, and an extended discussion of the theoretical issues raised, see Landers (1993) Chapter 1.

⁶ The concept of intermediate variables, or proximate determinants, has been employed extensively in fertility analysis where it has proved possible to develop a quantifiable frame of reference (e.g., Bongaarts 1978), unlike mortality analysis where the framework remains conceptual. Our system differs from that developed for present-day mortality analysis (Mosley and Chen 1984) in being pitched at the level of population variations rather than that of individuals or households.

Mortality at time t

At any given time the level of mortality prevailing in the population will reflect the balance between exposure to infectious agents and resistance to infection on the part of members of the population in question.⁷ These two variables are influenced by a variety of structural features of the population, but the latter interact with the character of the pathogens actually present in the population at any time: the population's pathogenic load.⁸ The quality of water supply for instance is of little relevance to mortality from bubonic plague. Nutritional status will have an important effect on the mortality of populations subjected to a high level of exposure to respiratory tuberculosis, but be of much less significance if plague or smallpox are major mortality factors (Lunn 1991). Hence we must distinguish between the population's structurally-determined resistance and exposure potentials and the realized levels which arise from the interactions of these with the pathogenic load.

The exposure variables

To begin with, it is useful to think of a human population and its environment as a network in which the points are individuals and the links are constituted by pathways allowing pathogens to move between individuals, and between them and any animal, or other, reservoirs of infection that may be present. It is important at this point to note that such pathways form a matrix with an existence analytically independent of the pathogenic load. The density of this matrix determines the population's potential level of exposure to infection which is realized to a varying degree through its interaction with the pathogenic load.⁹

It is ultimately determined by a wide range of factors including personal hygiene, population density, and many of the traditional 'public health' variables such as housing conditions, and the quality and quantity of water supplies. Where the youngest age-group is concerned methods of infant feeding are also of major importance. In the absence of a suitable term to denote this proximate determinant of the level of exposure to infection I shall refer to it as the degree of conduction present.

The resistance variables

A population's potential resistance to infection can be considered as a joint function of nutritional and immunological status, whilst realized resistance again depends on the interaction of these with the pathogenic load. Nutritional status itself also reflects the interaction of diet and morbidity.¹⁰

In this context the term **diet** refers to nutrients actually consumed and thus reflects methods of food preparation and storage as well as the quantity and quality of food available to the population.

⁷ For reasons explained above we are restricting attention to the determinants of infectious-disease mortality. In principle the framework developed here could be generalized to include non-infectious causes by a widening of the concept of 'exposure' and allowing for 'negative resistance', that is, propensity to develop disease in the absence of exposure.

⁸ The experience of subgroups of the population, defined geographically, socioeconomically, or in some other way, may have differed substantially in these respects. Under these circumstances the population's epidemiological regime is said to be 'segmented' (see Landers 1993, Chapter 1).

⁹ The most striking example of this is provided by pre-Columbian America. Here the high population densities attained in some regions created a high level of conduction, and thus exposure potential, but realized exposure apparently remained very low until the coming of Europeans transformed the pathogenic load.

¹⁰ A further determinant of individual nutritional status is the energy demand imposed by the work load, but it seems unlikely that this was sufficiently variable over time, or between populations (though it may well have varied socioeconomically within populations) to constitute a significant determinant of mortality. It would, however, be easy to include this effect, as well as those of other sources of stress such as temperature, by incorporating the intermediate variable stress as a further proximate determinant of resistance potential.

Morbidity affects nutritional status, and thus resistance to infection, through the energy demands it places on the organism and the longer-term effect that certain types of infection, particularly gastrointestinal, may have on the ability of the organism to absorb nutrients from the diets. Current resistance potential is thus affected by both past and present levels of morbidity.

Immunological status depends on a combination of nutritional status and the past history of exposure to infectious agents, whether by 'natural' infection or by artificial immunization. For the youngest age-group it is also affected by the methods of infant feeding employed.

Change over time

The proximate determination of mortality at a given time t is evidently highly complex. A number of the proximate determinants interact with each other, or appear more than once, and the role of morbidity means that the state of the system at t is not independent of its state at earlier points. Fortunately, these complexities are simplified when we consider change over time, for a number of the variables are endogenously determined and so drop out of the analysis.

Resistance variables

Of the resistance variables, the level of morbidity is, as we have seen, set internally to the system. This is also true of changes in immunological status which, in the absence of migration,¹¹ are determined by variations in pathogenic load and in nutritional status. Nutritional status reflects the interaction of diet and morbidity, and since morbidity is also endogenously determined only diet remains as the principal proximate determinant of changes on the resistance side of the balance.

Exposure variables

Realized exposure at a given moment depends on conduction and pathogenic load. Changes in conduction will thus affect levels of exposure and can occur through variations in any of the social, economic or ecological variables to which this variable refers.

Pathogenic load

Pathogenic load, as we have seen, affects the realized levels of both exposure and resistance to infection. The determinants of changes in this variable are more complex than those considered so far and reflect the interaction of two distinct sets of factors. It is evident that the pathogenic load will vary according to the ability of pathogens to remain in the population and the introduction of new pathogens. The first of these is determined by a variety of factors including population size and density, and the existence of animal or other reservoirs of infection, and it can be defined as a further proximate determinant of mortality which I shall term the degree of retention.

The introduction of new pathogens can occur either by their physical entry from the outside world—by means of trade, migration, warfare, or other forms of movement—or biologically, through genetic or chromosomal alterations to the existing stock. The latter represent truly 'autonomous' phenomena, but the former depend in part on the degree of physical movement between the population and the outside world. This depends on a variety of geographical, economic, and political variables and can be referred to as the population's degree of bounding.

¹¹ Migration will affect immunological status, at the population level, if the population from which immigrants are drawn differs in this respect from the host population in question. This appears to have been the case where early modern European cities are concerned.

Epidemiological regimes

Change over time thus reflects the effects of variations in resistance and exposure potentials, governed by the four proximate determinants diet, conduction, bounding, and retention. Changes in realized resistance and exposure depend on these but also on two factors outside the system which cause truly 'exogenous' variations in pathogenic load. The first of these includes biological events such as gene mutations, whilst the second refers to the character of the pathogenic load in the 'outside world' populations with which the regime is in contact.

Epidemiological regimes¹² can thus be characterized in terms of varying levels of the proximate determinants of mortality. Of the former, particular interest attaches to those characterizing the populations of large metropolitan centres in the pre-transitional era.¹³ Here high population densities combined with poor housing, sanitation, and water supplies to produce high levels of conduction and retention, whilst at the same time the nodal position of such centres in networks of trade and migration reduced the degree of bounding to a low level. The exposure potential of such regimes was thus high, and they appear to have functioned as endemic reservoirs of infection.

Mortality was correspondingly severe, but individuals who survived for any length of time acquired a degree of immunological resistance to a range of pathogens. Hence at any given time the level of immunological resistance characterizing such populations was likely to be substantially higher than that found in more thinly settled 'hinterland' regions where the degree of retention was much lower. The high exposure potential of such regimes was thus partly offset by a high resistance potential, and in many instances the latter was further enhanced by the economic and political centrality enjoyed by metropolitan centres which provided a food base that was more secure, and possibly also more abundant, than those of smaller settlements.

This configuration of proximate determinants can thus be described as a 'high potential' epidemiological regime. According to our model such a regime should manifest itself in a characteristic mortality pattern. Mortality would be severe, concentrated among children and recent immigrants who lacked previous exposure to urban infections, but levels should be relatively stable from year to year since the scope for epidemic outbreaks was limited by the high level of endemic infection, and major subsistence crises would be rare. In the remainder of this paper we shall see how far these predictions are borne out in a particular critical instance.

The case of London

The experience of London over the 'long eighteenth century' provides an important test case for the predictions of the high potential model of metropolitan epidemiological regimes. In 1700 London numbered around half a million inhabitants, some ten per cent of England's population at large and more than half of its urban population (Wrigley 1967). More than ten times the size of the largest provincial centre, London was an indisputable 'primate city'. The demands of the London market had played a crucial role in integrating the national economy over the previous century or more. It was

¹² The concept of the epidemiological regime is defined by analogy with that of the vital, or demographic, regime. It refers to the persisting structural relationships which govern levels of both morbidity and mortality in a population over time, although in practice the former are generally hidden from observation.

¹³ This is primarily because of the substantial burial surpluses characterizing such centres, which have long been a focus of attention among economic and demographic historians (Sharlin 1978; DeVries 1984; Landers 1987). The particular model outlined here is based on suggestions advanced by McNeill (1980) and is considered in greater detail in Landers (1990).

England's judicial and administrative centre and the hub of a nationwide system of long-range migration.¹⁴

In the course of the eighteenth century, England's urban hierarchy became more sophisticated with the development of a tier of large provincial cities and a range of new functional urban types (De Vries 1984). London's population in 1800 was, relative to that of England, no larger than it had been a hundred years earlier, but in absolute terms it had nearly doubled. What makes this growth the more remarkable is that it was accomplished in the face of an excess of recorded burials over baptisms every year throughout the century until the 1790s. Under these circumstances the city's continued existence, let alone its expansion, depended on a continuing stream of immigrants from the countryside and smaller urban settlements.

The study of London's demographic history over this period is greatly assisted by the existence of a remarkable source of aggregated vital data. These are the London *Bills of Mortality*, sets of burial totals broken down by cause of death, the parish in which they occurred and, from 1728, by age. This information was published every week, together with a global baptism total, and a consolidated annual Bill was issued at the end of each year.¹⁵ The Bills suffer from problems of under-registration, particularly where baptisms are concerned, and caution is required in their use, but for all their shortcomings they represent a source of demographic data on a scale unparalleled elsewhere before the beginning of the statistical era.

Tables 1 and 2 give the breakdown of the global burial totals by cause of death and by age for a series of subperiods. The cause-of-death terminology presents obvious difficulties, and only in the case of smallpox can we make an unequivocal identification with a category known to twentieth-century medicine. The category of 'fevers' was retained in the early decades of civil registration, after 1837, before being split into typhus and typhoid. It is likely that these two causes, particularly the former, predominated among fever deaths in the eighteenth century although many other infections were doubtless represented in addition. 'Diseases incidental to infancy' (henceforth simply 'infancy') groups together a number of causes of death and the total is roughly equivalent to that for burials under the age of two years.

Table 1
Contribution(%) of causes of death to total burials

Cause of death	1675–99	1700–24	1725–49	1750–74	1775–99	1800–24
Consumption	16.7	12.6	15.7	19.5	24.5	23.2
Fevers	15.2	14.9	15.3	13.9	11.3	6.9
Infancy	29.9	39.1	37.6	35.0	32.2	29.2
Smallpox	6.4	7.3	7.8	10.0	9.3	5.4
Other causes	31.8	26.1	23.6	21.5	22.7	35.3

Table 2
Age distribution of total burials

¹⁴ See Landers (1993), Chapter 2, for a general review of the economic and social character of London over this period and its relationship to that of England as a whole.

¹⁵ Marshall (1831) provides data from the annual Bills in a conveniently tabulated form and was used as the source for the analyses which follow.

Age group	1728–49	1750–74	1775–99	1800–24
0–1	36.3	33.6	33.2	28.4
2–4	8.6	9.1	10.1	10.6
5–9	3.6	3.6	3.8	4.2
10–19	3.1	3.4	3.5	3.3
20–29	7.8	8.0	7.2	6.9
30–39	9.5	9.3	9.0	9.4
40–49	9.8	9.7	9.9	10.3
50–59	7.9	7.9	8.5	9.2
60–69	6.2	8.1	7.2	8.1
70–	7.2	7.3	7.6	9.6

It is probable that most deaths from respiratory tuberculosis appear under the heading of ‘consumption’, but the latter term seems to have been used as a ‘catch-all’ for a variety of emaciating conditions and cannot be equated with this disease alone (Hardy 1988a). Table 1 shows the importance of infancy deaths and the two most ill-defined categories, consumption and ‘all others’ in the overall burial totals, but where change over time is concerned it is fever and smallpox, together with infancy, which contribute most to the decline in numbers at the end of the period.

Mortality levels and their stability

The age-specific burial totals underline the scale of mortality in the youngest age-groups and the eventual reduction in infant mortality. The distribution of burial totals by age is sufficient to give us an impression of the corresponding mortality risks, but for a more rigorous investigation we require a set of age-specific mortality probabilities (the life-table function q_x). These cannot be obtained directly from the material in the London Bills, since the latter contain no information on the underlying population at risk, but they can be calculated from nominal vital registers using the technique of family reconstitution. The rates in Table 3 were obtained from the registers of two of London’s six Quaker Monthly Meetings,¹⁶ and a comparison between these and similar figures obtained from reconstitution studies of parish registers elsewhere in England reveals the extent of London’s excess mortality in infancy and childhood.¹⁷

Table 3
Age-specific mortality rates (q_x) per thousand in 13 English parishes and two London Quaker Meetings

Cohort	Age-group	English parishes	London Quakers
1650–99	0	170	260
	1–4	101	244
	5–9	40	67

¹⁶ See Vann and Eversley (1992) for a general discussion of the registration system together with the Quakers’ social, economic and demographic characteristics. Data on the composition and location of the two meetings used in this analysis are provided in Landers (1991).

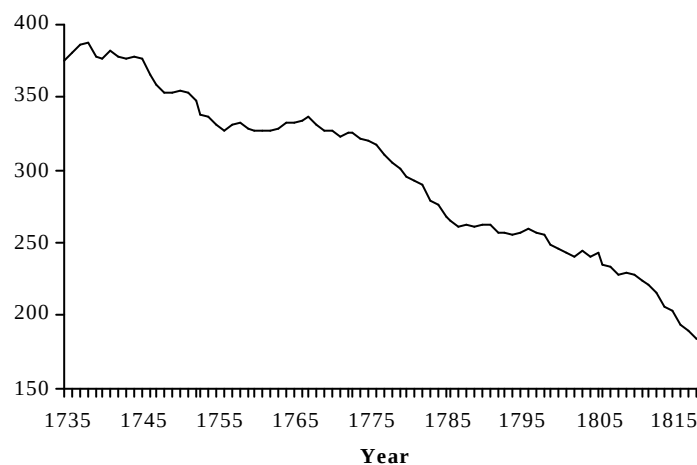
¹⁷ It is unfortunate that the methodology of family reconstitution does not lend itself at all well to the measurement of adult mortality (Wrigley 1968; Wrigley and Schofield 1983), and relatively little is known about its level in pre-industrial England. The limited information available, however, suggests that the differential between London and other English communities was much smaller than was the case with infant and child mortality (Landers 1990).

	1-9	137	295
	0-9	284	478
1700-49	0	195	342
	1-4	107	298
	5-9	41	95
	1-9	143	365
	0-9	310	582
1750-99	0	165	276
	1-4	103	253
	5-9	33	57
	1-9	133	296
	0-9	277	490

Source: Wrigley and Schofield (1983) and Reconstitution Tabulations.

As we have seen, the Bills of Mortality do not lend themselves easily to the analysis of age-specific mortality levels, but from 1728 at least, it is possible to construct a series of infant-mortality estimates on the basis of assumption about the under-registration of baptisms relative to burials and the distribution of ages at burial in the youngest of the Bills' age-groups (0-1 year).¹⁸ The results in Figure 1 are reassuringly close to those obtained from the Quaker records and suggest that, before the late eighteenth-century decline in mortality, over a third of London-born infants died before their first birthday.

Figure 1
Infant mortality rate in London Bills (11 point moving mean)



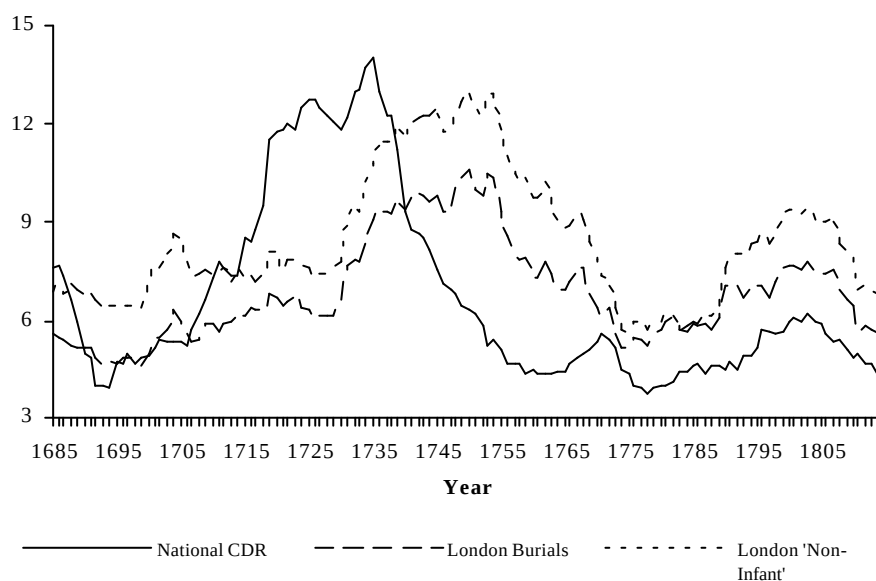
Metropolitan mortality levels, according to the high potential model, not only should be higher than those of hinterland populations, they should also display a greater short-run stability. This expectation can be tested quite easily using the Bills' burial series. The first step in the procedure is to remove the effects of medium- and long-term trends from the series. This 'de-trending' can be done in a number of ways, and I have employed the relatively simple method of dividing each annual burial

¹⁸ The major problem involved in this process is to estimate the shortfall in the registration of baptisms relative to that of burials. In order to do this a set of correction factors was constructed based on those used by Wrigley and Schofield (1981). For details see Landers (1993) Chapter 5.

total by the mean figure for the adjacent quinquennia and multiplying the result by a hundred. The resulting 'crisis mortality ratio' (CMR) thus measures the departure of each year's crude death rate from its short-run average, and by taking the mean absolute deviation (MAD) of the CMR series from its mean we have a convenient measure of the short-run stability of mortality.

Figure 2 plots 21-year moving averages for MADs calculated from the London Bills series (both including and excluding infancy deaths) together with a comparable series constructed from Wrigley and Schofield's estimates of the national crude death rate over this period (Wrigley and Schofield 1981). Other things being equal, the short-run stability of a mortality series should rise with the size of the population concerned, whilst for technical reasons CMRs constructed from crude death rates (CDR) should be less volatile than those derived from the corresponding burial totals being unaffected by short-run movements in population size. It is thus noteworthy that the MADs for the London burial series are actually lower than the national CDR for much of the high-mortality period and at comparable levels throughout.

Figure 2
Short-run instability of mortality



An examination of extreme CMR values leads to similar conclusions. In Figure 3a I have plotted values calculated from the national and London series, together with a further set based on the Bills of Mortality but excluding infancy deaths, for the years in which the national series registers its 15 highest values. In Figure 3b I have done the same for the years which fall into the 'top 15' in either of our London series without doing so nationally. The designation of 'mortality crises' has an inevitably arbitrary character in situations of this kind, but Flinn (1974) has suggested a CMR of 130 as the appropriate criterion when working on a regional scale. If we adopt this, for purposes of demonstration, we find that there were no 'crises' in London at any time during our period, using either of our two

series, although the threshold is passed once in the national series. The level of 120 is reached on only six occasions in the 'non-infant' series, and three times in the total burials. The national series also contains six values above this level but passes the 40 per cent mark in the crisis year of 1729.

Figure 3a
Crisis mortality ratios extreme values: national criterion

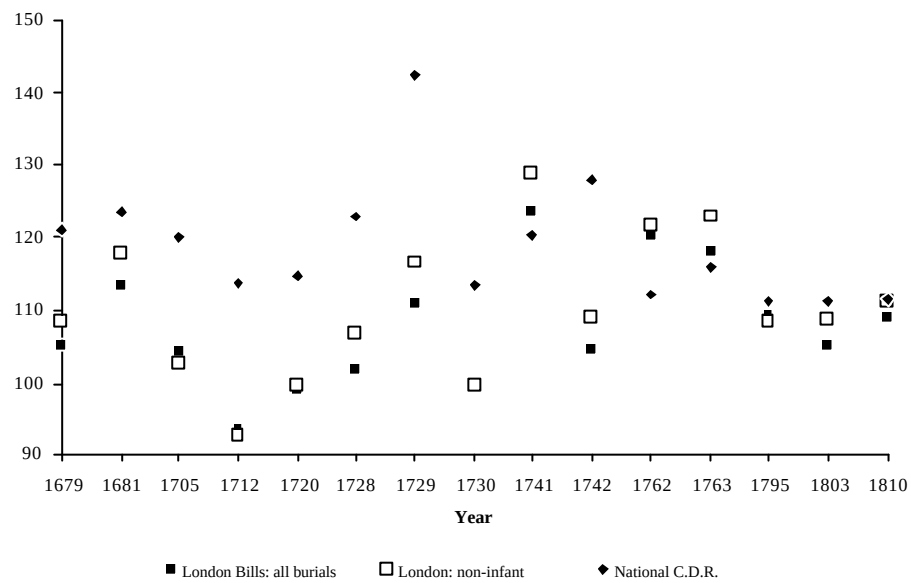
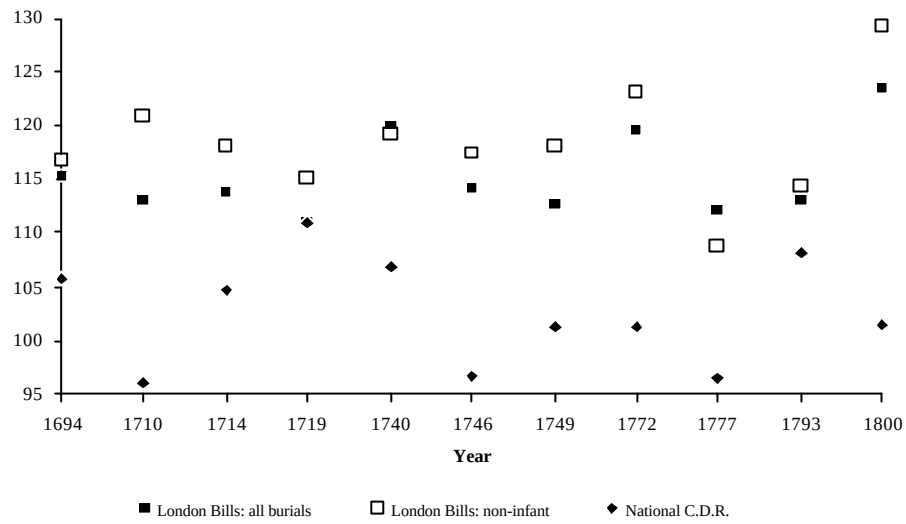


Figure 3b
Extreme values: London criterion

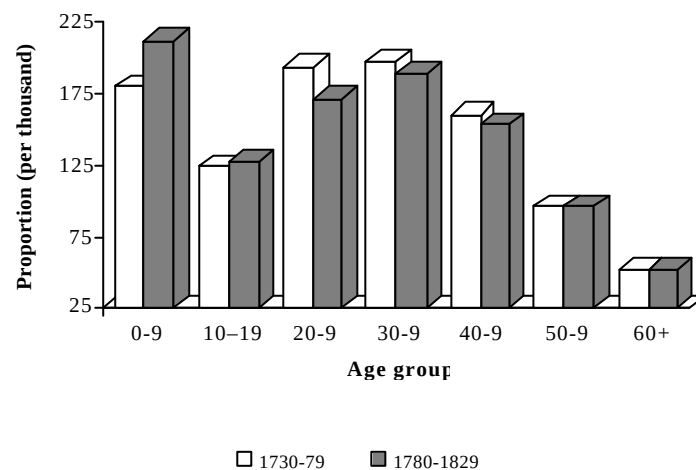


It is interesting to note that London's 'advantage' appears to have been most marked in the earlier part of the period when absolute levels of mortality were most severe. Of the 15 worst years in the national series, eight occur before 1740, but only in 1681 and 1729 does the London CMR pass the 110 mark. Similarly, only four of the years excluded on the national criterion, for this part of our period, feature among the 15 worst years in London. Furthermore, in two of these cases, 1710 and 1719, the divergence is probably less marked than appears at first sight, since both prefigure major upswings in the national series and the 1719 CMR for the latter is, in any case, of comparable magnitude to the London non-infant figure. The position after the mortality wave of the early 1740s is, however, very different. The London CMRs are now close to, or above, the national figure in each of the remaining five national crises, and six major upswings which are included on the national criterion.

Migration

The scale of London's burial surplus combined with the city's secular growth implies a substantial net inflow of migrants. Although the data do not allow us to estimate the absolute numbers involved to any useful degree of precision, it is possible to obtain a rough approximation of the age-structure of net immigration using the age-specific burial series. In order to do this we must first obtain a set of age-specific death rates (life table m_x) which are then divided into the burial totals so as to obtain estimates of the underlying population size. Death rates of this kind cannot be obtained directly from the Bills, but they can be estimated indirectly using model life tables together with the infant mortality estimates obtained earlier.¹⁹

Figure 4
Estimated age-structure for London



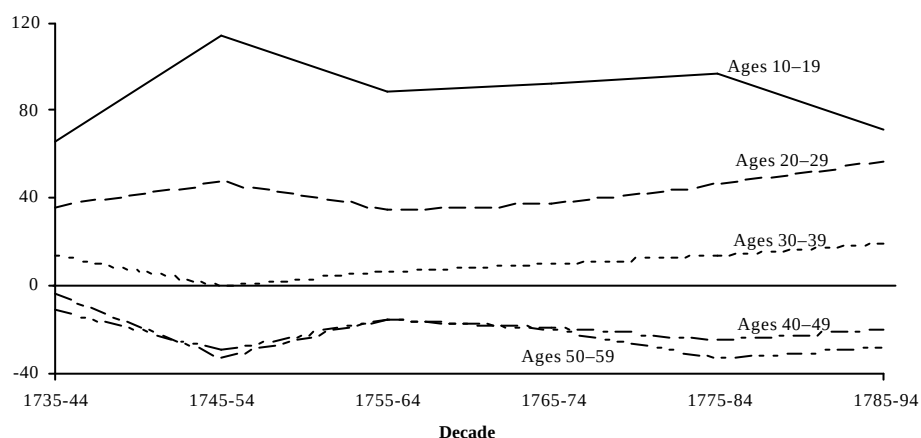
Applying these estimates to the burial totals, decade by decade, yields the results in Figure 4 in which the data for the decades of high mortality and mortality decline have each been pooled. The relative predominance of young adults in London's population emerges clearly from Figure 4 as does

¹⁹ The method used was based on the Brass two-parameter model life-table system (Brass 1971) and is discussed in detail in Landers (1993), Chapter 5.

the increase in the proportion of children as mortality declined. The uncertainties and approximations of the method are such that the detail of the results will not bear too much scrutiny, but these two general observations are both fairly robust and follow from the very distinctive age profile of the burial totals.

The age-structure of net migration can now be estimated by comparing the population total for each ten-year age-group with that for the next youngest age-group in the preceding decade.²⁰ The results in Figure 5 demonstrate the importance of adolescents and young adults in net immigration whilst suggesting an overall outflow of older adults. Paradoxically these figures are probably more reliable than are the decadal age-structure estimates from which they were derived because any systematic errors in the latter will be, at least in part, self-cancelling when the results for successive decades are compared.²¹

Figure 5
Net immigration by age (as percentage of total net immigration)



The high potential model predicts that, as a group, recent immigrants should show different immunological and epidemiological characteristics from native-borne adults, reflecting their limited exposure to urban infections. The Bills themselves do not provide a double-classification of deaths by age and cause, and so we cannot approach the problem directly using this material. The Quaker registers, however, use the same cause of death and age labels as the Bills and, since they identify individuals, allow us to construct a classification of this kind. The results in Figure 6 are based on an analysis of this kind carried out on smallpox burials. They reveal that a substantial minority of all

²⁰ Population estimates by age were calculated for each decade and were assumed to be valid for the midpoint of that decade. Net migration totals were then calculated by comparing the inter-decadal changes with the corrected age-specific burial totals for the intervening period.

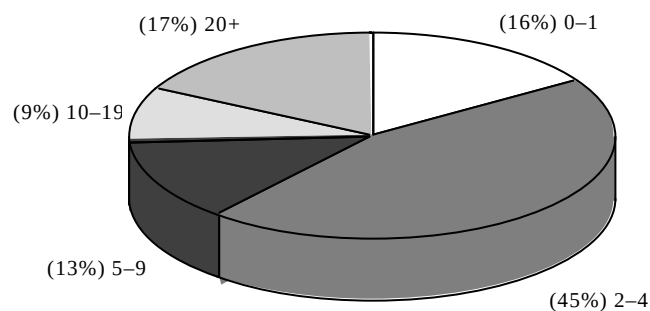
²¹ The most likely source of error in the age-structure estimates is an underestimate of mortality in early adulthood relative to childhood. This is because the model system, being based on more recent experience, does not take account of what are believed to have been the peculiar risks experienced by immigrants in this age-group to early modern cities. If this is so, the corresponding population estimates should be somewhat too high before the onset of secular mortality decline, but the discrepancy should be consistent from decade to decade.

deaths occurred among young adults and adolescents, despite the fact that smallpox was apparently a universal disease of childhood in London at this time.

The implication of this finding is that a significant proportion of the immigrants to London at this time had not been previously exposed to smallpox and were thus immunologically unprotected, as were the London-born children. We can pursue this question indirectly, with the aid of the material from the London Bills, by examining the relationship between movements in the age-specific CMR series. Figure 7 displays the principal components of the inter-correlation matrix for the latter.²²

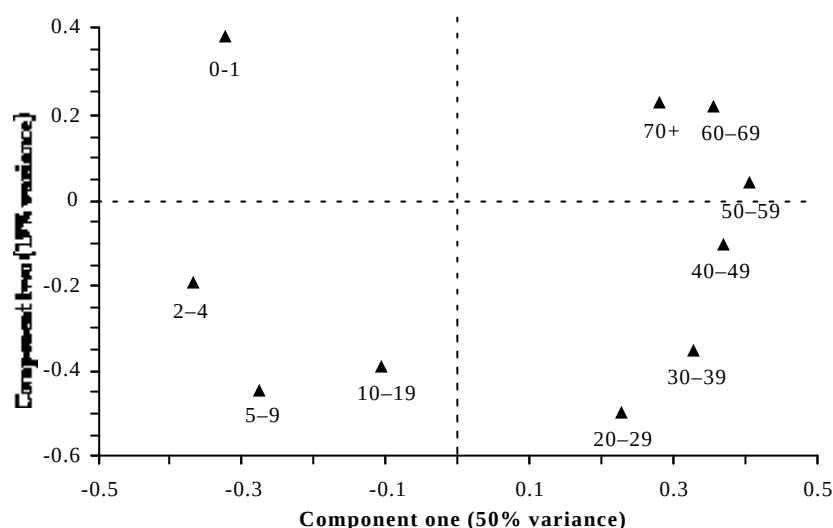
Figure 6

Age distribution of smallpox burials in two London Quaker Meetings 1700–49



The first of these accounts for half of the total variance and divides the age-groups fairly clearly into two clusters corresponding to the adult and childhood-teenage years respectively. The second component, however, produces a different grouping, distinguishing the age-groups 5-39 years from those aged 0-1 and 60+, with the 2-4 and 40-59 year olds having intermediate values. One interpretation of this finding is that the second component is picking out a pattern of disease susceptibilities common to children and recent immigrants, but distinct from those imposed by the physical frailties characterizing the oldest and youngest age-groups.

²² In order to remove the effects of shifts in the overall level of mortality from this analysis the raw age-specific annual burial figures were divided by the global burial total for that year prior to the calculation of CMRs.

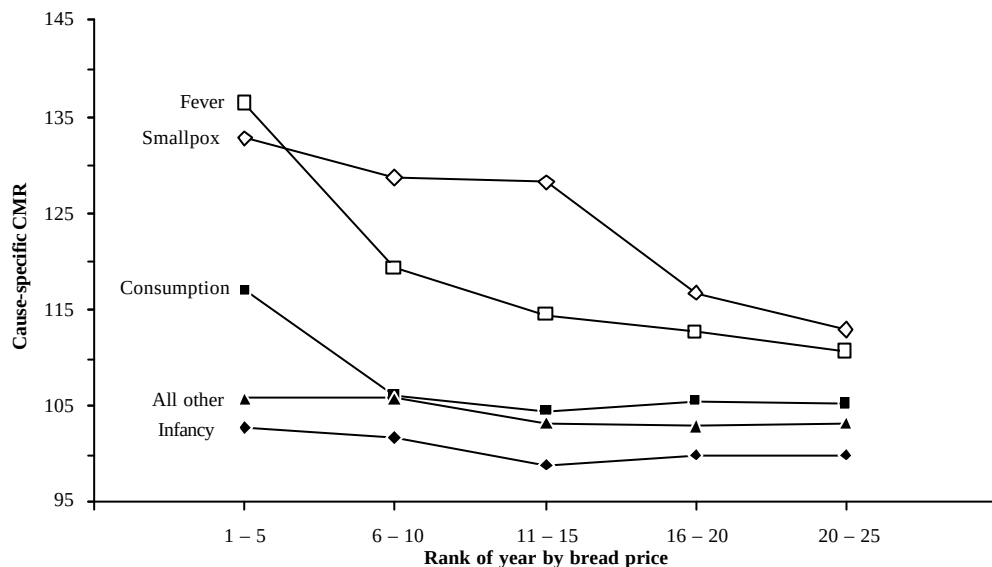
Figure 7**Short-run movements in burials by age 1733–1824: principal components analysis****Mortality fluctuations**

We have already seen that short-term mortality fluctuations in London were of relatively modest amplitude and that there appear to have been no real ‘mortality crises’ at any time during our period. There were, however, a number of marked upswings in mortality, and it is of some interest to examine the circumstances surrounding these. Orthodox ‘crisis theory’ distinguishes three factors capable of producing disturbances of this kind: price rises linked to crises of subsistence, outbreaks of epidemic disease, and war or civil unrest. We shall begin by examining the effects of movements in the price of bread on those in the burial series.

There are a number of ways in which this can be done, including some sophisticated econometric techniques, but I have adopted a relatively simple method. First the detrended annual burial totals, and the average annual price for a quartern loaf, were ranked in order of magnitude. The years having the top 1–5, 6–10, 11–15 etc. bread prices were then identified and the mean detrended burial total calculated in each case. Figure 8 displays the results of this analysis for the cause-specific burial series. These suggest that mortality in general was little affected by high bread prices but that these had a major impact on mortality from smallpox and, to a lesser extent, from fever.

One shortcoming of this analysis is that it can only detect effects which occur in the same year as the price rise itself, whereas it is likely that much of the impact will be ‘lagged’ by one or more years. An elementary means of examining this possibility is to repeat the analysis taking the detrended burial total for the year following the price rise should this exceed the value for the year in question. The results in Figure 9 reveal that this exercise strengthens the relationship substantially in the case of the fever and smallpox burial series, whilst having little effect on the others.

Figure 8
The effect of high bread prices on mortality in the London Bills



This finding is at first sight surprising, since the most straightforward interpretation of such associations is nutritional. On this basis higher prices would lead to a reduction in the quantity and quality of food intake, and in resistance to infection, on the part of the members of the population most affected. Smallpox, however, was an infection whose outcome was very little affected by the nutritional status of the victim, and nutritional status also seems to have been of secondary importance at most in the case of typhus, the disease likely to have been responsible for the bulk of the excess fever deaths.

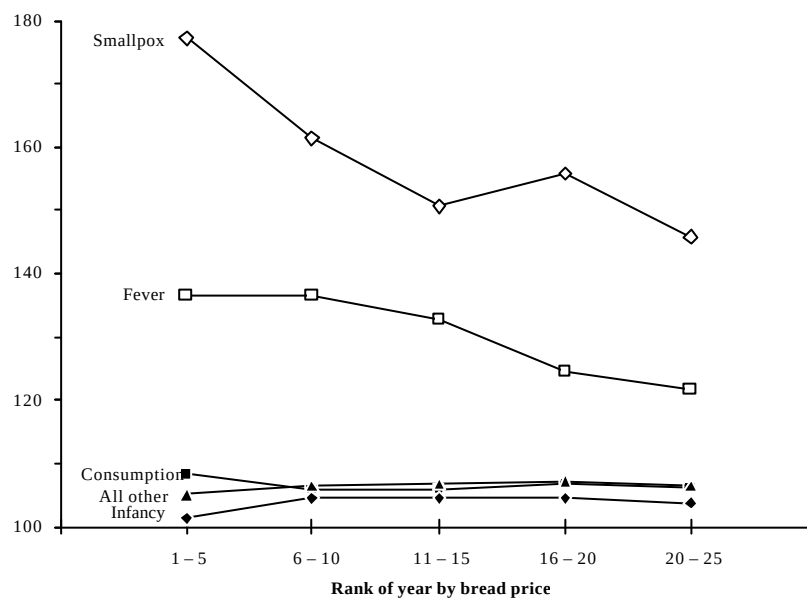
In this case an explanation must be sought in the variables governing exposure to infection; in particular, bounding and conduction. It is likely that high food prices in London generally coincided with periods of economic stress at least as severe in the country at large and that migration to London was one response to such circumstances. Many such migrants would, as we have seen, lack resistance to smallpox and in all probability to other urban diseases as well. Once in the capital they congregated in the suburban districts with their filthy and overcrowded lodging houses, further exacerbating an already chronic shortage of accommodation.

Under these circumstances the overall immunological resistance of the population was reduced by compositional change. Many new migrants were wretchedly housed, and their swelling numbers meant that the existing population experienced increased competition for the inadequate supply of accommodation, a factor of particular importance where economic disruption reduced employment opportunities in the capital for migrant and native alike.²³ It is this combination of increased movement

²³ For an analysis of the relationship between price shocks, social dislocation and increased mortality in pre-industrial Europe see Post (1985). Hardy (1988b) has demonstrated an analogous relationship between economic dislocation and increased housing densities, giving rise to epidemic disease, in Victorian London.

and effective population density under conditions of indifferent hygiene which can best explain the rise in smallpox and fever mortality.

Figure 9
The effect of bread prices on mortality (including one year lags)



The effect of this broader social and economic disruption can be seen in the mortality waves accompanying the transition from war to peace over this period. England's seventeenth century experience, under republic and Stuart monarchy alike, had endowed its political class with a hostility both to standing armies, and to the fiscal apparatus required to maintain them; so the cessation of hostilities was followed by the demobilization of tens, and later hundreds, of thousands of servicemen with no effective measures taken to reintegrate them into peacetime society. Many ended up on the roads of England, most of which led to London.

Postwar London was, however, ill-prepared to receive such an influx. Its labour market usually suffered badly from the transition between war- and peace-time economies, whilst the housing stock was adversely affected by wartime depression in the construction sector. The effects can be seen in Table 4 which sets out the detrended burial totals for fever, smallpox, and all causes, together with a comparable figure for bread prices and absolute numbers for the size of the armed forces. Only at the very beginning and end of the period, in 1698 and again in 1816, does the coming of peace coincide with a major price rise, but a mortality wave is apparent on each occasion, the only exception being the close of the American war. In the early part of the period it is apparently smallpox deaths which are the most sensitive to these conditions. This ceases to be the case from the last quarter of the eighteenth century, but fever deaths remain as responsive as ever down to the close of the Napoleonic wars. Fewer deaths, however, now constitute such a small proportion of the overall burial totals that these are largely unaffected.

Table 4
The consequences of peace

Year	Bread price	Burials (as % of trend)			Armed forces (thousands)
		Fever	Smallpox	Total	
1697	116	94	60	103	87
1698	124	102	209	99	36
1699	110	113	90	101	13
1712	97	90	108	93	153
1713	98	87	85	91	46
1714	100	135	145	114	29
1748	102	116	96	103	105
1749	102	135	145	113	75
1750	99	128	62	103	39
1762	87	122	115	120	151
1763	90	106	157	118	136
1764	103	121	97	102	48
1783	112	94	89	97	234
1784	110	80	99	90	57
1785	94	95	119	97	48
1802	87	120	115	104	274
1803	76	130	82	105	194
1815	76	126	85	102	326
1816	135	125	83	108	257
1817	108	123	143	105	203

Sources: For armed forces figures: Ashton (1959); Floud et al. (1990).
For bread prices: Mitchell and Deane (1962).

The seasonality of mortality

According to the high potential model, metropolitan centres function as endemic reservoirs of infection giving rise to intermittent epidemics in their more thinly settled hinterlands. Underlying this is the assumption that much of the excess mortality characterizing such centres reflected the action of airborne, or other, infections transmitted directly from person to person.²⁴ Such an assumption appears plausible given the high population densities they displayed, but it runs counter to the widespread attribution of such excesses to water- and food-borne pathogens which might not be so readily transmissible beyond the city itself.

The resolution of this question is hampered by the weakness of the cause-of-death classification contained in the Bills, but it can be approached by moving from the annual to the monthly level and analysing the seasonal incidence of mortality in London. The rationale of this procedure is that the seasonality of mortality tends to differ according to the mode of transmission of the pathogen concerned. Thus data from Victorian London show a strong summer peak for dysentery, diarrhoea, and

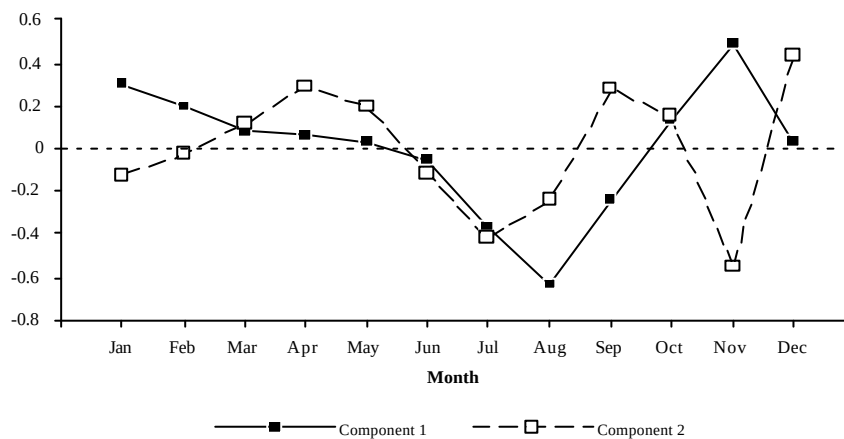
²⁴ For this purpose typhus, which is actually arthropod-borne, can be treated as a person-to-person infection since close physical proximity is usually required in order to transmit the disease.

non-respiratory tuberculosis mortality, with a corresponding winter peak for typhus, respiratory tuberculosis, and bronchitis. A third group of conditions, including diphtheria and scarlet fever as well as typhoid, were distinguished by a marked autumn seasonality.²⁵

A reconstruction of burial seasonality in London over our period, however, yielded unexpected results, for the overall pattern proves to have altered substantially in the course of the eighteenth century (see Table 3).²⁶ There appear to be two main changes, of which the first, accomplished between the late seventeenth and mid-eighteenth centuries, is the more far-reaching. This transforms an initial summer peak, strongly centred on August, into a much broader 'cold-weather plateau' stretching from November to April; the months of June and July now representing a trough in mortality, with only a moderate excess remaining in September. In the second phase, from 1775 to the end of the period, a markedly bimodal pattern emerges. The cold-weather plateau is much reduced, becoming clearly focused on the months of January and February, whilst a new burial peak emerges in November.

Figure 10a

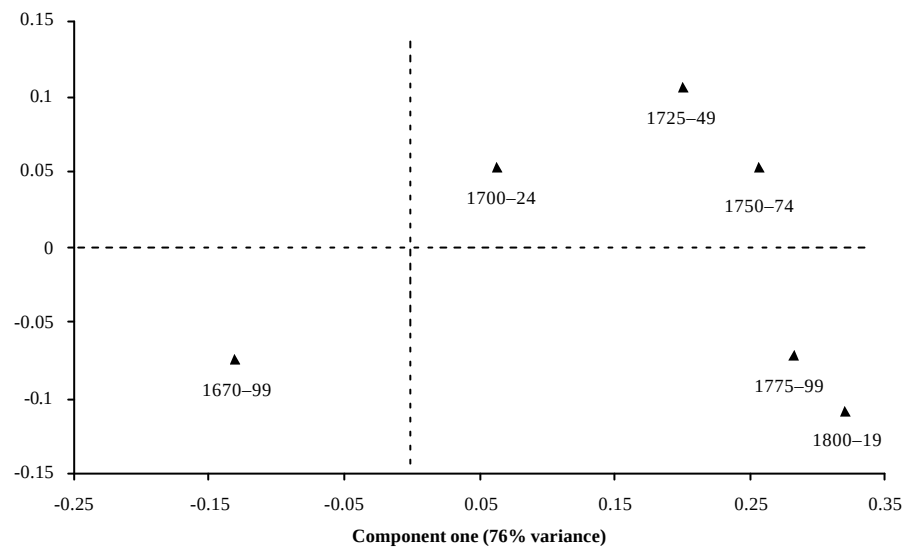
Bills of mortality: burial seasonality principal component coefficients



²⁵ The Victorian data are taken from Buchan and Mitchell (1875); see Landers and Mouzas (1988) for details of the methodology employed in the analysis of the Bills' seasonality, and the technical and conceptual problems encountered.

²⁶ The monthly burial indices given in the table express the observed burial total for each month as a percentage of what would be expected if the risk of mortality were to be distributed evenly across the year.

Table 10b
Principal component scores



These movements in seasonality were summarized in the form of a principal components analysis (Everitt and Dunn 1983), as in Figures 10a and 10b in which 97 per cent of the variation between the indices for the six periods was accounted for by the first two components. The plot of scores on the latter suggests that the first should be thought of as an axis of secular change: a change continuing throughout the period, though at a substantially reduced rate after 1750. Inspection of the coefficients reveals, as we might expect, a pattern of strong negative associations with the indices for the months in the third quarter of the year, especially August.

The second axis, by contrast, separates the 1700-74 periods from the remainder. The coefficients here are harder to interpret, but the combination of relatively strong positive values in December and April, with weak or negative ones for the intervening months, is consistent with the prevalence of the cold-weather plateau of mortality at this time. The first three-quarters of the eighteenth century thus stand out with a distinctive seasonal pattern of burials, over and above the longer-term movement from a summer to a winter peak of mortality.

The decline in the initial summer peak of mortality is apparently associated with a decline in deaths from food-borne gastro-intestinal conditions (particularly that known to contemporaries as 'gripping in the guts'), and it is most likely that this was due either to changes in the relevant pathogen or, conceivably, to the climatic change which occurred at this time.²⁷ The development of the 'cold-weather plateau', however, seems to have reflected an increase in the relative weight of deaths from diseases such as typhus and the bronchitis-influenza-pneumonia group in the global burial total.

²⁷ This assertion is based partly on negative evidence: the apparent absence of any significant improvement in sanitary conditions at this time; and partly on the occurrence of a similar decline in summer deaths in neighbouring rural districts (Greatorex n.d.)

The trend in overall mortality in London between the later seventeenth and early eighteenth century is difficult to establish with any certainty, but it seems very unlikely that it was declining at this time, implying that there was an absolute as well a relative increase in mortality from these causes of death. The pattern is further complicated by a rise in mortality from a further cause of death, or group of causes, with a marked autumn seasonality in the central decades of the eighteenth century, a phenomenon most marked in the case of the fever burial series. Outside London, however, there appears to have been a definite amelioration of mortality in south-eastern England in the early eighteenth century.

We can pursue this further with the aid of an alternative source of vital data: the parish registers. The advantage of these is that, in many cases, it is possible to distinguish the burials of adults from those of children and thus to perform an elementary analysis of age-specific patterns of burial seasonality. Against this, however, must be set the much more time-consuming nature of the preliminary data collection involved. Hence the analysis was limited to two decades, 1695-1704 and 1750-59, for which the available parish registers were examined and those distinguishing child burials included in the main sample.

If the assumptions of the high potential model are correct we should expect to find that most variation in burial seasonality occurs in the childhood age-groups, since it is here that individuals first encounter the pathogens present in the environment and either succumb or develop a degree of immunity. The results in Figures 11 and 12 generally bear out this expectation. Some change occurs in both adult and child burial seasonality, but the latter is far more pronounced suggesting that it is indeed in childhood that the major changes in cause-of-death structure are apparent.

Figure 11
Burial seasonality indices 1695-1704

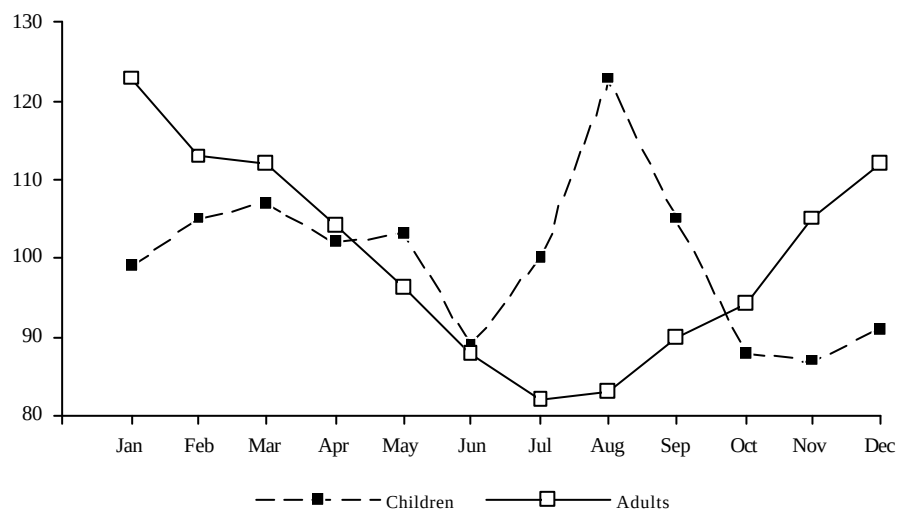
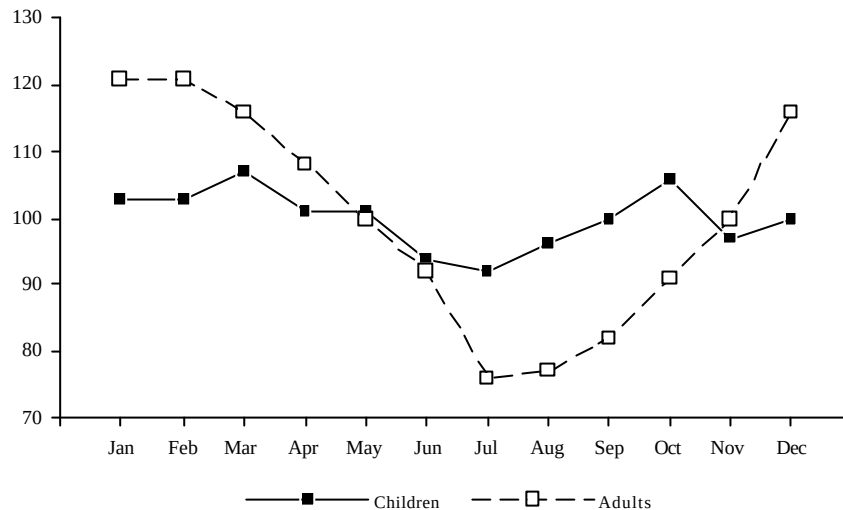


Figure 12
Burial seasonality indices 1750-59



Mortality change and its context

We can also use the parish register sample to investigate spatial variations in burial seasonality. Figures 13 and 14 depict the seasonal pattern of child burials for four groups of parishes in each of the two decades. There are some differences between the groups for the decade 1695-1704, but this is primarily a matter of quantitative variation in a common underlying pattern. In 1750-59, however, the position is transformed. The summer peak has evidently disappeared throughout London, but the new autumn excess is very unevenly distributed, being wholly absent from the wealthier western districts.²⁸

This emerging spatial differentiation in London's epidemiological regime seems to point to a factor underlying the latter's evolution over the course of our period. This was the problem of housing conditions. The peculiar forms of housing tenure and finance in the capital at this time gave rise to a pattern of house building which was highly speculative and characterized by a dramatic alternation of 'booms' and 'busts'.²⁹ Under favourable conditions the system was capable of turning out large numbers of houses in a short time, but the quality was often very poor and any deterioration in the economic climate, due to the outbreak of war or one of the period's recurrent financial crises, stopped activity dead in its tracks with a wave of bankruptcies and a crop of abandoned half-built houses (Ashton 1959; Hoppitt 1987).

Figure 13
Child burial seasonality indices, 1695-1704

²⁸ The evidence of baptism burial ratios suggests that child mortality was also substantially higher in districts with the autumn burial peak, and that spatial differences in mortality levels had increased markedly since the decade 1695-1704.

²⁹ See Landers (1993) for a survey of the literature on this subject. The standard accounts are those of Ashton (1959); George (1966), Summerson (1978) and Sheppard, Belcher and Cottrell (1979), whilst Clarke (1992) provides a detailed analysis from the standpoint of Marxist mode-of-production theory.

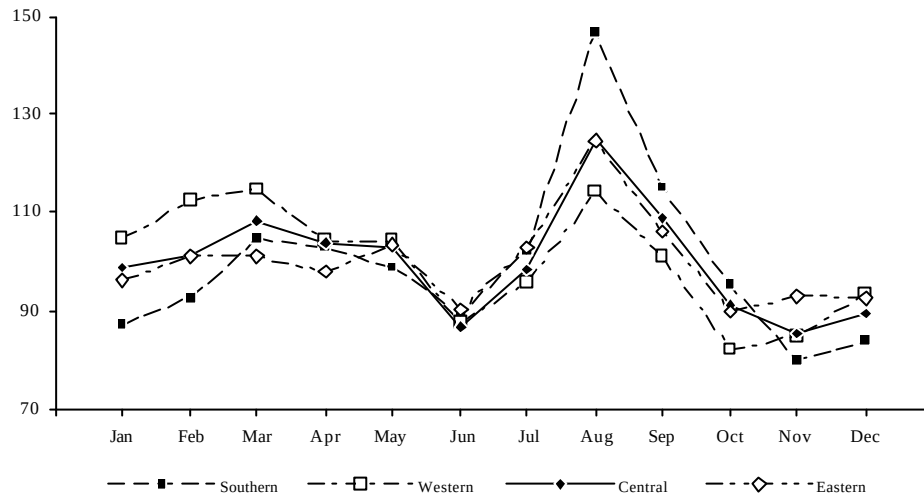
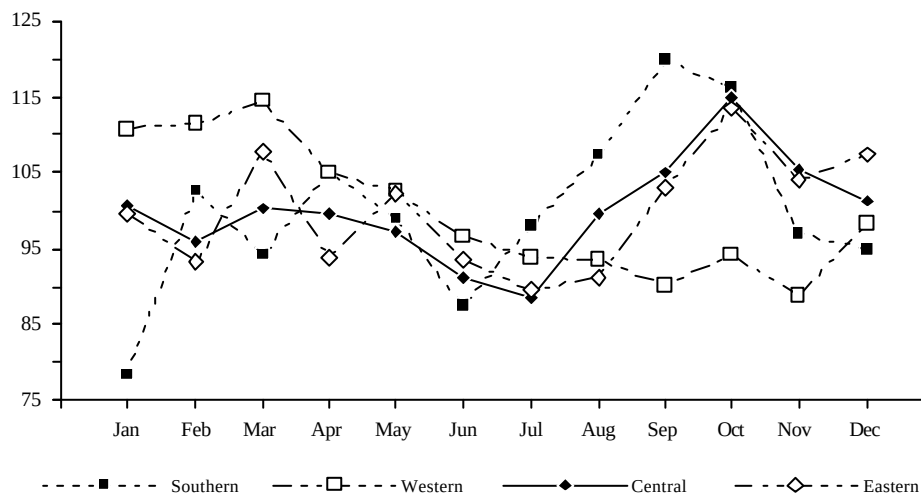


Figure 14
Child burial seasonality indices 1750–59



Equally significant from our point of view is the fact that the great bulk of new housing was built for the Žlite and the better off among those whom contemporaries termed ‘the middling sort’. The result of this was that the housing stock grew, in effect, ‘from the top down’ and increments to the accommodation available to the mass of the population came about through an involutionary process akin to ‘internal colonization’, properties being divided and subdivided as they were left behind by the westward movement of the better off.

Thus, Francis Sheppard, the historian of nineteenth-century London, describing the process at work shortly after 1800, noted

the absence, except in parts of East London, of new building specifically for the working class, for most early nineteenth century landlords, being anxious for the maintenance of the value of their property, intended that their estates should be occupied by the upper or the middle classes. But... many houses intended for the 'respectable' classes often degenerated into slums within a decade or two of their erection. Once this process had begun it was almost impossible to reverse it. Decaying houses intended for occupation by a single family, with cellars, large rooms and an inadequate water supply, were invaded by half-a-dozen or more families, for whom they provided utterly unsuitable accommodation (Sheppard 1971:94).

Such a system might have provided tolerable results as long as the built-up area was expanding, but this ceased to be the case, for any sustained period, for a generation after the financial crisis of the late 1720s. Under these circumstances it is likely that popular housing conditions deteriorated considerably, especially in the older inner-suburban districts and parts of the East End, with an accompanying increase in the two proximate determinants of exposure to infection which we have termed conduction and retention.

It is the latter which was probably responsible for the emergence of spatial variations in the seasonality of burials. The autumn peak in childhood mortality seems to be a reflection of a 'new' disease which had arrived in an epidemic form in the late 1730s and was probably a form of streptococcal infection (Landers 1993, Appendix 2). On its first appearance this had taken a substantial toll across the capital and among all social classes, but it apparently became endemic only in those districts worst affected by the economic stagnation of the period.

If this argument is correct, then the resumption of growth in the last quarter of the eighteenth century would have produced a corresponding amelioration of conditions capable of reducing mortality despite the contemporary decline in real wages (Schwarz 1985, Floud et al. 1990). It does not seem plausible, however, to attribute the whole of such a dramatic and sustained decline in mortality to this factor alone, and it is likely that some of the traditional explanations, based on contemporary economic changes, are of relevance. In particular, the wider availability of cotton cloth, and the increased output of soap, probably improved personal hygiene in a manner particularly important for typhus mortality (Chambers 1972; Razzell 1974).

At the same time, water supplies increased in volume, even if their quality was as bad as, or worse than, ever with important implications for the 'water washed' diseases such as dysentery (Bradley 1974; Rudden 1985; Hardy 1991). In addition there is evidence to suggest that improvements in street paving and cleaning, in some parts of London, may have reduced the volumes of filth and excrement in public spaces that had earlier fostered the breeding of flies and thus the transmission of disease (George 1966; Porter 1991). It is likely that changes in infant feeding regimes made a substantial contribution to the decline in infant mortality (Landers 1993, Chapter 4), and inoculation against smallpox may have been important for child mortality although the direct evidence on this point is equivocal at best (Vann and Eversley 1992:222–223).

Conclusion

The secular decline in mortality, the 'mortality transition', primarily reflects a reduction in mortality from infectious diseases, and it is this which explains much of the cross-sectional variation in contemporary death rates. Recent experience has shown that the relationship between infectious disease

mortality and socioeconomic change is more complex than was once thought, and if the problems thus uncovered are to be resolved we will have to gain a much deeper understanding of the factors responsible for the former's long-term dynamic. European historical demography provides an important basis for pursuing such an understanding, but in practice its study has been pursued with little reference to the question of mortality transition and its *sequelae*.

This in turn reflects theoretical difficulties which have arisen in historical mortality studies following the failure of explanations based on a narrow standard-of-living determinism. Recent work in historical epidemiology, however, has suggested that a broader range of structural characteristics of past populations, in particular their spatial structure, may be invoked to explain mortality change which need not be regarded as an 'autonomous' variable determined by factors outside the human world. The implications of this work can be drawn out with the aid of a proximate determinants framework, giving rise to the concept of structurally-determined 'mortality potentials'.

One such configuration of proximate determinants characterized large urban centres in pre-transitional Europe and defined a 'high-potential' epidemiological regime. According to this model such centres functioned as endemic reservoirs of infection with mortality levels which were correspondingly high, especially early in life, but relatively constant from year to year. Eighteenth-century London, a critical test case, displayed these characteristics in a highly developed form. Nonetheless, its mortality patterns were strongly affected by the political events of the period, in particular the alternation of war and peace, and by the social and economic organization of the construction sector. These factors were highly specific to the time and place in question, and they serve to underline the inadequacy of ahistoric conceptualizations of mortality change.

References

- Arriaga, E.E. 1989. Changing trends in mortality declines during the last decades. Pp. 105–130 in *Differential Mortality*, eds. L. Ruzicka, G. Wunsch and P. Kane. Oxford: Clarendon Press.
- Ashton, T.S. 1959. *Economic Fluctuations in England 1700-1800*. Oxford: Clarendon Press.
- Bengtsson, T., G. Fridlitzius and R. Ohlsson. 1984. *Pre-Industrial Population Change*. Stockholm: Almqvist and Wiksell.
- Bongaarts, J. 1978. A framework for analyzing the proximate determinants of fertility. *Population and Development Review* 4,1:105–132.
- Bradley, D.J. 1974. Water supplies: the consequences of change. In *Human Rights in Health*. Ciba Foundation Symposium no. 23. Amsterdam: Ciba Foundation.
- Brass, W. 1971. On the scale of mortality. Pp. 69–110 in *Biological Aspects of Demography*, ed. W. Brass. London: Taylor and Francis.
- Braudel, F. 1980. History and the social sciences: the *longue durée*. Pp. 25–54 in F. Braudel, *On History*, London: Weidenfeld and Nicolson.
- Buchan, A. and A. Mitchell. 1875. The influence of weather on mortality from different diseases and at different ages. *Journal of the Scottish Meteorological Society* 4:185–263.
- Bynum, W.F. and R. Porter eds. 1991. *Living and Dying in London*. London: Wellcome Institute for the History of Medicine.
- Chambers J.D. 1972. *Population, Economy and Society in Pre-Industrial England*. Oxford: Oxford University Press.
- Clarke, L. 1992. *Building Capitalism*. London: Routledge.
- Coleman, D. and J. Salt. 1992. *The British Population: Patterns, Trends and Processes*. Oxford: Oxford University Press.
- De Vries, J. 1984. *European Urbanisation 1500-1800*. London: Methuen.

- Dobson, M. 1989a. The last hiccup of the old demographic regime: population stagnation and decline in late seventeenth and early eighteenth-century south-east England. *Continuity and Change* 4,3:395-428.
- Dobson, M. 1989b. Mortality gradients and disease exchanges: comparisons from old England and colonial America. *Social History of Medicine* 2,3:259-298.
- Everitt, B.S. and G. Dunn. 1983. *Advanced Methods of Data Exploration and Modelling*. London: Heinemann Education Books.
- Flinn, M.W. 1974. The stabilisation of mortality in pre-industrial western Europe. *Journal of European Economic History* 3:285-318.
- Floud, R., K. Wachter and A. Gregory. 1990. *Heights, Health and History: Nutritional Status in the United Kingdom, 1750-1980*. Cambridge: Cambridge University Press.
- George, D. 1966. *London Life in the Eighteenth Century*. London: Penguin.
- Greatorex, I. n.d. Burial seasonality in early modern England. Unpublished Manuscript in library of ESRC Cambridge Group, Cambridge UK.
- Hardy, A. 1988a. Diagnosis, death and diet: the case of London 1750-1909. *Journal of Interdisciplinary History* 18:387-401.
- Hardy, A. 1988b. Urban famine or urban crisis? Typhus in the Victorian city. *Medical History* 32:333-357.
- Hardy, A. 1991. Parish pump to private pipes: London's water supply in the nineteenth century. Pp. 76-93 in *Living and Dying in London*, eds. W.F. Bynum and Roy Porter. London: Wellcome Institute for the History of Medicine.
- Hoppitt, J. 1987. *Risk and Failure in English Business 1700-1800*. Cambridge: Cambridge University Press.
- Horiuchi, S. 1989. Some methodological issues in the assessment of the deceleration of the mortality decline. Pp. 64-78 in *Differential Mortality*, eds. L. Ruzicka, G. Wunsch and P. Kane. Oxford: Clarendon Press.
- Kunitz, S.J. 1983. Speculations on the European mortality decline. *Economic History Review* (2nd ser.) 36:349-364.
- Kunitz, S.J. 1987. Making a long story short: a note on men's heights and mortality in England from the first through the nineteenth centuries. *Medical History*. 31: 269-280.
- Landers, J.M. 1986. Mortality, weather and prices in London 1675-1825: a study of short-term fluctuations. *Journal of Historical Geography* 12:347-364.
- Landers, J.M. 1987. Mortality and metropolis: the case of London 1675-1825. *Population Studies* 41:59-76.
- Landers, J.M. 1990. Age patterns of mortality in London during the 'long eighteenth century': a test of the 'high potential' model of metropolitan mortality. *Social History of Medicine* 3:27-60.
- Landers, J.M. 1991. London's mortality in the 'long eighteenth century': a family reconstitution study. Pp. 1-28 in *Living and Dying in London*, eds. W.F. Bynum and Roy Porter. London: Wellcome Institute for the History of Medicine.
- Landers, J.M. 1993. *Death and the Metropolis: Studies in the Demographic History of London 1670-1830*. Cambridge: Cambridge University Press.
- Landers, J. and A.J. Mouzas. 1988. Burial seasonality and causes of death in London 1670-1819. *Population Studies* 42:59-83.
- Lee, R.D. 1986. Population homeostasis and English demographic history. Pp. 75-100 in *Population and Economy*, eds. R.I. Rotberg and T.K. Rabb. Cambridge: Cambridge University Press.
- Lunn, P.G. 1991. Nutrition, immunity and infection. Pp. 131-145 in *The Decline of Mortality in Europe*, eds. R.S. Schofield, D. Reher and A. Bideau. Oxford: Clarendon Press.
- Marshall, J. 1831. *Mortality in the Metropolis*. London.
- McKeown, T.R. 1976. *The Modern Rise of Population*. London: Edward Arnold.

- McKeown, T.R. and R.G. Record. 1962. Reasons for the decline of mortality in England and Wales during the nineteenth century. *Population Studies*. 16:94-122.
- McNeill, W.H. 1980. Migration patterns and infections in traditional societies. Pp. 27-36 in *Changing Disease Patterns and Human Infections*, eds. N.F. Stanley and R.A. Joske. London: Academic Press.
- Mitchell, B.R. and P. Deane. 1962. *Abstract of British Historical Statistics*. Cambridge: Cambridge University Press.
- Mosley, W.H. and L.C. Chen. 1984. An analytical framework for the study of child survival in developing countries. Pp. 25-48 in *Child Survival: Strategies for Research*, eds. W.H. Mosley and L.C. Chen. Supplement to *Population and Development Review*, Vol. 10.
- Perrenoud, A. 1984. Mortality decline in its secular setting. Pp. 41-69 in *Pre-Industrial Population Change*, eds. T. Bengtsson, G. Fridlitzius and R. Ohlsson. Stockholm: Almqvist and Wiksell.
- Perrenoud, A. 1991. The attenuation of mortality crises and the decline of mortality. Pp. 18-37 in *The Decline of Mortality in Europe*, eds. R.S. Schofield, D. Reher and A. Bideau. Oxford: Oxford University Press.
- Porter, R. 1991. Cleaning up the Great Wen: public health in eighteenth century London. Pp. 61-75 in *Living and Dying in London*, eds. W.F. Bynum and Roy Porter. London: Wellcome Institute for the History of Medicine.
- Post, J.D. 1985. *Food Shortage, Climatic Variability and Epidemic Disease in Pre-Industrial Europe: The Mortality Peak of the Early 1740s*. Ithaca: Cornell University Press.
- Razzell, P.E. 1974. 'An interpretation of the modern rise of population in Europe' – a critique. *Population Studies* 28:5-17.
- Rotberg, R.I. and T.K. Rabb, eds. 1986. *Population and Economy*. Cambridge: Cambridge University Press.
- Rudden, B. 1985. *The New River: A Legal History*. Oxford: Clarendon Press.
- Ruzicka, L., G. Wunsch and P. Kane, eds. 1989. *Differential Mortality: Methodological Issues and Biosocial Factors*. Oxford: Clarendon Press.
- Schofield, R.S. 1985. The impact of scarcity and plenty on population change in England, 1541-1871. Pp. 67-94 in *Hunger and History*, eds. R.I. Rotberg and T.K. Rabb. Cambridge: Cambridge University Press.
- Schofield, R.S., D. Reher and A. Bideau, eds. 1991. *The Decline of Mortality in Europe*. Oxford: Oxford University Press.
- Schwarz, L.D. 1985. The standard of living in the long run: London, 1700-1860. *Economic History Review* (2nd ser.) 38:24-41.
- Sharlin, A. 1978. Natural decrease in early modern cities: a reconsideration. *Past and Present* 79:126-138.
- Sheppard, F. 1971. *London 1801-1870: The Infernal Wen*. London: Secker and Warburg.
- Sheppard, F., V. Belcher and P. Cottrell. 1979. The Middlesex and Yorkshire deed registries and the study of building fluctuations. *London Journal* 5:176-217.
- Stanley, N.F. and R.A. Joske, eds. 1980. *Changing Disease Patterns and Human Infections*. London: Academic Press.
- Summerson, J. 1978. *Georgian London*. London: Peregrine.
- Szreter, S. 1988. The importance of social intervention in Britain's mortality decline c1850-1914: a re-interpretation of the role of public health. *Social History of Medicine* 1:1-38.
- Vann, R. and D.E.C. Eversley. 1992. *Friends in Life and Death: The British and Irish Quakers in the Demographic Transition*. Cambridge: Cambridge University Press.
- Wrigley, E.A. 1967. A simple model of London's importance in changing English society and economy 1650-1750. *Past and Present* 37:44-70.
- Wrigley, E.A. 1968. Mortality in England: Colyton over three centuries. *Daedalus* 97:546-580.
- Wrigley, E.A. and R.S. Schofield. 1981. *English Population History 1541-1871: A Reconstruction*. London: Edward Arnold.

Wrigley, E.A. and R.S. Schofield. 1983. English population history from family reconstitution: summary results.
Population Studies 37:157–184.

Contours of death : disease, mortality and the environment in early modern England



Mary J. Dobson

Wellcome Unit for the History of Medicine, 45–47 Banbury Road, Oxford OX2 6PE, UK

Abstract

Interest in health and the environment dates back many thousands of years and is particularly associated with the Hippocratic works, *On Airs, Waters and Places*. The seventeenth and eighteenth centuries saw a resurgence of Hippocratic ideas. Physicians and topographers began to collect sets of medical and environmental observations and to ask why diseases varied according to locality or season, why certain environments seemed more conducive to ill-health than others, and, in turn, whether such knowledge could be used to intervene, ameliorate, manage or avoid unhealthy sites. This paper seeks to readdress these environmental issues using the tools and techniques of historical demography and a multifactorial approach. Mortality and epidemiological data from a very large number of parishes in Southeast England are analysed, as a way of understanding the influence of 'airs, waters and places' on early modern populations. The paper concludes that features of the natural environment accounted for some of the spatial variations in disease and mortality, but many epidemiological patterns were far more complex, reflecting the significance of a range of environmental, social, economic, biological and demographic variables. In trying to reconstruct past epidemiological landscapes, it is now time to 'move beyond the real wage', to move beyond 'airs, waters and places', and to avoid the temptation to search for any single determinant of mortality patterns and their changes over time and space.

Introduction: airs, waters and places—beyond the real wage

Economic and demographic historians of the early modern period are now familiar with the concept of the 'real wage' and its role as a determinant of population change in the era of the demographic transition (Wrigley and Schofield 1981). From the time of Malthus to the post-McKeown era, scholars have grappled with such variables as wages, income, labour supply and demand, subsistence resources, grain prices and levels of nutrition as determinants of both short- and long- term fluctuations and changes in mortality (Malthus 1986 [1798]; McKeown 1976; Appleby 1978; Wrigley and Schofield 1981; Walter and Schofield, 1989). The papers by Landers and Kunitz and Engerman in this volume discuss the contributions of historians to the real-wage and standard-of-living debate in the context of mortality change. These papers show that findings of recent generations of quantitative historians remain far from conclusive: the usual expectation of a secular relationship between changing real wages and changing mortality was not as Kunitz and Engerman write always realized. It is now time to move 'beyond the real wage'.

To move 'beyond the real wage' requires, as Landers (1986, 1987, 1990) has shown in his recent works, that we examine critically the position of mortality both within the domain of historical demography and within the cognate fields of economic and social history. Landers has done much to reassert the role of mortality, to rescue it from the confines of 'autonomy' or 'exogeny' (Chambers 1972; Lee, 1985) and to look again at the correlates, the flows and linkages between mortality, patterns of disease transmission and economic, social and environmental explanatory variables (Landers 1990). It is not the intention of this paper simply to reiterate these current concerns though the present author shares Landers's view that mortality deserves renewed recognition. Instead, this paper steps back in time and looks, first, at some of the environmental ideas that have preoccupied a succession of physicians and writers for many hundreds of years before and after the time of Malthus. It then examines how these same preoccupations can now be fitted into a framework of historical epidemiology. It takes as its theme the Hippocratic heritage of 'airs, waters and places'.

Patients and physicians have for centuries tried to understand the ebb and flow and geographical patternings of epidemic and endemic diseases. The idea that atmospheric and environmental influences might affect patterns of disease and mortality dates back many thousands of years but is particularly associated with the works of Hippocrates in Greece in the fifth century BC (Sargent 1982). The Hippocratic concept of 'airs, waters and places' received significant and renewed attention in the seventeenth and eighteenth centuries. Physicians began to ask why diseases varied according to locality or season, why certain environments seemed more conducive to ill-health than others, and, in turn, whether such knowledge could be used to intervene, ameliorate, manage or avoid unhealthy sites (Jordanova 1979; Riley 1987). The lie of the land, the nature of the terrain, the type of soil, the proximity to stagnant, salt or fresh water sources, moderated or tempered by the direction of the wind, the temperature of the air and the quality of the atmosphere all appeared to influence and impinge on the health and well-being of a locality and its residents.

One of the most striking of these environmental images was the notion of 'bad air' and its association with marshland topography. Mortality levels in marshland localities were reckoned to be unusually high and residents and visitors complained bitterly about the marsh fevers or agues they encountered in such localities. Many physicians and writers in the early modern period were convinced that it was the noxious stench that appeared to emanate from the stagnant marsh waters, which were primarily responsible for such hazardous conditions. In Romney Marsh

the large quantity of stagnating waters ... engenders such noxious and pestilential vapours as spread sickness and frequent death on the inhabitants ... the sickly countenances of them plainly discovering the unwholesome air they breathe in (Hasted 1797-1801 vol. 7:253-254).

and in the Fens 'awful reservoirs of stagnated water' were observed 'which poisons the circumbient air ... and sickens and frequently destroys many of the inhabitants' (Parkinson 1811:21). The link between marsh fever and stagnant water was, in fact, so clear that anyone could predict with considerable certainty, upon examining a locality, whether or not the residents were subject to that affliction (Forbes 1836:183). Indeed, the stench, the miasmas and the fevers became such a recognized feature of marshland terrain that it was this miasmatic explanation of the marsh fever which gave rise to the word 'malaria' or 'mal'aria', literally meaning 'bad air'.

Elsewhere, topographers and physicians searched for further links between disease patterns, bad airs and stagnant waters. Many remained convinced that even beyond the English marshes there was 'something in the air'. They, accordingly, looked towards terrestrial and atmospheric effluvia emanating from rivers and streams or from stagnant sources of decay, putrefaction and dirt. In urban

settings writers shifted their attention away from the natural environment, elements of altitude, soils, terrain, exposure, wind direction, and focused instead on attributes of the man-made environment in conjunction with the conditions of its population. Towns and the confined spaces of the poor and sick were often viewed as places of 'a thousand stinks'. Steams of effluvia and noxious vapours were believed to arise from open sewers, churchyards, slaughter houses, butchers' shops and lanes, dead flesh, burial grounds, cesspools and from every other sort or putrefaction, excrement, decay, human and animal filth. The offensive and 'notorious fountains of stench' corrupted the air, created terror amongst the inhabitants, and made 'the people sick and faint as they pass by' (George 1966 [1925]:349). Putrid exhalations might also arise and be contained in such closed spaces as cellars, garrets, cells, common lodging houses, tenements, courts, alleys, alehouses, ships, prisons, and workhouses. These were the pestilential black-spots, the 'pest-houses of concentrated contagion', the sinister spaces where 'darkness, dirt and stagnant air combine to augment all the evils resulting from such a situation' (George 1966:64, 96). These were the pockets of high mortality.

Fresh airs and running waters were the antithesis to bad airs and stagnant waters. Country airs and sea airs, distant from marshland exhalations, were said to be very healthy, especially

dry, open situations meanly elevated, neither like beacons on tops of lofty mountains, nor like reeds in the marshy vallies, are above all others (*caeteris paribus*) the healthiest; for such habitations have a free, pure, open air (Short 1973 [1750]:13).

Spring waters and water from fast-flowing streams and rivulets were praised for their purity of taste and smell. In urban settings, social reformers and military doctors began to focus more and more of their attention on the advantages of fresh airs and pure waters. Freely circulating air, the removal of all foul substances, the supply of fresh water in the towns, in the apartments of the poor, in the hulks of the ships, in the prisons and elsewhere became as important an objective as the removal of bad airs from the marshes (Riley 1987).

In order to substantiate their environmental hunches, physicians began to collect and collate a wide range of information on disease and mortality, atmospheric and environmental circumstances, clinical and meteorological observations (Cassedy 1974; Jordanova 1979; Sargent 1982; Riley 1987). With such information, men like Thomas Short remained convinced that they would be able to

shew the rise, progress, extent, severity or mildness, duration, seasons, and degrees of mortality in sundry places, by endemics and epidemics ... which diseases have their frequentest returns, and what places and soils are most liable to them, or suffer slightest or sharpest by them (Short 1973:110).

Extensive lists and tables were produced and published; mortality data from the Bills of Mortality and parish registers were compared across cities, countries, regions and villages; histories of epidemic diseases and meteorological fluctuations were compiled; arithmetical calculations, ratios, proportions, life tables, and simple measures of association were introduced in order to measure and understand the geographical and environmental influences affecting human health.

The medical topographies, environmental and epidemic histories produced by these writers contain a wealth of astute and interesting observations. But in trying to make sense of disease causation and the environmental links, many writers became increasingly aware of the complexities of their epidemiological landscapes. 'Airs, waters and places' were not always sufficient to explain and understand the seemingly complex and elusive patterns of epidemic disease. They added to their list of 'airs, waters and places' a whole constellation of geophysical, economic, medical, and social influences. They considered associations concerning both places and people, situations that generated diseases,

situations that brought people into contact with a disease, situations which predisposed individuals to the disease (Hamlin 1992:52). A large number of different forces were discussed: meteors and comets, states and prices of the fruits of the earth, cattle distempers, the character and customs of the citizens, diet and alcohol, cleanliness, religion, clothing, modes of employment, housing, crowding and ventilation, trades and manufactures, manner of life, wealth, populousness and circumstances of the inhabitants, local therapies and medical practices, hospitals and quarantine establishments, the role of infection and means of dissemination; any or all of these influences might hold the key to the complex and elusive patterns of epidemic disease.

The role of environmental influences: a framework for historical demography

The data compiled by these writers tell us much about the outline images of the worlds of disease and death that surrounded them while their attempts to gather more and more data, to explore the relationship between mortality patterns and extensive lists of variables serve to remind us of the complexity of epidemiological associations in times and places of disease and death. In spite of their aims and ambitions, in the end they were unable to quantify and unravel the complexity of the relationships. As Thomas Sydenham, the English Hippocrates of the seventeenth century, had been forced to admit:

much and diligently as I have observed the different characters in respect of the manifest atmospheric changes of different years with a view to detecting therein the reasons for the discrepancy amongst epidemic diseases I confess that I cannot find I have proceeded one single inch on my way (Sydenham 1848-50 vol. 1:33).

Today, as historical demographers, we have infinitely superior tools of computational analysis. We also have a more complete picture of the pathogens, the vectors, the vehicles and human intricacies of disease transmission. In particular, we are aware that different diseases have different modes of transmission. Environmental forces can be important but they are disease-specific. The influences that shape and determine one disease may be quite different from those that control another. We are aware, too, that human response in the face of infection can vary enormously depending on such factors as age, gender, previous exposure, nutritional experience, level of immunity and resistance. And again, these will vary in turn, according to the particular infection. Smallpox, plague, malaria, typhus, typhoid, dysentery, tuberculosis, venereal disease, influenza, infantile diarrhoea and a phlethora of other afflictions of pre-industrial times all had their own distinctive patterns and paths of human interaction and epidemiology. With this knowledge and with the statistical and computing packages now available, can we begin to understand the spatial patterning of the complex array of diseases that afflicted early modern societies: to appreciate why certain localities, environments, seasons, and periods of time seemed to experience a heavy toll of epidemic disease and mortality while some remained relatively free from the frequent onslaughts of plague, smallpox and fevers? Indeed, how far are we able to confirm that the gradients, the contrasts and the fluctuations in health conditions of different localities and seasons, as described by topographers, actually existed in England at this time?

Historical demographers and epidemiologists have made surprisingly little inroad into these spatial and environmental issues. Indeed, the approach of some econometric historians who seek for a unidirectional relationship between one factor, such as temperature or harvest prices, and national or aggregate mortality data, actually runs counter to what we now know about the complexity of local and individual disease patterns and their environmental influences (Lee 1981). In comparison with the outpouring of historical discussion surrounding Malthus and the role of the real wage, recent work in

the field of historical demography has done little to substantiate, refute or develop the environmental ideas of seventeenth and eighteenth century medical topographers (Riley 1987).

If we are to move beyond the limitations of the early writers, then we must devise strategies for mapping and explaining disease and mortality patterns which take into account local variations and complexities of past epidemiological landscapes. A large data set, incorporating demographic indices for over 1000 parishes in three counties of Southeast England, mortality data for some 600 parishes extending over a period of 200 years (1600 to 1800), and a wide range of possible demographic and epidemiological influences, has been constructed from seventeenth and eighteenth century sources and analysed using cross-sectional and time-series computer packages. Mortality rates and disease patterns were compared across a diverse array of local environments and over two centuries of time, and a multivariate approach was adopted, which included aspects of the natural as well as the social, economic and human environment of each locality and each season of time; by these means a framework was devised in which to evaluate the influence of a range of variables on the health patterns of early modern populations (Dobson 1980, 1989b, forthcoming).

Bad airs and stagnant waters: the geography of malaria and the role of mosquitoes

The most outstanding epidemiological divide within Southeast England was not between rural and urban communities but along the bounds of marshland and non-marshland terrain. It was here that the natural environmental or ecological features proved to be the critical determinants of the patterning of disease and death. In the area of marshland topography and 'bad airs', the seventeenth and eighteenth century writers had written with remarkable clarity and perception. They had sensed with their noses, they had realized through their experiences and their ill-health that some unique and peculiar quality of the marsh air gave rise to frequent suffering and premature death. The belief that it was the 'mal'aria' of the marshes which caused the high levels of mortality and sickness was not, of course, strictly correct. It was not the 'bad air', *per se*, that contributed to marshland mortality. Rather it was an anopheline mosquito vector, capable of transmitting a parasitic disease to humans, that was the culprit in this mortal landscape. Seventeenth and eighteenth century men, women and children were observing, witnessing and falling victim to the true plasmodium malaria. But they were unaware of the real ecological parameters of this disease.

In reconstructing the demographic and epidemiological landscapes of early modern Southeast England, considerable attention is given to the role of malaria in this setting (Dobson 1982). Mortality indices, based on parish register material for over 600 parishes in Essex, Kent and Sussex, show time and again that death rates in marshland parishes were excessively high compared to other places of early modern England. Average crude death rates lay above 50 per 1000, infant mortality rates exceeded 250 or 300 per 1000, and, for at least two centuries, burials remained in excess of baptisms in many marsh parishes. Seasonal and annual fluctuations in mortality presented a near-perfect match against meteorological controls. Cool wet summers allowed marshland parishioners some respite from their deadly fevers but in the hottest months and years of the seventeenth and eighteenth centuries death rates reached exceptionally unfavourable levels. Only a constant flow of newcomers to these mortal areas, 'lookers' or marshmen in search of high wages, prevented their complete demographic decline.

The study, moreover, shows how sharp the divide was between the high mortality levels of the marshy parishes and the lower rates of adjacent 'uplands' or the favourable levels of non-marshland coastal vicinities. Places beyond the 'noisome smells arising from the salt marshes' were far more healthy and experienced significantly lower mortality levels. Indeed, while once flourishing ports like Rye and Sandwich declined in importance over the seventeenth and eighteenth centuries as silting,

stagnant waters and malaria took hold, other coastal locations, free from the marsh vapours, like Brighton, Eastbourne and Margate rose to prominence in the later eighteenth century as fashionable seaside resorts, fulfilling the demands of Georgian society for healthy seabathing and good quality 'airs'. One contour up, one stretch upstream, one mile along the coast, and malaria, with all its mortal implications, ceased to exist.

These boundaries fit in precisely with what we now know about the distribution and ecological habits of the English malaria mosquito vector, *Anopheles atroparvus*. It is exactly in those coastal and estuarine areas of slightly saline stagnant waters that this vector breeds most readily and lives in sufficiently close association with humans to act as an efficient vector of plasmodium malaria. And since the mosquito has a limited flight range it rarely transmitted the parasite beyond the marshes. The biting patterns and flight range of this vector and its climatic requirements are all discussed in more detail elsewhere (Dobson 1980, 1989a, forthcoming) where a variety of evidence from the past and present is pieced together to show, conclusively, that *vivax* malaria was endemic in the coastal marshlands and Fens of England. But, as we now know, it was a bite from a mosquito rather than a breath of 'bad air' that caused the marshland malaria.

Malaria was unique in its geography; it was one of the few major endemic diseases of Southern England that was confined, by its vector, to certain environmental limits. It was also ironically one of the few diseases of this period for which a specific therapy, cinchona bark or quinine, had been introduced and recognized to be effective in controlling the symptoms of marsh fever. Its neat ecological boundaries and striking seasonal and annual fluctuations contrasted markedly with the more widespread and erratic nature of many other diseases and afflictions of early modern England. For most diseases and in other localities certain features of the physical environment, certain aspects of climatic variation, and some attempts to remove sources of 'bad air' played an important but by no means the only role in shaping the paths and rhythms of disease and mortality. Even within the marshes of Southeast England, the history of malaria must be understood alongside a whole set of other social, demographic and economic factors (Dobson, 1980). In trying to explain the far more complex patterns and influences of the past epidemiological landscapes of Southeast England, 'airs, waters and places' filter through to the surface contours but are met by a number of other confounding or complicating variables. In our search for explanations, we can cling, like Short and others, so far to the Hippocratic heritage, but ultimately recognize that there were many diverse influences, a whole range of variable forces and parameters affecting disease and mortality, that in the end transcended the natural bounds of time and place.

Altitude and drainage: water-borne diseases and the role of human pollution

One striking and repeatedly observed characteristic of the Southeast England data was the apparent significance of altitude and natural drainage in determining variations in death rates. Low-lying communities, especially those close to rivers and streams, while not as mortal as coastal and estuarine marshland parishes, nevertheless had consistently higher death rates than 'dry upland' settlements, defined as those situated above three or four hundred feet where there was often an absence or scarcity of surface drainage, and water was obtained from wells or natural springs.

Low-lying riverine parishes had average background mortality rates of the order of 30 to 40 per 1000; infant mortality rates between 150 and 200 per 1000; life expectation at birth in the thirties seasonal mortality peaks in winter or spring with a significant rise in deaths during a cold winter; a second minor peak in autumn becoming more pronounced after a hot, dry summer; and an unstable or irregular annual mortality series (as defined by annual deviations around the mean) with little or no obvious relationship between annual mortality, harvests and prices but displaying fairly frequent

epidemics of plague (until 1667), smallpox and autumn 'fevers', probably dysentery and typhoid rather than typhus (Dobson 1982, 1989b, c).

The upland scene was, by contrast, quite different and, indeed, it was these areas along the North and South Downs, the chalk hills of Essex and High Weald of Kent and Sussex which appeared outstandingly healthy by comparison with the typical image of early modern England. Individuals lived to ripe old ages and examples of longevity were often cited as evidence of the healthiness of these places: in Little Canfield, wrote one eighteenth century observer,

the situation we may venture to say is healthy from instances of longevity in some of its inhabitants ... Richard Wyatt arrived to the age of 101 years and upwards ... a predecessor ... died here at the age of 90 ... Thomas Wood was church clerk 78 years and died in May 1738 aged 106. He kept his bed but one day and could see without spectacles to the last (Muilman 1769-72 vol. 3:264).

Our own mortality estimates substantiate these claims. Death rates were generally less than 25 or 20 per 1000; infant mortality rates scarcely rose above 100 per 1000; life expectancies approached 50 or 55 years; seasonal and annual burials deviated less sharply from year to year; there was a notable absence of autumn mortality peaks in High Wealden parishes though occasional summer mortality peaks were experienced in chalk downland parishes after unusually dry summers; major mortality surges resulting from epidemic visitations (rather than famine) occurred in upland areas but they were irregular and, on the whole, less frequent than in other parts of Southeast England (Dobson 1982).

An environmental hypothesis immediately seems to equate with these differential patterns and, as with marshland mal'aria, it is hardly surprising that seventeenth and eighteenth century topographers again believed that the quality of the airs and the waters had something to do with these contrasting situations. While they imagined morbid particles suspended in the atmosphere or deadly vapours emanating from telluric effluvia and sluggish waters, we can now look back and implicate a range of bacterial and viral organisms, which might be identified with diseases such as typhoid, paratyphoid, dysentery, viral meningitis and other gastro-enteric or water-borne 'fevers' (Dobson 1982, 1989b). Rivers and streams at low discharge, as well as stagnant marshes, may well have provided the ideal conditions for trapping such organisms and, in spite of our limited knowledge of the insanitary behaviour of early modern societies, given what we know of their habit of throwing 'divers unhealthy bodies called in English "blude, garbage and guttes"' (Dobson 1982:286) into local water courses, it is quite likely that villagers and townsfolk adjacent to such sites invariably suffered 'from a want of a sufficient supply of good water ... when the river is nearly stagnant and always unfit to drink' (Dobson 1982:287).

Just as some writers, especially from the mid-eighteenth century onwards, began to recognize and emphasize the role of human excrement, filth and decay as additional causes of 'bad air', so we can confirm that it may well have been human pollution and contamination, rather than some mysterious telluric or atmospheric effluvia, that contributed to the unhealthy nature of many riverine localities, and probably added a further pathogenic load on the already mortal marshlands. Proximity to water courses, fouled by human contact, led to the invasion of all sorts of morbid and potentially fatal water-borne pathogens, and seventeenth and eighteenth century accounts of 'autumnal fevers' in these low-lying districts (their symptoms, their seasonality, and their epidemiology) point to an unusually high prevalence of water-borne diseases (Dobson 1982). Access to natural spring or well water, by contrast, may have provided a supply of relatively pure drinking water and it is pertinent that it was during the driest summers, when the water table on the chalk downlands was low and people were 'forced to go from door to door to beg a pail of water' or carry water from the lowlands to the high chalk parishes 'for the barest necessities of life' (Dobson 1982:305), that summer mortality levels, in these otherwise healthy locations, peaked sharply (Dobson 1982). The environmental association moves from 'good and bad airs' to 'good and bad waters'. But it also moves along the chain from simple ecological cause and effect (marsh mal'aria equals mosquitoes equals malaria) to one which begins to take into account elements of human behaviour and the quality of the 'human' environment.

Urban environments and pastoral settings: the role of human behaviour and population movements

Contaminated water supplies and water-borne diseases, particularly following hot summers, appear to have played a significant role in the epidemiological regime of early modern populations, long before they rose to prominence in the cholera era of the nineteenth century. But further epidemiological evidence indicates visitations of all sorts of other diseases, implicating factors beyond 'bad waters'. Plague, smallpox, typhus, venereal disease, influenza, measles, scarlet fever, diphtheria, bronchitis, pulmonary tuberculosis, botulism, salmonella and food poisonings, brucellosis, bovine tuberculosis, trichinosis, worms, and various occupational hazards, were just some of the other afflictions that stand out on the epidemiological map of early modern Southeast England. We can assume that the chain of disease causation was entered by many sources of contamination, besides water, such as food, milk, faeces, manure, dust; diverse noxious substances, including alcohol, opium, lead and chemical pollutants; and different mechanisms of disease transmission, such as rodents, livestock, flies, fleas, lice, and person-to-person contact. A broader environmental framework could, thus, be extended to explain and include disease pathways associated with all sorts of insanitary conditions, all range of foul habitats and habits, and any number of hazardous practices and crowded occasions that would have brought together humans and their disease pathogens to make certain localities or seasons more unhealthy than others.

Within the marshland-non-marshland and riverine-lowland-upland divides, there were, indeed, further local variations in mortality levels which, on the surface, would suggest that crowded and filthy conditions were likely to produce higher levels of mortality and more varied patterns of disease. The rural-urban divide is an obvious one and while many historians have tended to present the situation, outside London, as a dichotomy (provincial towns were unhealthy, countryside was healthy) the Southeast England data, which include information for about 50 urban places, suggest that within the provincial urban hierarchy there were significant differences between towns which were not simply related to size and population numbers, but more critically associated with function, location, population movements, and the type and conditions of the urban environment. The rapidly expanding ports and docklands of Kent, for example, experienced higher mortality rates and a far more erratic pattern of epidemic mortality than many inland market towns or the cathedral city of Canterbury, while even quite small towns close to London showed a distinct disadvantage over larger but more distant towns.

The Thames and Medway ports were particularly unhealthy right through the seventeenth, and eighteenth centuries and well into the nineteenth. They maintained their excesses of burials over baptisms, their high levels of mortality, and striking autumnal seasonal peaks, at a time when other towns were witnessing reductions in death rates. In the early decades of the nineteenth century, burials still exceeded baptisms in some of these ports, infant mortality rates remained above 250 per 1000 and life expectancies were in the low thirties at a time when some towns in Southeast England were experiencing a life expectancy above 40 years (Dobson 1982, forthcoming). Adjacent to the low-lying estuarine marshes the ports were, like smaller parishes along the north shore of Kent, undoubtedly subject to 'bad airs' and 'bad waters'. Malaria and a range of water-borne infections were especially prevalent in these localities. Alcohol and opium were consumed in excessive quantities as a way of coping with the constant invasion of fevers. But contemporary accounts indicate that these were only part of a whole complex set of endemic and epidemic diseases. Southall, as early as 1730, made an interesting observation concerning bedbugs and dirt. He noted that 'not one sea-port in England is free, whereas in inland towns, bugges are hardly known' (Southall 1730). The epidemiological evidence suggests that typhus, a disease transmitted by body lice, was highly concentrated in the congested areas

of the Thames ports, the docklands and suburban London while almost unknown in other parts of Kent. Plague epidemics in the seventeenth century also showed a striking preference for the ports and outlying areas of London. Massive epidemics of plague occurred in places like Greenwich, Chatham, Sittingbourne, Gravesend, Rochester, Woolwich and Deptford. Air-borne infections, especially epidemics of smallpox, also struck these types of urban environment with particular force, frequency and intensity. Pulmonary tuberculosis was another disease concentrated in areas of crowded living and working quarters and it is likely that a considerable proportion of the winter or spring rise in mortality, especially during severe winters, was attributable to this and other respiratory causes of death in the densely populated neighbourhoods of ports and towns close to the metropolis. Indeed, the spread of many infectious diseases, scarlet fever, diphtheria, measles, pneumonia and others, would have progressed rapidly in urban areas where human interaction reached its most intimate.

When we turn to contemporary descriptions of the squalor, the filth, and the overcrowding, especially along the dockland areas of north Kent, by comparison with the more spacious and salubrious settings of some country towns, it is not difficult to imagine why they harboured and propagated the appropriate vectors, conditions and pathogens for major outbreaks of disease. Moreover, alongside their unfavourable, crowded and insanitary environments they were more frequented, more accessible by road and water, and populated by a greater mix of peoples than many inland communities. People lived, worked and interacted in close proximity with each other; they came into frequent contact with travellers and visitors from London and from overseas; they attracted throughout the seventeenth and eighteenth centuries a constant influx of new occupants and young migrants, merchants, sailors, traders and buyers; they housed a complex mix of poor and rich, young and old, temporary and permanent residents; they generated pockets of squalor and set up institutions for the diseased, the criminal and the poor; they became noted for their shifting maritime populations and notorious for attracting some of the more destitute members of Kentish society: rough seafaring men, oyster dredgers, smugglers and alehouse keepers, all of whom left their mark on the mortality statistics. They provided the ideal conditions for mixing population groups with different prior experience of disease: foreigners and residents; immunes and non-immunes; the contagious, the sick and the healthy; carriers and potential victims.

These ports were also in a prime location to receive not only people, goods, and merchandise from a range of origins but also a plethora of old and new infections, their hosts and vectors, from places like the West Indies, the Americas, Africa and Europe. This was a zone where people and pathogens constantly interacted, where travellers, immigrants and residents were caught up in the 'unification microbienne du monde' (Le Roy Ladurie 1978) as global exchanges of disease followed the channels and routes of overseas trade. The ebb and flow of population movements, with all their concomitant complexities and ramifications, added an indeterminate but highly relevant role to the paths and patterns of disease and mortality.

Within the rural areas of lowland and upland Southeast England other more subtle but perhaps equally significant contrasts are manifest in the mortality statistics, contrasts which again suggest that gradients of disease and death were moulded by local characteristics and also affected by wider population movements. One such example existed between pasture and arable farming areas. Mortality levels in parishes with large herds of livestock tended to be slightly higher than those based on grain or market gardening. There was also a greater prevalence of autumnal fevers, especially during the later seventeenth and early eighteenth centuries (Dobson 1989b). All sorts of possible influences may have accounted for these patterns, some of which we can measure, others we can only infer. It would be interesting to know, for example, whether factors such as diet varied locally. Did market gardening areas benefit from the availability of fresh fruit and vegetables? Were pastoral communities more likely

to suffer from the harmful effects of contaminated meat and milk? We need more information on rural habits of sanitation and the disposal of human and animal excrement. Did some villages heed contemporary advice and keep all foul substances at a distance from their cottages? Were there differences in the arrangement and type of domestic and animal living quarters which influenced health in arable and pasture areas? If we are to extend the link between filth, crowding and population movements then we could speculate that proximity to animals, their dung, rodents, flies and other disease-carrying vectors, and the crowded domestic living conditions of dairy farms, were important epidemiological controls. In many pastoral areas, farmers in the later seventeenth century were encouraging agricultural servants to live in for a year or so rather than pay them a daily wage and, at the same time, the farmers of southeast England were increasing their herds of livestock as the price of grains started to fall. Kussmaul (1981) has described the shared sleeping and eating arrangements of servants and masters, and cites seventeenth-century references to chambers over oxhouses and servants' beds in stables. The effect of very close human and animal contact on the level of mortality may have been significant. Moreover, many of the living-in servants were hired and arrived in their new household during the autumn at a time when, especially following hot dry summers, certain pathogens and their vectors were most active. It is possible that some of these young people were moving into new types of environment and encountering new sets of diseases for which they had little prior experience, resistance or immunity. The combination of seasonal population movements and the mixing of humans, animals and disease agents in certain localities and at particular points in time may have contributed to the elevated mortality and autumn fever peaks of lowland pasture parishes.

Healthy airs and upland communities: the role of local influences and patterns of outward migration

At the opposite end of the mortality spectrum, we need to look for factors that explain the unusually favourable background, seasonal and annual mortality patterns of many upland localities. Historians traditionally enjoy focusing on the darker sides of life, the blacker pockets of human mortality, but a striking feature of this regional survey is the remarkable persistence of much lower levels of disease and death across wide areas of upland Kent, Sussex and Essex for many decades over the seventeenth and eighteenth centuries. In an era of poor and inadequate sanitation, filthy living conditions, almost non-existent medical care, low standards of nutrition, how was it possible for some communities to achieve infant mortality rates below 100 per 1000 and life expectancies as high as 50 years in the early modern period? Were there specific environmental or local features that acted as protective mechanisms, or were these places sufficiently isolated from the main flows of population to avoid some of the worst visitations of disease and death?

Eighteenth-century topographers emphasized repeatedly the exceptionally healthy air of upland parishes. These parishes were certainly well elevated and far enough away from the marshes to avoid the 'bad airs' of malaria, and, as has already been suggested, the absence of surface drainage, or 'bad waters', along the dry chalk hills may have to some extent protected them from the seasonal impact of water-borne diseases. In downland areas, arable farming and extensive sheep grazing were more important than dairying and cattle-rearing, and lesser levels of contact between animals and humans, lower densities of flies and insects and fewer sources of manure and cattle dung may have reduced the transmission of certain parasitic, bacterial or viral infections. The ready availability of local fuel supplies, in the form of timber and charcoal, may have acted as an additional protective mechanism in some upland communities and it is interesting to note that exceptionally cold winters did not produce an upsurge in mortality in the heavily wooded Wealden areas as they did in some other parts of the country. The distribution of settlements within these areas may have been less conducive to the spread

of infections than in other types of locality. Population densities along the Downs were low, there were no towns, few manufacturing industries, and little evidence of overcrowding. The High Wealden parishes covered much larger areas but within their sprawling bounds the settlements, hamlets and farmsteads were scattered over relatively wide distances. Few institutions were set up to house the diseased or the destitute in these parts and interactions between the sick and the healthy may have been kept at much lower levels than in more densely populated lowland areas.

The list of possible local influences could be extended. However, when we turn to some of the material conditions and characteristics of the inhabitants and their habitations, there is a startling inverse relationship between levels of wealth and health. These upland communities were not only areas of exceptional health but also regions of depressing or acute poverty. For the North and South Downs, the poverty was reflected in the infertile soils, the unproductive farming economy, the drab and dreary appearance of the countryside, the poor stoney and narrow state of the roads, the mean and shabby condition of the houses. Cottages were made of local flints and wood and thought old-fashioned, and the peasants themselves were described as equally rough and uncultured as the soil they tilled. In the High Weald, poverty resulted from a different set of circumstances. Population pressure and an economic dependence on a formerly profitable but declining textile industry combined to present tremendous difficulties for communities in these districts by the seventeenth century. Frequent expressions of hardship were heard at this time when the declining cloth trade of the Weald gave rise to 'the loud and heart piercing cries of the poor ... and the disability of the better sort to relieve them through the total decay and subversion of the trade' (Dobson 1989b:409).

Levels of poverty, as indicated by the proportion of exempt householders in the 1670s Hearth Tax returns, were depressingly high with some two-thirds of textile parishes classifying at least 45 per cent of their householders as non-chargeable. A century and a half later those same parishes continued to exhibit marked signs of economic malaise and poverty and, throughout the period, observers commented on the mean state of the cottages, the appalling condition of the clay soils and roads, parched in summer and deep muddy tracks in winter, and the impenetrable backward nature of the countryside where the tenantry were

as poor, weak and spiritless, as their lands, drawn down, as for ages they have been, with exhausting crops without a sufficiency of stock, or of extraneous manures, to make up for this endless exhaustion (Marshall 1798 vol. 2:133).

We do not, as yet, have detailed local evidence on standards of living, food supply, diet, clothing, sleeping arrangements, habits of cleanliness, ways of dealing with human excrement, patterns of breast feeding, infant care, or welfare of the elderly in these contrasting regions but if we are to assume that these poverty-stricken communities could not afford as many basic provisions and necessities as their contemporaries then such inadequacies were not apparent in the mortality statistics of upland communities. Pockets of poverty appear to have matched pockets of disease in many urban environments, especially by the nineteenth century (Woods and Woodward 1984; Slack 1985; Landers 1990), but within the rural sphere in the early modern period, the 'real wage' does not seem to have influenced mortality gradients. Indeed, the epidemiological consequences of this type of poverty worked in their favour. The very fact that these regions were poor, backward, impenetrable, relatively isolated and inaccessible gave them a distinct advantage over their more frequented and busy lowland counterparts. For the 200 years of this study both the Downs and the High Weald were areas of substantial outmigration. Labourers and young people moved away from these unproductive localities in steady streams to seek better fortunes in places like London, Europe and the New World (Dobson 1989b, c). They left behind an aging population and this in itself may have accounted for the instances

of longevity in these parts. But they also left behind villagers whose daily, seasonal and annual rhythms of work brought them into less frequent contact with the mixing and movements of microbes.

Indeed, it is an irony of the migrational patterns of this region that the areas of Southeast England which attracted the greatest numbers of new migrants were those that already had the least favourable environmental living conditions (the marshlands, the low-lying riverine settlements, the ports and towns) whereas the healthy but less prosperous upland environments (the Downs and Weald) were the areas shedding their populations in this period. As more and more people moved to the mortal zones, so the exchange, mixing and susceptibility to micro-organisms were constantly intensified, deepening the pools of disease and death. As fewer people entered or returned to the healthy localities, so those who stayed behind, lived or were born in these parishes enjoyed a relative freedom from the continuous invasion of old and new infections.

A chronology of epidemics: the role of complicating variables

One other perspective of this geographical survey focuses on the timing and spread of those epidemics which crossed the usual gradients of health and mortality. We have already moved from malaria, a disease bounded by natural or ecological features, to the dynamic complexities of other major diseases of the period: their local preferences and patternings, their environmental influences and their associations with the shifting behaviour and movements of early modern populations. But beyond these striking gradients of disease and death, were other more elusive patterns. Some epidemics came and went with seeming irregularity. Outbreaks of smallpox, plague, and spotted fevers, while far more prevalent in certain communities than in others, could also, from time to time, be carved haphazardly or randomly into the topographical landscapes or urban-rural hierarchies. Some healthy spots could be visited while the traditional black spots might avoid the epidemic. 'New' diseases could affect isolated, inland localities while leaving ports and busy thoroughfares unscathed. Seasons of scarcity could lead to famine and starvation in some places but not in others. Wet and cool summers could prove healthful to marshland environments but dangerous to others. It was the elusive patterning of some of the epidemic visitations across the divides of airs and waters, across the seasons of want and plenty, which continued to puzzle and frustrate the topographers in their search for simple environmental causal associations.

A reconstruction of the geographical, seasonal and annual pattern of epidemic disease, using the descriptions left by seventeenth and eighteenth-century writers and matched against the Southeast England mortality statistics, shows that the main impact and movement of epidemic disease fitted the contours of lowland, marshland, upland, and the contrasting gradients of town and countryside, already outlined (Dobson 1987). But it also highlights the contrary nature of some of the epidemics of this period; epidemics that appeared to have moved against the usual contours and gradients of space and time. The last part of this paper looks not at the familiar trajectories of plague, smallpox and fevers—the chronology of epidemic disease as well as the following quotations are presented and cited elsewhere (Dobson 1987)—but at some of the contradictions and perplexities of the epidemiological landscape that finally take us beyond the traditional bounds of 'airs, waters and places'.

The year 1638 stands out as the time of the very worst mortality crisis on the Southeast England chronology of disease and death. John Graunt, the seventeenth-century pioneer of epidemiology, also found that this was the most 'mortal' year in his Hampshire parish between 1570 and 1660. The exact nature of the epidemic remains puzzling: it was a 'malignant fever ... which raged so fiercely about harvest that there appeared scarce hands enough to take in the corn'. It was widespread in the sparsely populated agricultural regions of Kent and Sussex and, ironically, it was in Wadhurst in the Weald of Sussex that it was reported: 'this yeare was an infectious summer, so that verie many died in many places here in Sussex also speciallie on the Downs'. Another exceptionally sickly time for Southeast

England was the fever years of 1657 and 1658. Many were ill 'in their brain and nervous stock', all complained of their head being 'grievously distempered', in some 'little broad and red spots' appeared and then disappeared, followed by 'a benumbedness of the senses and a sleepiness'. Again, it was the geographical peculiarity of this 'new' fever which struck observers, such that by August 1657

it began to spread far and near, among the people, that in every region and village many were sick of it, but it was much more frequent in the country, and smaller villages than in cities or towns.

At the other end of our period, the years 1779–1780–1781 mark the final major regional mortality peak of the two centuries. The epidemic ague, the 'new' ague or the 'plague ague', as it was called in Kent, was a prolonged, widespread and peculiarly protean fever epidemic. The epidemic was said to harass the upland villages more than communities in adjacent valleys and to afflict all male labourers in the fields, while leaving women nearly exempt. In the downland areas of east Sussex, almost one half of the parishes experienced a rise in burial levels. Deaths in the little village of Patcham, Sussex trebled and a note in the parish register recalls: 'this year was remarkable for a violent distemper which carried off the person afflicted in the space of five days ...'. Places which might normally expect to enjoy low levels of mortality and long life expectancies could still be suddenly and tragically hit by epidemic disease.

Epidemics of influenza were widespread; they visited places simultaneously, regardless of topography, terrain or economy. This was an infection which could invade large areas in a dramatically short space of time. In the spring of 1658 influenza was universal and prevalent in many parts of the world, such that one reporter believed that 'a third of mankind, almost should be distempered with the same in the space of a month'. Indeed, so suddenly did this distemper arise that it was 'as if sent by some blast of the stars, which laid hold on very many together, that in some towns, in the space of a week above a thousand fell sick together'. The influenza epidemic of the winter of 1732–33, likewise, was later described as 'the most universal disease on record'. It visited every country in Europe and raged in America and the Caribbean: 'the uniformity of the symptoms of the disease in every place was most remarkable'. With such a striking global epidemiology, it is not surprising that observers looked beyond atmospheric vapours or person-to-person contagion to some ethereal or extraterrestrial influence, to comets and meteors, a theory that even today finds favour with some scholars (Hoyle and Wickramasinghe 1979).

Plague epidemics showed a striking preference for large urban centres (Slack 1985), but isolated outbreaks, sometimes of severe intensity, were recorded in a number of country parishes during the seventeenth century. A church memorial in the tiny downland Sussex parish of Pyecombe commemorates its alarming visitation in 1603 when 15 per cent of the population died. In some places, plague was simply confined to a few households. The tragic story of the Gale family of the small hamlet of Kemsing in 1636–37, as related by a sole survivor some years later, shows how plague could afflict a whole family but spread no further through the settlement. On this occasion, several women 'laid forth' the dead, 'no manner of clothes were taken out of the chamber', 'a great many people visited the house', and yet 'all this while no one took the distemper of or from us'. In other localities, measures would be taken 'to use the best meanes we can, both to God, and by all outward instrumentall meanes as shall be fittinge' to avoid a visitation of plague; some communities took serious quarantine action by segregating houses, burning contents, removing occupants to a pest house and restricting population movements and, yet, these parishes would subsequently be affected by the plague. Bubonic plague was nearly always associated with hot summers in England, but in the southeast one of the most severe visitations of plague occurred during 1625–26, a time of 'unseasonably' cold, wet springs and summers.

Again, it is ironic that the final outbreak of plague in seventeenth century Southeast England occurred in the healthy High Weald in the community of Biddenden where in June 1667 '12 were buried at Betnams Wood of the plague' and '12 more had plague sores which recovered'. The distribution and fluctuations of plague in seventeenth century Southeast England displayed a random quality which added terror to its frightening and unexpected impact.

Other diseases could appear 'to reign everywhere spread far and wide take hold of whole households', and yet make little inroad into the mortality statistics. Meteorological signs might presage some imminent disaster: predictions, based on past experience, that 'our beautiful kingdom will ingender more strange and incurable diseases and infecte the whole nation', but the ensuing season could remain free from any major onslaught. At other times epidemics would arrive 'suddenly, unlooked for and unawares' cutting off the unsuspecting. Some would 'creep from house to house, infecting with the same evil most of the same family'. Smallpox epidemics could be 'confluent of the worst kind', 'never in the memory of man so fatal' and, on other occasions, be 'mild and seldom fatal'. Short noted that 'there is no general constitution of weather wherein the smallpox are not epidemic somewhere'. Some epidemics would be fatal to 'the aged, the weak, the consumptive, the gross bodied, the infirm' or those 'afflicted with or lately recovered from intermittents', others would affect those 'familiarily conversing with the sick', some would decimate those 'in their prime of life' or rage 'mostly among children and youths'. Even marsh fever was selective in its attack: 'it is far from being mortal to natives but to strangers and to persons accustomed to a pure air it proves particularly severe, and sometimes fatal' while, similarly 'in persons of an ill habit of body it often proves very dangerous'. Epidemics could run their course in a matter of weeks or months or linger for several years. The ebb and flow of epidemic diseases, their waxing and waning over time and space, produced patterns which went beyond natural or environmental boundaries, fluctuations that did not fit any neat meteorological associations, vicissitudes that proved unexpected and unexplained.

The chronology of epidemic disease and mortality for Southeast England highlights the range and variability of the many disorders that afflicted the early modern world. In the epidemiological landscapes of Southeast England, few diseases beyond malaria were entirely confined to specific localities and few people were entirely isolated by geography from epidemic visitations. The role of population movements and their simultaneous passage of vectors, pathogens, carriers and victims, explains many of the dynamic and elusive patterns of infectious disease outbreaks. We cannot track or quantify with any precision the intricate and diverse paths of all these interactions. But even for those who eked out their lives in the more isolated reaches of Southeast England there must have been a certain amount of coming and going: trading, conversing, or socializing with people from other parts of the countryside. There was a world which was apart but never entirely cut off from the epidemiological linkages and flows of the early modern world.

Below the surface channels of local, seasonal and annual migratory movements, other aspects of human populations and their diseases added to the final complexity of epidemiological events. We can, thus, draw on our knowledge of the peculiarities, the distinctive patterns and different modes of transmission of individual diseases; we can raise questions about the movements of disease vectors, or changing virulence of infections over time; we can address such issues as biological defences, immunity levels, genetic predisposition, age and gender differences, varying responses in accordance with nutritional status, encounters with previous or concurrent infections; we can look at the effectiveness of measures to avoid, contain, cure or quarantine diseased individuals and localities in past times. All of these factors must have played important and complicating roles in the passage and paths of human disease and mortality (Slack 1985; Landers 1990; Hope-Simpson 1992). Whether we are dealing with diseases with strong environmental associations, like malaria (and cholera in the nineteenth century) or

epidemics, such as smallpox, which had a less striking seasonal and spatial patterning, the final outcome, the recreation of the epidemiological maps and chronologies of the past, must inevitably reflect the intricate and complex nature of human disease.

Conclusion: from medical confusions to epidemiological complexities

While seventeenth and eighteenth-century physicians searched for associations with ‘airs, waters and places’ or extended their lists to explore a whole range of environmental, social, economic and epidemiological circumstances and events, they also included in their range of possible influences such attributes as the ‘constitution’, the ‘pre-disposition’ or ‘peculiar nature’ of individuals, the relative ‘contagiousness’ of different diseases, the ‘benefit’ or ‘danger’ of certain treatments. Today, historical epidemiologists can refine or reinterpret these ideas and speculate, for each individual epidemic, the likely interplay of the many different variables, concerning both places and peoples, which might have accounted for the varied and diverse patterns of epidemic diseases in the past. In formulating our ideas and explanations, we can move beyond our predecessors to imagine a world where microbes, vectors and agents of disease transmission, population movements and behaviour, age and sex structures, human biology, maternal and infant care, individual circumstances and environmental conditions interlocked and interacted in all sorts of complex ways (Landers 1990). We cannot always account for individual idiosyncracies in the outcome of disease, but we can explore such questions as the role of immunity, nutritional status, multiple infections and infant feeding practices on the patterns of health and sickness of past populations. We can emphasize complexity where early writers saw confusion.

In extending the field of historical epidemiology, there are all sorts of questions, patterns and findings which still remain to be understood. A broad multivariate approach, which takes into account local variations, seasonal fluctuations, epidemiological subtleties, and human variability will prove more revealing, even if more taxing, than analyses of aggregate series or simple bivariate statistical correlations. This present study has already illustrated the ease with which certain patterns can be explained and the difficulties of explaining others. Many puzzles will remain and many influences will defy precise quantitative measurement. In a sequel to this paper, secular trends in mortality across different parts of Southeast England will be explored in a way that, once more, highlights and contrasts the apparent simplicity of avoiding or removing ‘bad airs’ and ‘bad waters’ in ‘places’ like the English marshes with the far more enigmatic role of other environmental strategies devised to ‘teach men to choose their dwellings for their better health’. In each of these dimensions, it is now time to ‘move beyond the real wage’, even to move beyond ‘airs, waters and places’, to avoid the temptation to search for any one single determinant of mortality patterns and changes over time, as McKeown and other scholars have done, and to explore, instead, a whole set of environmental, social, economic, biological and demographic changes: that complex and confusing constellation of associations that absorbed but perplexed the early modern physicians and medical topographers.

References

- Appleby, A.B. 1978. *Famine in Tudor and Stuart England*. Liverpool: Stanford University Press.
- Cassedy, J. 1974. Medicine and the rise of statistics. Pp 283–312 in *Medicine in Seventeenth Century England*, ed. Allen G. Debus. Berkeley: University of California Press.
- Chambers, J.D. 1972. *Population, Economy, and Society in Pre-industrial England*. Oxford: Oxford University Press.
- Debus, A.G. ed. 1974. *Medicine in Seventeenth Century England*. Berkeley: University of California Press.
- Dobson, M.J. 1980. ‘Marsh fever’: a geography of malaria in England. *Journal of Historical Geography* 6:357–389.

- Dobson, M.J. 1982. Population, disease and mortality in Southeast England, 1600-1800, D.Phil thesis, Oxford University.
- Dobson, M.J. 1987. *A Chronology of Epidemic Disease and Mortality in Southeast England, 1600-1800*. Cheltenham: Institute of British Geographers.
- Dobson, M.J. 1989a. History of malaria in England. *Journal of the Royal Society of Medicine* 82:3-7.
- Dobson, M.J. 1989b. The last hiccup of the old demographic regime: population stagnation and decline in late seventeenth and early eighteenth century south-east England. *Continuity and Change* 4:395-428.
- Dobson, M.J. 1989c. Mortality gradients and disease exchanges: comparisons from old England and colonial America. *Journal of the Social History of Medicine* 2:259-298.
- Dobson, M.J. forthcoming. *Contours of Death, Contours of Health. Death and Disease in Early Modern Southeast England*. Cambridge: Cambridge University Press.
- Forbes, J. 1836. On ague. *Transactions of the Provincial Medical Association* 4.
- George, M.D. 1966 [1925]. *London Life in the Eighteenth Century*. London: Penguin (first published 1925).
- Hamlin, C. 1992. Predisposing causes and public health in early nineteenth-century medical thought. *Journal of the Social History of Medicine* 5:43-70.
- Hasted, E. 1797-1801. *History and Topographical Survey of the County of Kent*. Canterbury: W. Bristow.
- Hope-Simpson, R.E. 1992. *The Transmission of Epidemic Influenza*. New York: Plenum Press.
- Hoyle, F. and Wickramasinghe, C. 1979. *Diseases from Space*. London: Dent.
- Jordanova, L.J. 1979. Earth science and environmental medicine: the synthesis of the late Enlightenment. Pp 119-146 in *Images of the Earth. Essays in The History of the Environmental Sciences*, eds. L.J. Jordanova and R.S. Porter. Chalfont St.Giles: British Society for the History of Science.
- Jordanova, L.J. and R.S. Porter, eds. 1979. *Images of the Earth. Essays in the History of the Environmental Sciences*. Chalfont St. Giles: British Society for the History of Science.
- Kussmaul, A. 1981. *Servants in Husbandry in Early Modern England*. Cambridge: Cambridge University Press.
- Landers, J. 1986. Mortality, weather and prices in London 1675-1825: a study of short-term fluctuations. *Journal of Historical Geography* 12:347-364
- Landers, J. 1987. Mortality and metropolis: the case of London 1625-1825. *Population Studies* 41:59-76
- Landers, J. 1990. Age patterns of mortality in London during the 'Long Eighteenth Century': a test of the 'high potential' method of metropolitan mortality. *Journal of the Social History of Medicine* 3:27-60.
- Lee, R.D. 1981. Short-term variations in vital rates, prices, and weather. Pp. 356-401 in *The Population History Of England 1541-1871: A Reconstruction*, eds. E.A. Wrigley and R.S. Schofield. London: Edward Arnold.
- Lee, R.D. 1985. Population homeostasis and English demographic history. *Journal of Interdisciplinary History* 15:635-660.
- Le Roy Ladurie, E. 1978. Un concept de l'unification microbienne du monde (xive-xviiie siècles). In E. Le Roy Ladurie, *Le Territoire de l'Historien*. Paris: Gallimard.
- Malthus, T.R. 1986 [1798]. *An Essay on the Principle of Population*. Reprinted in *The Works of Thomas Robert Malthus* 8 vols., eds. E.A. Wrigley and D.C. Souden. London: W. Pickering (first published 1798).
- Marshall, Mr. 1798. *The Rural Economy of the Southern Counties*. 2 vols. London: G. Nicol.
- McKeown, T.R. 1976. *The Modern Rise of Population*. London: Edward Arnold.
- Muilman, P. 1769-72. *A New and Complete History of Essex*. 6 vols. Chelmsford: L. Hassall.
- Parkinson, R. 1811. *Agriculture of the County of Huntingdon*. London: Board of Agriculture.
- Riley, J.C. 1987. *The Eighteenth Century Campaign to Avoid Disease*. Basingstoke: Macmillan.

- Sargent, F. 1982. *Hippocratic Heritage. A History of Ideas about Weather and Human Health*. New York: Pergamon Press.
- Short, T. 1973 [1750]. *New Observations on City, Town and Country Bills of Mortality*. Reprint with introduction by R. Wall. London: Longman (first published 1750).
- Slack, P. 1985. *The Impact of Plague in Tudor and Stuart England*. London: Routledge, Kegan Paul.
- Southall, J. 1730. *A Treatise on Bugges*. London.
- Sydenham, T. 1848-50. *The Works of Thomas Sydenham*, 2 vols. Translated by R.G. Latham. London: Sydenham Society.
- Walter, J. and R.S. Schofield. 1989. *Famine, Disease and the Social Order in Early Modern Society*. Cambridge: Cambridge University Press.
- Woods, R. and J. Woodward. 1984. *Urban Disease and Mortality in Nineteenth Century England*. London: Batsford Academic and Educational.
- Wrigley, E.A. and R.S. Schofield. 1981. *The Population History of England 1541-1871: A Reconstruction*. London: Edward Arnold.

Influenza in Sri Lanka, 1918–1919: the impact of a new disease in a pre-modern Third World setting*



C.M. Langford and P. Storey

Department of Population Studies, London School of Economics, Aldwych, London WC2A 2AE, UK

Abstract

In 1918–1919 there was an unusually serious influenza pandemic. The main object in this paper is to establish the course and impact of this outbreak in Sri Lanka using census and registration data. Influenza probably entered the country through the port of Colombo and possibly also through the port of Talaimannar. As elsewhere there was a mild first wave followed by a virulent second wave characterized by fatal pneumonic complications. Women suffered heavier mortality than men and young adults more (relatively) than other age groups. Fertility fell. Probably about 1.1 per cent of the population died.

Introduction

There had almost certainly been many influenza pandemics before that of 1918–1919: Patterson (1986:83) lists nine between 1700 and 1900. Usually, though, these were fairly benign: a very large proportion of the population tended to catch the disease but very few to die from it. Moreover, where there were deaths these tended to be among the elderly or the very young. What marked the 1918–1919 pandemic from earlier (and later) pandemics was its terrible virulence: one authority (Burnet 1979) has even suggested that influenza was probably responsible for 50–100 million deaths worldwide at this time. The lower figure of this range is certainly not difficult to believe: Davis (1951:237), for example, has estimated that there were about 20 million deaths in India alone. Another remarkable feature of the 1918–1919 pandemic was its tendency to kill disproportionately those in the prime of life rather than the elderly or the very young: one of the contributors to the official British report on the outbreak noted that mortality seemed to be concentrated among those aged 20–40 and especially those aged 25–35 (French 1920:90). Thus the influenza pandemic of 1918–1919 tends to invite comparison with such other great historical pestilences as the Black Death in the fourteenth century and the Plague of Justinian in the sixth century, many accounts ranking it third in mortality terms after these two but some even putting it in second place. (See, for example, Ministry of Health 1920:182; Cliff, Haggett and Ord 1986:1; Patterson 1986:1.)

Accounts of the 1918–1919 outbreak usually speak of two and often three distinct waves of influenza. The first wave was in spring-summer of 1918 and was apparently fairly mild; the second was in autumn-winter of 1918 and was very far from mild, showing a terrible propensity to lead on to pneumonic complications and death. According to French (1920:69) in his contribution to the official British report on the outbreak, and obviously talking in round numbers, 20 per cent of those contracting the disease developed pneumonic complications and eight out of these 20 per cent died. The third wave, where there was one, tended to come in the early part of 1919; it involved the same serious form of the disease as the second wave but its overall impact was much less.

*This paper is based upon work carried out with the support of a grant from the Wellcome Trust.

According to Crosby (1976) the first wave of the 1918–1919 pandemic began in the United States in March 1918 and then spread around the world over a four-month period. Most other accounts essentially agree, though some point out that the disease may have arisen elsewhere but not been especially noticed. Beveridge (1977) suggests that the first wave of the pandemic may have begun in the United States or it may have begun in China. It apparently struck Britain in June and July (Ministry of Health 1920) and reached Bombay in June 1918 (Mills 1989). The second wave has been seen by some observers as a puzzle, in that it apparently arose in more than one place at the same time. Crosby (1976) reports that the first outbreaks of this more serious form of influenza occurred in the last week of August 1918 in Brest in France, in Boston in the United States and in Freetown in Sierra Leone. This apparent observation thus presents the difficulty that either the more virulent form of the disease arose or developed from the mild form independently but at much the same time in a number of different places around the world, or it travelled at faster than human speed between them; neither of which seems very likely. Hoyle and Wickramasinghe (1977, 1978) have suggested, with reference to this and other outbreaks, that the scattering of infective debris from comets might be responsible, thus producing simultaneity. However, by no means everyone accepts that there were such simultaneous outbreaks. Beveridge (1977) suggests that the severe second wave may have evolved from the first or resulted from the invasion of a new virus from Russia or Africa. Stuart-Harris, Schild and Oxford (1985) simply mention West Africa as the starting point of the second wave of the disease. In any case mortality in this more virulent second wave of the pandemic reached a peak in October 1918 in the United States, in October or November 1918 in different parts of Europe, and in November 1918 in most of India, though the peak was in October in Bombay Presidency and in December in Bengal (see Ministry of Health 1920; Crosby 1976; Mills 1989).

The intention in this paper is to trace the course and assess the impact of the 1918–1919 influenza outbreak in Sri Lanka (then Ceylon). This effectively means the second wave of the disease since it is possible to examine data only on mortality and not on morbidity. There is certainly a need for such an account since although a number of assessments of the impact of influenza in 1918–1919 on various developed countries are available, almost no detailed studies of Third World countries have been carried out, virtually the only exception being Mills's (1989) account for India; of course in many cases there are few or no data which would permit this. The data used in this paper come from vital registration and the census and have been taken from the *Reports of the Registrar General of Ceylon on Vital Statistics*, from issues of the *Ceylon Government Gazette* and from census reports. There are no doubt errors in these data, both of coverage and of accuracy. However, if they are used with care, a great deal of information can be had from them.

In the next section of the paper a brief account is given of the geography of Sri Lanka as well as of certain features of the demographic and the economic situation in the country in the early part of the twentieth century, in order to provide some indication of the context in which the 1918–1919 influenza outbreak occurred. In subsequent sections of the paper the influenza outbreak itself is examined in some detail.

The geography, demography and economy of Sri Lanka in the early twentieth century

Sri Lanka is an island not far north of the equator (5°55'–9°50'N), just off the south-east tip of India. It is about 140 miles across at its widest point and 270 miles from north to south. The south-central part of the island is mountainous, ranging from about 1000 feet to more than 7000 feet above sea level: this is where most tea is grown. The south-western part of the country is well watered (the wet zone); the remainder, the dry zone, is not (see Figure 1 and Table 6). Historically, that is until the late 1940s,

malaria was especially important in morbidity and mortality in Sri Lanka: this was particularly true of the dry zone in most of which malaria was endemic. Most of the population, 63 per cent of the approximately 4.1 million people in 1911, lived in the wet zone. In the early part of the twentieth century, as indeed now, the Sri Lankan economy was heavily dependent on the output of the estate (plantation) sector, principally tea but also rubber and coconuts. Sri Lanka imported most of its food at this time.

Basic demographic measures for Sri Lanka in the early part of the twentieth century, and for some later years for comparison, are shown in Table 1. Birth and death rates were computed in every case for three-year periods centred on census years. The overall picture seems to be one of no secular trend at all in the crude birth rate over the entire period 1901–1953 and of a decline in the crude death rate only after 1921. The data are consistent with the idea, in other words, that in the first two decades of the twentieth century in Sri Lanka there was a ‘pre-demographic transition’ situation of more or less constant fertility and mortality. The small differences between the figures for 1911 (in fact 1910–1912) and the other years are probably genuine. There was a malaria epidemic in 1911–1912, so it is not at all surprising that both the crude death rate and the infant mortality rate were highest at this time. Nor is it surprising that the crude birth rate was lowest around 1911: an upsurge in mortality is commonly followed by a temporary downturn in fertility. In this case an examination of the yearly totals of registered live births for successive years clearly indicates a marked fall in 1912.

Table 1

Crude birth rates, crude death rates and infant mortality rates for Sri Lanka around census years, and census populations, 1901–1953

	1901	1911	1921	1931	1946	1953
Crude birth rates	38.5	36.7	39.1	37.9	38.4	38.6
Crude death rates	28.0	31.5	29.7	22.7	18.9	11.2
Infant mortality rates	173	202	188	165	126	74
Population	3,566	4,106	4,499	5,307	6,657	8,098

Note: crude birth and death rates were calculated by dividing one-third of the number of births or deaths in the three-year period centred on the census year by the census population and infant mortality rates by dividing infant deaths in the three-year period by live births in the same period; crude birth and death rates are shown per 1000 total population and infant mortality rates per 1000 live births; census populations are shown in thousands.

The appearance of influenza in Sri Lanka in 1918

In the 18 years between 1900 and 1917 the maximum number of deaths registered as due to influenza in any one year in Sri Lanka was 256 in 1909; the minimum number was 44 in 1903; the average number of deaths from influenza per year in this period was 115. In 1918, on the other hand, the number of deaths registered as due to influenza in Sri Lanka was 19,102 of which, moreover, 18,887 were recorded in the last quarter of 1918. Clearly, the 1918–1919 influenza pandemic had reached Sri Lanka.

According to the Principal Civil Medical Officer of Ceylon in his report for the year 1918 (Ceylon 1919:Part IV, B1-B14) the first cases of influenza appeared in June of 1918 in Colombo, the capital city and main port, among harbour workers, and the disease spread from there.

The spread of the disease was rapid, and the gravity of the symptoms increased as the disease increased. By September and October nearly every Province and district in the

Island was affected. Notable features of the disease were the rapid onset of pneumonia in a large percentage of cases, mostly of the broncho-pneumonia type (Ceylon 1919:Part IV, B2)

Despite the absence in this account of any clear reference to mild and virulent waves of influenza there may well have been such waves. According to the Medical Officer the initial epidemic in Colombo was in June and July; however the data indicate that mortality did not rise there appreciably until September. Moreover, the Government Agent for the Northern Province, in his report for the district of Jaffna for the year 1918, wrote:

There was a wide prevalence of influenza in the district in common with the rest of the Island, or rather with the whole world. It swept over the country in two waves: one mild, and the other a severe outbreak, attended with high mortality. The first occurred about the middle of August, and appeared to be dying out towards the end of September, when it re-appeared in greater virulence and spread everywhere (Ceylon 1919:Part I, D4).

The total numbers of deaths registered each year in Sri Lanka, from all causes, for years between 1900 and 1925 are shown in Table 2. It may be seen that 1918 was the second worst and 1919 the worst year in mortality terms during the whole of this period. The worst years previously had been 1906, 1911 and 1914, all years in which serious malaria outbreaks had occurred. A more detailed picture is provided by the data presented in Table 3 on the total deaths registered in each quarter of 1918 and 1919. It may be seen that the peak of mortality was in fact in the last quarter of 1918, though the first quarter of 1919 was not very far behind; 1919 as a whole was a worse year than 1918 because not only the first quarter in 1919 but also the second and third quarters had exceptionally high mortality. Just how serious was the mortality in the last quarter of 1918 and the first quarter of 1919 may be seen from the fact that these were, respectively, the worst and second worst quarters in mortality terms over the whole period 1900–1925; moreover the next worst quarter, the third quarter of 1911, had only two-thirds the number of deaths in the last quarter of 1918.

Table 2

Deaths registered each year 1900–1925 in Sri Lanka, and indices showing the relative mortality in different years

Year	Total deaths	Relative to 1919 = 100	Rank order of year ^a
1900	100,873	60	22
1901	98,813	59	24
1902	99,680	59	23
1903	96,084	57	25
1904	93,940	56	26
1905	108,160	64	21
1906	136,271	81	7
1907	119,377	71	15
1908	117,982	70	16
1909	122,969	73	11
1910	110,195	65	19
1911	143,380	85	3
1912	134,383	80	8
1913	119,956	71	14
1914	136,831	81	6
1915	109,818	65	20

1916	120,162	71	13
1917	113,389	67	18
1918	149,407	89	2
1919	168,323	100	1
1920	132,955	79	9
1921	140,749	84	5
1922	126,820	75	10
1923	141,891	84	4
1924	122,958	73	12
1925	117,543	70	17

^aYear with most deaths 1900–1925 = 1, year with fewest deaths = 26.

That the upsurge in mortality in Sri Lanka during 1918–1919 was largely due to the influenza outbreak is very clear from the data on causes of death shown in Table 3. Not only were there many deaths during this period from influenza and—almost a defining characteristic of the 1918–1919 outbreak—from pneumonia, but there were also more deaths than usual from respiratory diseases other than pneumonia and it seems quite likely that this rise may also have been influenza-related. It may be seen, moreover, that there was an increase in deaths from pyrexia, meaning pyrexia (fever) of unknown origin, during this time. Although in the Sri Lankan context such deaths are usually taken to indicate deaths from malaria and there was indeed a malaria outbreak in some areas of Sri Lanka during this period, it seems extremely likely nevertheless that many of the deaths ascribed to pyrexia during 1918–1919 were influenza deaths.

A more detailed picture of the timing of the 1918–1919 influenza epidemic in Sri Lanka is provided by the data shown in Table 4 on the number of deaths registered each month during this time. It may be seen that mortality first rose, albeit only slightly, in September 1918 and that this was then followed by a dramatic rise in October. Mortality reached a peak in November 1918, when the number of deaths registered was almost three times the average for November during 1915–1917, and then declined, remaining, however, at very high levels through the first three months of 1919 before dropping back to still-above-average levels at the end of the year.

Table 3

Total deaths and deaths from particular causes registered in each quarter of 1918 and 1919 in Sri Lanka, and indices showing the relative mortality in these quarters compared with all quarters 1900–1925 or 1900–1921

	Year and quarter							
	Jan– Mar.	Apr– June	1918 July– Sept.	Oct– Dec.	Jan– Mar.	Apr– June	1919 July– Sept.	Oct– Dec.
Total deaths	30,137	23,929	27,027	68,314	58,522	40,227	37,512	32,062
Relative to worst quarter 1900– 1925 = 100	44	35	40	100	86	59	55	47
Quarter rank 1900–1925 ^a	43	94	71	1	2	9	12	30
Influenza deaths	32	20	163	18,887	12,324	5,498	3,897	1,095
Relative to worst quarter 1900– 1921 = 100	0	0	1	100	65	29	21	6
Quarter rank 1900–1921 ^b	34	71	14	1	2	3	4	6
Pneumonia ^c deaths	1,239	1,089	1,886	12,556	2,834	2,702	3,419	1,797

Relative to worst quarter 1900–1921 = 100	10	9	15	100	23	22	27	14
Quarter rank 1900–1921 ^b	20	37	8	1	3	4	2	9
Malaria ^d and pyrexia deaths	4,074	3,457	3,951	8,011	8,917	6,340	5,136	4,034
Relative to worst quarter 1900–1921 = 100	32	27	31	63	70	50	40	32
Quarter rank 1900–1921 ^b	55	79	62	7	5	19	31	58
Deaths from:								
all respiratory diseases ^e (excluding pneumonia)	1,099	842	1,086	2,236	1,817	1,365	1,368	1,010
pyrexia	3,784	3,209	3,647	7,728	8,357	5,943	4,836	3,736
malaria ^d	290	248	304	283	560	397	300	298
unspecified causes (excluding pyrexia)	1,232	1,009	1,091	1,358	1,749	1,616	1,613	1,506

^aQuarter with most deaths = 1, quarter with fewest deaths = 104

^bQuarter with fewest deaths = 88

^cIncluding broncho-pneumonia

^dIncluding malarial cachexia

^eThe 'respiratory diseases' category does not include influenza

The suggestion that mortality rose in September 1918 as a result of the influenza outbreak is supported by the monthly figures on deaths by cause, available only for certain causes of death for Sri Lanka as a whole. In August 1918 in Sri Lanka only eight deaths were registered as due to influenza; in September this rose to 145. Moreover, the number of deaths recorded as due to pneumonia or broncho-pneumonia in September was 977, up from 500 in August and 409 in July, and an average of 388 per month in the first six months of the year. The suggestion that mortality reached a peak in November 1918 because of the influenza outbreak is also supported by the monthly figures on causes of death. There was a very marked peak in November 1918 in the number of deaths registered as due to influenza (8,253 deaths) and a very marked peak also in the number recorded as due to pneumonia or broncho-pneumonia (6,082 deaths).

Table 4

Deaths registered each month in 1918 and 1919 in Sri Lanka, and indices showing deaths each month in relation to the average number in the same month during 1915–1917

Month	Deaths	1918 Relative to 1915–1917 = 100	Deaths	1919 Relative to 1915–1917 = 100
January	11,243	99	21,124	187
February	9,987	100	20,097	201
March	8,907	92	17,301	179
April	7,870	90	12,423	142
May	8,166	84	13,325	137
June	7,893	85	14,479	155
July	8,674	90	13,968	145
August	9,000	94	12,103	126
September	9,353	109	11,441	133
October	23,453	258	10,724	118

November	26,231	286	10,223	111
December	18,630	195	11,115	116

The spread of the disease through the island

The numbers of deaths registered each month in the 21 administrative districts of Sri Lanka during 1918 and 1919, compared with the average number of deaths in the same month during 1915–1917, are shown in Table A1 in the appendix. Certain features of these data are also indicated on the maps presented as Figure 2 which are intended to show the progress of the disease through the island. Names of districts are indicated on the map presented as Figure 1.

In Colombo, Kalutara, Kandy, Kegalle, Mannar, Mullaittivu, Puttalam and Ratnapura mortality rose in September 1918 whereas in the other districts mortality did not increase until October. What this amounted to in geographical terms may be seen from Figure 2, part (A). The districts affected earliest by the epidemic were apparently in two clusters, one around Colombo in the south-west of the island and the other in the north, though not including the northernmost district of all, Jaffna.

The further progress of the disease through the island may be seen from Figure 2, part (B), which indicates for every district the worst month in mortality terms during 1918. Mullaittivu with joint peaks in October and November has been counted as October and Kurunegala with joint peaks in November and December has been counted as November. In twelve districts the peak month for mortality in 1918 was October and in eight districts it was November; in Batticaloa it was December. It may be seen that, with the exception only of Kandy, all of the districts affected earliest by the epidemic, that is, where mortality rose in September, also experienced an early mortality peak, in October. Moreover, other districts which experienced an October mortality peak were all districts adjoining these. The evidence suggests, in other words, that influenza spread out in Sri Lanka from Colombo and from some other point further north, possibly having arrived there from Colombo or possibly from elsewhere. Initially, coastal districts in the north, north-east and west and districts in the south-west near to Colombo were affected; subsequently the disease spread to the interior of the island and to districts in the south; finally it reached Batticaloa district on the east coast of the island.

Figure 1
The districts of Sri Lanka

Figure 2

The timing of the rises in mortality in the different districts of Sri Lanka during 1918–1919

Note: These maps are based upon the data presented in Table A1 which show the number of deaths each month relative to the average number in the same month during 1915–1917; names of districts are shown in Figure 1.

That Colombo and some other districts nearby should have been affected very early in the epidemic is hardly surprising and is consistent with the suggestion of the Principal Civil Medical Officer of Ceylon already cited that influenza entered the country through the port of Colombo. As for

the other cluster of districts, some doubt attaches to the observation since each of the districts in question had a rather small population: all three districts had populations below 40,000 and Mullaittivu had a population of only about 18,000. However, there are reasons for suspecting that the observation might be genuine. The port of Talaimannar in Mannar district was the Sri Lankan end of the sea crossing between Sri Lanka and India which linked the railway systems of the two countries. In 1918 in Sri Lanka, Talaimannar was second only to Colombo in terms both of the number of vessels using the port and the total size of their crews (Ceylon 1919:Part II, A42). Not far short of 200,000 people passed through Mandapam quarantine camp in south India during 1918 en route to or from Sri Lanka via Talaimannar (Ceylon 1919:Part I, N8). Trincomalee on the north-east coast was not at this time in use as a military or naval base (Ceylon 1920:Part I, E12).¹

The information presented in Figure 2, part (C), points to another feature of the situation. Although in most districts the worst month in mortality terms during the whole of 1918–1919 was in fact in 1918, in a few districts the peak month was in 1919. These were Anuradhapura (peak in February), Batticaloa (January–February), Chilaw (March), Kurunegala (February), Negombo (February, June) and Puttalam (February). In these districts the interval between the first appearance of influenza as indicated by rising mortality and the worst month for mortality during 1918–1919 was typically four or five months (disregarding for the moment the second 1919 peak in Negombo district which will be discussed subsequently) compared with, usually, one month or less in the other districts. Was it not, then, the case that influenza simply moved more slowly in some districts than others so that the height of its impact tended to come rather later, moreover in some cases after an earlier peak had apparently been reached? This may have been so; however the situation seems also to have been complicated by the fact that there was a serious outbreak of malaria in some districts at this time.

The report of the Principal Civil Medical Officer of Ceylon for 1918 refers to there having been only one ‘extensive epidemic’ of malaria in Sri Lanka during 1918, that is, in the North-Western Province comprising Kurunegala, Puttalam and Chilaw districts (Ceylon 1919:Part IV, B2). The seriousness of this outbreak is confirmed by the local official reports. The Government Agent for the North-Western Province in his report for 1918 wrote (with reference to Kurunegala district):

About the end of November the epidemic of influenza abated, but about the same time a severe outbreak of malarial fever occurred, attended again with unusual mortality. The explanation probably is that the victims were weakened by a previous attack of influenza.

The report for Puttalam and Chilaw districts for 1918 referred to malaria having ‘set in in an epidemic form with unwonted severity’ and went on ‘Many cases of malaria were complicated by pneumonic sequelae, probably due to influenza. By the end of the year almost every part of the district had begun to suffer’ (Ceylon 1919:Part I, F2, F12). It is perfectly clear, moreover, from the official reports that the malaria epidemic persisted in these districts through the early part of 1919 (Ceylon 1919:Part I, F2, F10; Ceylon 1920:Part I, F2).

There seems to be no mention in the official reports of a similar outbreak of malaria in either Batticaloa district or in Negombo district and in the case of Anuradhapura district there are only vague hints of this possibility in the form of references to ‘influenza and fever’ and ‘fever and influenza’

¹ It might be feared that the apparent timing of the upsurge in mortality in Mannar district was due not to deaths in the local population but to deaths among those in transit. However, it is clear that this was not the case since the overwhelming majority of those passing through would have been Tamil estate labourers, yet the pattern of month-by-month changes in the numbers of registered deaths in Mannar district was much the same for Tamils and non-Tamils; moreover the upsurge in mortality was somewhat more marked for non-Tamils.

epidemics (Ceylon 1919:Part I, G1-G2). It seems quite likely, however, that Anuradhapura district was also affected: it adjoined districts of the North-Western Province; it had the same kind of climate (all these districts are in the so-called ‘dry zone’ of Sri Lanka); and the peak month for mortality was February 1919, just as in Kurunegala and Puttalam districts. Moreover, it will be seen that the overall rise in mortality in Anuradhapura district during 1918–1919 was second only to that in Kurunegala district.

Was the explanation, then, for monthly mortality reaching a peak in early 1919, at least for some districts, that as the influenza outbreak was beginning to subside it was succeeded by an outbreak of malaria, so that mortality rose still further? The material presented in Table A2 in the appendix suggests that it was not as simple as this. These data seem to show that just those districts (with the exception to some extent of Negombo) which reached a mortality peak in early 1919 also reached the high-point of influenza-related mortality in early 1919. In Anuradhapura, Batticaloa, Chilaw, Kurunegala and Puttalam districts the number of deaths registered as due either to influenza or pneumonia reached a peak in the first quarter of 1919 whereas in all other districts there was a peak in the last quarter of 1918. Moreover, in the case of Negombo district, though the high-point of mortality from influenza and pneumonia was in the last quarter of 1918 there were almost as many such deaths (93% of the number) in the first quarter of 1919.

It might be objected that at a time when influenza was prevalent there was a danger that malaria deaths would be misreported as influenza deaths. For this reason data on deaths from pneumonia and deaths from all respiratory diseases (which include pneumonia but do not include influenza) in these districts are presented in Table 5: these causes are likely to be influenza-related but are presumably not confusable with malaria. These data clearly indicate a 1919 peak in influenza-related mortality in Batticaloa district, where the high-point was apparently in the second quarter of the year, and in Chilaw district; and they suggest that there was probably a 1919 peak also in Anuradhapura district. In Kurunegala district and in Puttalam district the peak of influenza-related mortality was apparently in late 1918 though in neither case was the first quarter of 1919 far behind. The data for Negombo district are not very helpful.

In some of these districts, then, the high-point of influenza-related mortality was almost certainly in 1919 and it is possible, if the reporting of influenza deaths was valid, that there was a 1919 peak in virtually all of them, the possible exception being Negombo. Some of these districts were also badly affected by malaria during this time. It is possible that the high-point of influenza mortality was delayed until 1919 in some districts simply because the infection travelled more slowly in those districts, perhaps because the population was more scattered. Another possibility, however, which fits with the apparent observation that in some districts influenza seemed to be subsiding but then re-emerged, is that the mortality rate among those contracting influenza was actually worsened in some districts by the appearance of malaria. Observers at the time remarked that the malaria outbreak in the North-Western Province was particularly deadly because people were already weakened by influenza. It is perfectly possible that the reverse was also true: that those who contracted influenza who were already suffering from malaria, or had done so recently, were more likely than the average influenza sufferer to die from the disease.

Table 5
Deaths from pneumonia and from all respiratory diseases registered in the last quarter of 1918 and each quarter in 1919 in certain districts of Sri Lanka, and indices showing the relative mortality in these quarters compared with all quarters 1900–1921.

District

	Anura- dhapura		Batticaloa		Chilaw		Kurunegala		Negombo		Puttalam	
	N	(R)	N	(R)	N	(R)	N	(R)	N	(R)	N	(R)
Pneumonia												
4th quarter 1918	95	(1)	80	(21)	47	(2)	351	(1)	15	(50)	60	(3)
1st quarter 1919	59	(3)	98	(11)	77	(1)	242	(2)	29	(13)	59	(4)
2nd quarter 1919	38	(7)	201	(1)	34	(4)	148	(3)	18	(38)	63	(2)
3rd quarter 1919	14	(29)	119	(7)	18	(16)	95	(4)	17	(41)	24	(15)
4th quarter 1919	18	(16)	120	(6)	29	(7)	61	(10)	27	(18)	46	(6)
All respiratory diseases												
4th quarter 1918	211	(8)	289	(4)	102	(2)	551	(1)	38	(23)	140	(2)
1st quarter 1919	229	(3)	341	(2)	180	(1)	535	(2)	46	(15)	131	(3)
2nd quarter 1919	180	(13)	390	(1)	71	(13)	299	(9)	38	(23)	95	(4)
3rd quarter 1919	151	(26)	204	(8)	45	(57)	225	(27)	25	(52)	50	(25)
4th quarter 1919	111	(59)	165	(12)	61	(23)	146	(50)	41	(22)	64	(13)

Notes: N = number, R = quarter rank 1900–1921; quarter with most deaths = 1, quarter with fewest deaths = 88.
Pneumonia includes broncho-pneumonia.
Respiratory diseases include pneumonia (but do not include influenza).

The third wave of the epidemic

For some time after the 1918–1919 outbreak in Sri Lanka mortality from influenza and from pneumonia continued at much higher levels than had obtained before the epidemic. Even by 1925 the number of deaths registered as due to influenza (1,532) was still six times the number in the worst year before 1918 (since 1900). The number of deaths registered as due to pneumonia or broncho-pneumonia in 1925 (7,371) was 47 per cent higher than the number in the worst year before 1918; the excess varied between 33 per cent and 78 per cent over the period 1920–1925. In other words, there may well have been a number of further ‘waves’ of influenza following the first outbreak of the virulent form of the disease in 1918–1919. That being said, the monthly figures on deaths by cause for the whole island (these data are not available for districts) suggest that there was indeed a ‘third wave’ of the epidemic, in June and July 1919. The number of deaths registered as due to influenza each month increased to a maximum in November 1918 and declined thereafter. However, the number quite clearly rose again, though to nowhere near the original peak level, in June and July 1919. In the case of mortality from pneumonia and broncho-pneumonia there was a similar upsurge in June and July 1919 (beginning in May) though with another, even more marked, in September 1919.

The data on deaths from all causes each month by district shown in Table A1 in the appendix suggest that there may well have been a resurgence of influenza sometime in the period between May and August 1919 in the following districts: Badulla, Batticaloa, Colombo, Galle, Kalutara, Kandy, Kegalle, Matale, Matara, Negombo, Puttalam and Ratnapura. The same may have been true also of Chilaw, Hambantota, Mannar and Mullaittivu districts though the evidence is less clear in these cases. In Nuwara Eliya district there was a very sharp increase in mortality in September 1919. This was clearly associated with a resurgence of influenza since the third quarter of 1919 in this district was second only to the last quarter of 1918 (over the period 1900–1921) in both influenza and pneumonia plus broncho-pneumonia deaths. The upsurge in this district was undoubtedly largely responsible for the marked rise in pneumonia and broncho-pneumonia deaths in September in the island as a whole. Kandy district also showed a marked increase in mortality in September 1919, in addition to an earlier

peak in June. In the cases of Anuradhapura, Jaffna, Kurunegala and Trincomalee there seems to be no evidence of any ‘third wave’ of the epidemic, at least during 1919.

The impact of influenza on different subgroups in the population

Districts

Table 6 presents ratios (expressed as percentages) of observed deaths in 1918–1919 to average deaths in 1915–1917 in different districts: the impact of influenza varied substantially between districts. These variations seemed particularly large, judged on the basis of the mortality in the worst month during 1918–1919, which clearly may reflect not only seriousness of overall impact but also shortness-and-sharpness of attack. However, there were clear differences even when a longer time period is considered. The ratio of deaths observed during the worst month in 1918–1919 to average deaths in the same period during 1915–1917 varied from 190 per cent in Matara to 828 in Nuwara Eliya; the mean value was 387 per cent. For the whole six-month period comprising the last quarter of 1918 and the first quarter of 1919 these ratios ranged from 140 in Galle to 402 in Kurunegala with a mean of 231. For the period comprising the last quarter of 1918 and all of 1919 the ratios varied between 127 in Jaffna and 253 in Kurunegala with a mean of 173.

It is very difficult to say what may be the reasons for these differences between districts; however, a few points are probably worth noting. The first is that of course not all of these differences were due only to influenza. The districts showing the greatest excess mortality, on a six-month or a 15-month view of the data, were Kurunegala and Anuradhapura, in at least one of which and most probably both of which there was a serious outbreak of malaria during this period. Chilaw and Puttalam, which also suffered from malaria, were high up on the list, too, in terms of excess mortality. There is some suggestion in the data that in general (again over six or 15 months) districts in the dry zone may have tended to suffer more than those in the wet zone and districts where a substantial proportion of the population lived on estates (mainly tea estates) more than non-estate districts; the estates in Badulla and Matale would have been in non-dry parts of these districts. It is possible that the dry zone tended to suffer more because it was generally less healthy, less developed and less well provided with facilities than the wet zone and so may have experienced higher mortality from influenza. The dry zone was also much more prone to malaria than the wet zone and it is possible that this might have contributed to higher mortality. In the estate sector it is possible that congested living conditions led to higher mortality from influenza. The largely Indian Tamil labour force typically lived in close-packed ‘lines’ of small adjoining single-storey units. (On the other hand there is no real suggestion in the, admittedly rather simple, data of Table 6 that the ‘urbanness’ of districts was important.) It is also conceivable that cold might have been a factor on estates: tea is typically grown at high or relatively high altitudes, from about 1,000 feet to more than 6,000 feet above sea level. On the other hand it might simply be that the relatively poor state of general health of Indian Tamils (discussed below) predisposed them to higher mortality from influenza.

Table 6

The rises in mortality over different periods of time during 1918–1919 in Sri Lanka and districts, with some socioeconomic and climatic information about districts

District	Observed/expected ^a deaths (and rank ^b)			1921 %	Wet or dry zone d
	Worst month during 1918–1919	October 1918 – March 1919	October 1918 – December 1919		

	Urban ^c		Estates						
Colombo	262	(17)	153	(20)	138	(19) }			
							31.9	1.5	wet
Negombo	196	(20)	167	(17)	150	(17) }			
Kalutara	310	(15)	156	(19)	143	(18)	7.5	11.4	wet
Kandy	364	(12)	206	(14)	162	(13)	10.4	34.8	wet
Matale	427	(7)	283	(3)	188	(6)	6.7	29.9	dry
Nuwara Eliya	828	(1)	244	(10)	183	(7)	4.5	62.8	wet
Galle	213	(19)	140	(21)	132	(20)	13.5	3.6	wet
Matara	190	(21)	159	(18)	156	(14)	10.9	2.2	wet
Hambantota	370	(11)	236	(11)	175	(10)	6.5	0.2	dry
Jaffna	306	(16)	174	(16)	127	(21)	12.8	0.0	dry
Mannar	623	(2)	253	(7)	167	(12)	14.5	0.0	dry
Mullaittivu	389	(10)	276	(5)	202	(3)	13.4	0.4	dry
Batticaloa	256	(18)	213	(13)	176	(9)	8.3	0.5	dry
Trincomalee	391	(9)	248	(9)	153	(16)	27.6	0.4	dry
Kurunegala	574	(3)	402	(1)	253	(1)	2.9	4.5	dry
Puttalam	346	(13)	253	(8)	180	(8)	23.9	1.9	dry
Chilaw	444	(5)	281	(4)	201	(4)	6.5	3.2	dry
Anuradhapura	402	(8)	325	(2)	232	(2)	8.1	1.4	dry
Badulla	430	(6)	255	(6)	189	(5)	4.4	35.7	dry
Ratnapura	316	(14)	197	(15)	154	(15)	3.5	23.9	wet
Kegalle	498	(4)	229	(12)	169	(11)	1.3	21.0	wet
Sri Lanka	286		216		166		12.9	12.6	

^a‘Expected’ means average for same month or period during 1915–1917. Deaths in the period October 1918 to March 1919 are compared with one-third of the total deaths in first and fourth quarters during 1915–1917. Deaths in the period October 1918 to December 1919 are compared with one-third of the total deaths during 1915–1917 plus one-third of the total deaths in fourth quarters during 1915–1917.

^bRank indicated in brackets. District showing greatest rise = 1, district showing smallest rise = 21.

^c‘Urban’ includes the three municipalities of Colombo, Galle and Kandy, 21 local board areas and the towns proclaimed under the Births and Deaths Registration Ordinance.

^dThis matter is by no means completely straightforward and some classification schemes categorize some districts as ‘intermediate’ between wet and dry. This classification has been taken from United Nations (1976:35).

Ethnic groups

According to the 1911 census of Sri Lanka 66 per cent of the population were Sinhalese, 13 per cent were Ceylon Tamils, 13 per cent were Indian Tamils and six per cent were Ceylon Moors. The Sinhalese, Ceylon Tamils and Ceylon Moors were all long-established populations, whereas the Indian Tamils were immigrants from south India mainly working on tea estates. The Sinhalese are overwhelmingly Buddhist and the Tamils Hindu, though there are Christian minorities in both cases; Moors are Muslims.

On a ‘worst month’ view of the data there seemed to be very substantial differences between ethnic groups so far as the impact of influenza was concerned (see Table A1 in the appendix). In the worst month for Indian Tamils there were more than six times (611%) the average number of deaths in the same month for this group during 1915–1917; the corresponding ratio for the Sinhalese in the worst month was less than two-and-a-half (238%). The figures for Moors and Ceylon Tamils were 336 per cent and 281 per cent, respectively. However, viewed over a longer period of time the differences between ethnic groups were very much less and the positions of groups relative to each other changed somewhat. For the period comprising the last quarter of 1918 and the first quarter of 1919 the ratios of observed to average 1915–1917 deaths (percentages) were: for the Sinhalese 213; for Ceylon Tamils

188; for Indian Tamils 261; and for Moors 212. For the period comprising the last quarter of 1918 and all of 1919 the figures were: Sinhalese 170; Ceylon Tamils 143; Indian Tamils 184; and Moors 155.

It seems, then, that ethnic groups differed very considerably in terms of whether they suffered extremely heavy mortality from influenza in a relatively short period of time or somewhat lower mortality over a longer period, but that the longer-run differences between them, that is over six or 15 months, were very much smaller. Indian Tamils apparently suffered particularly badly in a relatively short period of time. This may well have been due to the fact that they were heavily concentrated in a residential sense in that they lived in small areas on estates and also fairly concentrated in a geographical sense in that estates were themselves concentrated in a certain part of the country (see Table 6 and Figure 1). The Sinhalese, by contrast, experienced much lower peak mortality but many more months not far off that peak, so that on balance over 15 months they were not much better off than the Indian Tamils. This may well have reflected the fact that the Sinhalese were much more scattered both in being more likely to live in small rural communities, through which influenza may have moved more slowly, and in being distributed across many more districts in the island, where influenza arrived at different times. In the case of the Sinhalese, though, malaria would have been a factor as well as influenza; the overwhelming majority of the population in the North-Western Province, and in Anuradhapura district, were Sinhalese. It should not be overlooked, however, that even on a 15-month view of the data Indian Tamils apparently suffered more than other groups in the population.

Why should Indian Tamils have experienced higher mortality from influenza than other sections of the population? Congested living conditions and low temperatures due to altitude may have played a part. However, the most important factor may well have been simply the poor state of health of Indian Tamils even in ordinary times. Indian Tamils were drawn from the lowest echelons of south Indian society; and their conditions of life and work in Sri Lanka were arduous. Their mortality tended to be considerably higher than that of other groups in the Sri Lankan population. The mean expectation of life at birth for Indian Tamils computed for the combined period made up by 1900–1902, 1910–1912 and 1920–1922 was 24.0 years; the corresponding figure for the Sinhalese was 33.6 years; the figures for Moors and Ceylon Tamils were 30.3 and 31.5 years respectively. (The figures cited are averages of the life expectancies of males and females; figures for Ceylon and Indian Tamils were derived as in Table A1 in the appendix.) Thus it may be that Indian Tamils suffered more than others in 1918–1919 mainly because they were already more debilitated and so at greater risk of death from influenza in the epidemic. However, it is also conceivable that the impact of influenza on this group may then have been further exacerbated by the serious food shortage that affected Sri Lanka in late 1918 and during 1919 (discussed below) since this may well have had more effect on those, like estate workers, who depended on wages and were remote from subsistence agriculture, than on others in the population.

Age groups

Table 7 shows the impact of influenza during 1918–1919 on different age groups in the Sri Lankan population. It may be seen that, as has been noted for other countries, young adults suffered more, in one sense at least, than any other age group. The greatest proportional increases in mortality rates occurred among men aged between 20 and 35 and among women aged between 15 and 35. By contrast, infant mortality and mortality among those aged 45 and over were apparently not very much affected by the epidemic. But a somewhat different picture emerges when absolute rather than relative changes in mortality rates are considered. For both males and females it was those in the first year of life who suffered the greatest absolute increase in the mortality rate with the 1–4 age group experiencing the second greatest rise. The adult age groups which suffered the greatest proportional increases in mortality rates experienced lower absolute increases than these two groups. Thus, despite the fact that the infant mortality rate apparently only rose nine per cent during 1918–1919 the implied number of

additional deaths in infancy was still more than 25 per cent of the implied additional number in the entire 15–34 age group. Moreover, the implied number of additional deaths in the 1–4 age group was more than half the implied additional number in the 15–34 age group.²

² In general the estimated age-specific death rates for Sri Lanka in 1918–1919 shown in Table 7 may be biased upwards slightly because of the way the denominators were estimated. This was done by interpolating 75 per cent of the way between the 1911 and 1921 censuses. No account was taken of the fact that the 1921 population was itself somewhat diminished precisely because of the influenza epidemic; moreover arguably a figure of a little higher than 75 per cent might have been used. However, in the case of the 1–4 age group the upward bias may well be greater than in other age groups because the population aged 1–4 in 1921 would have been affected also by the fall in the number of births in the aftermath of the epidemic. On the basis of the figures presented in Table 7 the implied number of additional deaths in the 1–4 age group because of the epidemic was almost two-thirds the implied additional number in the entire 15–34 age group. The extremely cautious expression ‘more than half’ used in the text easily allows for the maximum conceivable upward bias in the 1–4 rate.

Table 7

Estimated age-specific mortality rates by sex for Sri Lanka in 1918–1919, ratios (%) of estimated to 'normal' rates and differences between estimated and 'normal' rates

Age group	Males			Females		
	Estimated rate ^a (per 1000)	Estimated/ normal rate ^b (%)	Estimated - normal rate ^c (per 1000)	Estimated rate ^a (per 1000)	Estimated/ normal rate ^b (%)	Estimated - normal rate ^c (per 1000)
0	211	109	17	198	108	16
1–4	53	131	12	62	127	13
5–9	15	113	2	18	114	2
10–14	9	102	1	11	116	1
15–19	13	122	3	16	166	6
20–24	17	146	6	23	153	8
25–34	20	143	6	30	150	10
35–44	26	126	5	29	131	7
45–54	33	107	2	29	111	3
55+	77	103	2	91	99	-1

^aNumbers by age group were estimated by linear interpolation (75% of the way) between the censuses of 1911 (taken on March 10) and 1921 (taken on March 18). Age-specific death rates were computed by dividing half the registered deaths in the age group in question during 1918 and 1919 by the estimated number in the age group. For the first year of life an infant mortality rate rather than an age-specific death rate was used. This was calculated by dividing the infant deaths in 1918 and 1919 by the live births in this period.

^bThe rates taken as 'normal' were for the combined period made up by 1900–1902, 1910–1912 and 1920–1922. Age-specific death rates were computed by dividing one-third of the total number of deaths in the age group in question during these nine years by the sum of the 1901, 1911 and 1921 census counts of the age group. Infant mortality rates were calculated by dividing the infant deaths during these nine years by the live births in the same period.

^cThese figures may contain rounding errors since they were obtained by subtracting rates expressed to the nearest whole number per 1000.

Women

It is clear from the data presented in Table 8 that women suffered heavier mortality than men during the 1918–1919 influenza outbreak in Sri Lanka. The ratio of male to female deaths fell markedly in the last quarter of 1918 (to 0.93) and the first quarter of 1919 (0.95). Only in five previous quarters (since 1900) had the ratio of male to female deaths dropped below one and only in two quarters had this ratio dropped below 0.99 (in one instance to 0.93 and in the other to 0.96), in both cases when there was a serious outbreak of malaria. It may be seen from Table 7 that female mortality apparently rose rather more than male mortality in all age groups between 15 and 45, and possibly between ten and 55.

One possible factor in this heavier mortality among females is the risk associated with pregnancy and childbirth. According to Beveridge (1977:15),

In most pandemics up to and including that of 1918–19, there were reports of abortions and stillbirths due to influenza. In 1918–19, one series of 1350 pregnant women who had influenza were observed: abortion, stillbirth or premature labour occurred in 26% of those without pneumonia and 52% of those with pneumonia. The prognosis was said to be serious for the women who aborted or went into labour.

Table 8

Live births, maternal deaths, ratios of male to female deaths, stillbirths^a and stillbirth^a rates, by quarter, for Sri Lanka during 1918, 1919 and 1920

Year and quarter		Live births	(Relative to average number 1915–17 = 100)	Maternal deaths	Ratio male/female deaths	Stillbirths ^a	Stillbirths ^a per 1000 live births
1918	1st	52,623	(114)	988	1.01	266	56
	2nd	42,731	(103)	706	1.03	254	63
	3rd	42,869	(109)	775	1.03	263	64
	4th	45,164	(97)	1,541	0.93	303	71
1919	1st	50,448	(110)	1,332	0.95	296	61
	2nd	41,077	(99)	988	1.00	186	45
	3rd	31,165	(79)	649	1.02	210	65
	4th	38,716	(83)	695	1.07	264	62
1920	1st	45,465	(99)	810	1.05	258	57
	2nd	38,494	(93)	714	1.05	265	59
	3rd	38,054	(97)	638	1.05	285	62
	4th	41,707	(89)	731	1.07	376	73

^aStillbirths and ratios of stillbirths to live births are for the 33 towns in which stillbirths were registered at this time.

It may be seen from Table 8 that there was a marked upsurge in maternal mortality (mortality associated with pregnancy or childbirth) in the last quarter of 1918 and the first quarter of 1919 in Sri Lanka, indicating that influenza did indeed lead to complications in pregnancy as Beveridge (1977) suggested. However, far from there being an increase in the stillbirth rate at this time, there seems to have been a slight drop in the rate, at least during 1919. In the years 1912–1921, the first ten years stillbirths were registered in Sri Lanka, in certain urban areas, the ratio of stillbirths to live births in these areas varied between 58 per 1,000 in 1919 (and in 1915) and 71 per 1,000 in 1917, with no particular trend over time. It may be seen from Table 8 that this ratio was particularly low—at 45 per 1000—in the second quarter of 1919; this was noticeably lower than for any other quarter during the period 1912–1921, the next lowest figure being 53 per 1,000. It is possible that the stillbirth rate fell at this time because influenza, in tending to cause miscarriage, and even the death of the mother, had the effect of bringing forward potential difficulty in some cases to an earlier stage of pregnancy.

However, the heavier mortality experienced by females during the 1918–1919 epidemic could not have been entirely explained by pregnancy-related deaths. What other reasons might there have been for the higher mortality among females? The explanation may be that in general at that time in Sri Lanka mortality tended to be higher for females than males. The mean expectation of life at birth early in the twentieth century, based on the ‘normal’ rates referred to in Table 7, was 32.7 years for males and 30.2 years for females: mortality rates were higher for females than males at all ages except during infancy and in the 45–54 age group; and early in the twentieth century also apparently in the 15–19 age group, though there are reasons for believing this may not have been genuine. It may have been then, much as has been suggested in relation to Indian Tamils, that women tended to suffer heavier mortality than men in the 1918–1919 epidemic partly because they were already in somewhat worse health than men.

This of course begs the question of why this should have been so in the first place. It has been suggested that the phenomenon of excess female mortality in the Indian sub-continent (for it is by no means confined to Sri Lanka) essentially reflects discrimination against women in such areas as nutrition, medical attention and general care. Whether this was so in Sri Lanka or whether other factors such as differences between the sexes in environment or lifestyle, or differences in their susceptibility to particular diseases, were responsible, is an interesting and debatable matter, but one which is unfortunately beyond the scope of this paper.

It may be seen from Table 8 that fertility clearly fell in response to the influenza outbreak: there was a marked drop in the second half of 1919 and it looks as though fertility was somewhat depressed throughout 1920 (though it recovered in 1921). This probably occurred for a number of reasons including the death of prospective mothers or their partners, an increase in foetal loss, a reduction in coital frequency because of illness or social disruption and the postponement of marriages because of death in the family.

The effect of the epidemic on agricultural production, and the possible significance of food availability for mortality in the epidemic

There is little doubt that the influenza epidemic disrupted agricultural production in Sri Lanka. According to the Director of Agriculture in his report for the year 1918 (Ceylon 1919:Part IV, C1) influenza 'seriously handicapped agricultural operations.' Tea exports were 12.5 million pounds, in weight, lower in 1918 than in 1917, down from 193 million pounds.

This ... was brought about to some extent by unfavourable weather conditions ... but was mainly the result of the influenza epidemic, which made it impossible to gather the whole of the north-east monsoon crop (Ceylon 1919:Part IV, C1).

'Paddy crops during the year were, on the whole, satisfactory' (p. C4) However,

In the North-Central Province and in parts of the Central Province sowing of the maha [major harvest] crop could not be carried out owing to the severe influenza epidemic, and shortage of paddy crops is to be expected in these parts. In some areas of the Southern Province cultivation was also prevented by the influenza epidemic (Ceylon 1919:Part IV, C4).

Local official reports from various parts of the island also mention difficulties of this kind towards the end of 1918 and early in 1919. The Assistant Government Agent for the Northern Province in his report for the district of Mannar in 1918 (written in March 1919) reported that 'The prevalence of influenza during the latter part of the year has very seriously interfered with the sowing of the paddy fields for the 1918–19 crop' and that the area cultivated 'is, it is estimated, probably not much more than half that cultivated for 1917-18' (Ceylon 1919:Part I, D14). The Government Agent for the Province of Uva, reporting in April 1919, wrote of Badulla district:

The evil effects of the epidemic have, I fear, by no means passed. The health of many villagers has been impaired, and their ability to work decreased. Cultivation of chenas [shifting plots of temporarily-cleared jungle] was in many cases interfered with, and in one case, where a whole village was struck down at once, the village chena was not watched, and has been totally destroyed by elephants, and the village has been left entirely destitute of food (Ceylon 1919:Part I, H1-H2).

It seems extremely likely, then, that there would have been some shortfall in the availability of locally-grown food in Sri Lanka during 1919 as a result of the influenza outbreak. However, the food shortage that developed in Sri Lanka during 1919, and indeed during the last few months of 1918, had relatively little to do with the situation in Sri Lanka and a great deal to do with the situation in India at the time. According to the Principal Collector of Customs of Ceylon in his report for the year 1919:

During the first four years of the war this Island, which imports over three-quarters of the foodstuffs required for its population, had been in the happy position of ability to obtain practically all its requirements. In 1918, however, the failure of the monsoon and the depletion of the normal reserves in India, coupled with the difficulty in obtaining freight, led to considerable anxiety. By the end of that year the export of all rice from South India had been prohibited, and Ceylon was thrown entirely on Rangoon and Calcutta for its supplies. At the beginning of 1919 the position was therefore serious, and the crisis rapidly came to a head in April, when the Indian Government informed Ceylon that the exports from India would be curtailed to 140,000 tons for the first nine months of the year, as against 30,000 tons a month, the normal requirements of this Island. On urgent representations a further supply of 50,000 tons was allowed to be ear-marked for the use of estate labourers and other Indians resident in Ceylon. As the stock of rice at any one time in the Island never exceeds two months' supply, a system of food control and rationing of supplies had to be inaugurated at once (Ceylon 1920:Part II, A3).

It is clear that the shortage of imported food was already serious during the last few months of 1918. The Assistant Government Agent for the Western Province in his report for the district of Kalutara for 1918 referred to a 'rice crisis' in the district from September onwards (Ceylon 1919:Part I, A16-A17). In some villages in Puttalam district there were food riots (Ceylon 1919:Part I, F10). There were many references to food shortages and rising food prices in local official reports.

At the time of the 1918–1919 influenza outbreak in Sri Lanka, then, there was quite a serious food shortage in the country. This may have been exacerbated, at least during 1919³, by the epidemic itself through its debilitating effects on the labour force, but in the main it was the result of the situation in India. Did this food shortage itself contribute to the mortality in the influenza epidemic? Certainly some observers at the time believed it had made things worse, particularly when coupled with the poor state of the economy at the time. According to the Assistant Government Agent for the Western Province in his report on Kalutara district for the year 1918,

The year under review ... has been a period of continued decline, staple products being low in price or almost unsaleable, while the cost of the necessities of life steadily rose. This produced a considerable amount of unemployment and hardship for the poorer classes, especially in the towns. To crown these misfortunes came the influenza epidemic, producing in places positive distress. ... Between September and the end of the year matters were made worse by a serious shortage of imported rice. Prices rose to about double the normal rates, making rice almost prohibitive in cost for the poorer classes. Steps were taken to guard against looting or any disturbances ... Coming on the top of this

³ The influenza epidemic may well have caused some drop in food production in Sri Lanka in the early part of 1919. It is rather unlikely, however, that output over the whole year was adversely affected. There was a great deal of scope in Sri Lanka for late planting and planting of alternative crops. Moreover, the high prices which obtained during 1919 because of the shortage of imported food would have provided an incentive for increased output; and there are some indications that this was indeed what happened.

condition of high prices and scarcity of employment, the influenza epidemic hit the poorer people hard and caused temporary acute distress in some places. The universality of the catastrophe appeared at first to paralyze effort. Whole families and hamlets were struck down together, and there was often none to procure food or attend on the sufferers (Ceylon 1919:Part I, A16-A17).

It is obviously possible that in Sri Lanka in 1918–1919 the population was indeed somewhat weakened because of the food shortage and the poor state of the economy and that this led to a higher death toll in the influenza epidemic than otherwise would have been the case. It might even be thought rather unlikely that there would have been no effect at all of this kind; though it should perhaps be emphasized that the measures taken for the procurement and distribution of food worked quite well and that by no stretch of the imagination was there famine in Sri Lanka at this time. Those dependent on wages from their labour and remote from subsistence agriculture might well have been particularly vulnerable; this would include labourers in urban areas and on tea estates. It must be emphasized, though, that it is simply not possible to gauge the likely scale of any effect that shortage of food may have had on influenza mortality at this time.

Conclusion

The outbreak of influenza in Sri Lanka in 1918–1919 had all of the features for which the 1918–1919 influenza pandemic is famous. It took a terrible toll in mortality, at least as bad as the worst malaria epidemic in Sri Lanka up to that time (since 1900), an important reason being that sufferers tended to develop fatal pneumonic complications. It progressed through a number of waves, the first innocuous, the second extremely severe and the third a damped version of the second. It showed a preference, in a certain sense, for victims who were in the prime of life rather than very young or old.

Influenza was first reported in Sri Lanka, in its mild form, in June 1918. The severe second wave began in September and reached a peak in November 1918. The evidence is consistent with the disease having entered the country through Colombo, the capital city and main port, and possibly also through Talaimannar, the railway terminus and port on the rail-sea route to India, and having spread out from there. Although the disease was sometimes referred to in Sri Lanka at the time as ‘the Bombay fever’ or ‘Bombay influenza’ there is no particular reason to think that influenza did indeed arrive from there. In fact, so far as one can tell, the timing of the epidemics in Bombay and Colombo was the same, with influenza first appearing in June and the severe outbreak beginning in September and reaching a peak in October; for details of the epidemic in India see Mills (1989).

In most parts of the island the peak of the epidemic, so far as mortality was concerned, came towards the end of 1918. However, in a few districts the peak was in early 1919. It is possible that influenza simply moved more slowly through some districts than others. However, it is also possible that in some districts influenza mortality was exacerbated and prolonged by an outbreak of malaria at the same time.

The impact of influenza in Sri Lanka, so far as mortality was concerned, varied between areas and between ethnic groups. At the time it was widely remarked that the Indian Tamil labour force on estates suffered particularly heavy mortality in the epidemic. This was probably true, possibly reflecting particular circumstances on estates, possibly the generally poor health of Indian Tamils, and possibly in some degree the food shortage of 1918–1919. Yet people’s perception was almost certainly entirely fashioned by the especially short-and-sharp nature of the outbreak in this group, which was in some degree misleading and might have been entirely so. Clearly, it is necessary to be cautious about accepting too readily at face value the perception of people at the time of such matters.

A similar point might be made in relation to the impact of influenza on different age groups in the population. A few observers reported at the time in Sri Lanka that the epidemic seemed to strike those in the prime of life especially hard. They might have been led to this conclusion by experiencing the fact that mortality rates rose proportionally more among young adults than for any other age group. However they were probably also influenced by a sense that young adults were the group who were normally least affected in epidemics, that for their mortality to rise was unusual. In any case, as has been noted, their perception of the matter was in danger of overlooking the fact that the greatest absolute rises in mortality rates occurred among those under age five so that there was a very substantial number of deaths among children during the influenza epidemic. It might be interesting to re-examine the data for developed societies from this point of view.

Women suffered heavier mortality than men during the influenza outbreak in Sri Lanka and maternal mortality rose and fertility fell. Contrary to expectation, however, the stillbirth rate did not rise but fell slightly at this time; though it is perfectly possible that the miscarriage rate may have increased during the epidemic. The heavier mortality among females was no doubt due in part to the particularly damaging impact influenza had on pregnant women; in addition the fact that female mortality was usually higher than male mortality in Sri Lanka at that time, and female health therefore presumably less good, may also have played a part.

The influenza outbreak undoubtedly caused a great deal of disruption in Sri Lanka. Morbidity and mortality were substantial; both plantation agriculture and subsistence agriculture were adversely affected. There were many local reports of difficulty or distress because of the temporary breakdown of family and village support systems. But many communities seem to have responded to the situation fairly quickly and there was a proliferation of 'relief committees' organizing support for victims of the outbreak. Such activity may have been quite important in ameliorating the impact of the epidemic.

During the last few months of 1918 and during 1919 there was quite a serious food shortage in Sri Lanka. This arose because India restricted rice exports, on which Sri Lanka was heavily dependent, and had little or nothing to do with any drop in agricultural production there might have been in Sri Lanka itself as a result of the influenza epidemic. It is possible that this food shortage worsened the mortality from influenza during the 1918–1919 outbreak. However, it is not possible to gauge the likely scale of any effect of this kind.

How many deaths were there in Sri Lanka during 1918–1919 as a result of the influenza outbreak? The number of additional deaths during these two years implied by the differences between the estimated age-specific death rates for 1918–1919 and the 'normal' rates shown in Table 7 is about 51,000.

This is probably a reasonable estimate, though taking account of the fact that the age-specific rates for 1918–1919 may be slightly too high (see footnote 2) might reduce this figure by two or three thousand. However, it does depend on the arbitrary assumption that the excess mortality during 1918–1919 which was due to malaria rather than influenza is exactly offset by the excess mortality due to malaria included in the so-called 'normal' rates as a result of the fact that two of the nine years on which they are based, 1911 and 1912, were bad years for malaria.

In round numbers, then, some 50,000 people probably died in Sri Lanka during 1918–1919 as a result of the influenza outbreak, about 1.1 per cent of the population. Expectation of life at birth fell from the usual level of 31.5 years to 26.7 years: these values are based upon the 'normal' and the estimated 1918–1919 mortality rates of Table 7. This was heavier mortality than was experienced in

England and Wales, where possibly one in 200 (0.5%) of the population died in the epidemic.⁴ However, it was very much lighter than the mortality in India where according to Mills (1989:256) almost 5.5 per cent of the population died.

⁴ The figure of 0.5 per cent in fact relates to the female population of England and Wales. There are horrendous problems in attempting to produce such an estimate for the male population, not least the difficulty of estimating the population exposed to risk during a period including the closing months of the First World War and the first few months of peace. According to the Registrar-General for England and Wales there were 100,000 female deaths during 1918–1919 in England and Wales which were attributable to the influenza epidemic (this included deaths from more causes than just influenza); the estimated mid-1918 population of females was 19,697,600. See *The Eighty-First Annual Report of the Registrar-General of Births, Deaths, and Marriages in England and Wales*, HMSO, London, 1920; and the supplement to that report entitled *Report on the Mortality from Influenza in England and Wales during the Epidemic of 1918–19*, HMSO, London, 1920.

Appendix

Table A1

Ratio of deaths each month during 1918 and 1919 to the average number of deaths during the same month 1915–1917, for the whole of Sri Lanka, for the different ethnic groups, and for administrative districts

1918	Jan	Feb	Mar	Apr	May	Jun	Jul	Month Aug	Sep	Oct	Nov	Dec
All Sri Lanka	99	100	92	90	84	85	90	94	109	258	286	195
Sinhalese	104	108	97	95	83	82	94	96	112	224	220	180
Ceylon Tamils	82	77	80	73	78	88	85	97	92	281	253	184
Indian Tamils	95	93	91	94	92	93	84	98	98	322	611	301
Moors	97	96	81	80	77	86	89	76	120	336	238	172
Colombo	108	107	96	91	85	84	96	89	132	262	158	117
Negombo	131	121	99	123	88	74	102	92	98	178	143	127
Kalutara	101	99	95	96	104	100	111	108	120	310	211	126
Kandy	99	109	88	92	86	87	87	93	119	248	364	248
Matale	102	104	89	91	82	71	87	86	100	324	427	378
Nuwara Eliya	87	103	76	100	82	101	86	111	68	138	828	139
Galle	116	111	95	105	90	84	91	96	105	213	171	122
Matara	120	113	99	115	94	84	92	100	98	180	190	140
Hambantota	93	107	110	107	87	73	80	107	110	196	370	249
Jaffna	76	70	77	74	83	97	89	94	90	306	243	168
Mannar	85	94	82	92	82	82	86	89	137	623	397	172
Mullaittivu	94	91	99	47	58	127	76	95	127	388	389	296
Batticaloa	100	92	84	67	67	72	71	91	82	115	205	226
Trincomalee	88	99	99	69	81	68	119	94	87	391	264	232
Kurunegala	98	100	96	83	68	71	78	80	99	198	291	288
Puttalam	92	97	67	69	78	66	120	104	149	243	187	199
Chilaw	113	104	88	82	66	76	75	89	106	222	169	153
Anuradhapura	70	71	64	63	79	87	89	110	102	244	319	267
Badulla	77	86	105	98	96	89	99	87	94	269	430	303
Ratnapura	96	93	78	76	78	84	95	97	141	316	261	158
Kegalle	129	131	127	94	81	94	82	93	135	498	293	203
1919												
All Sri Lanka	187	201	179	142	137	155	145	126	133	118	111	116
Sinhalese	203	238	216	160	149	166	153	142	124	117	115	120
Ceylon Tamils	163	138	137	122	128	148	130	100	86	81	82	103
Indian Tamils	139	137	88	99	109	140	125	101	216	165	122	113
Moors	191	186	173	125	115	121	158	106	98	98	102	110
Colombo	131	136	127	117	145	163	127	114	116	105	129	129
Negombo	179	194	186	156	156	196	172	133	114	110	103	107
Kalutara	107	96	102	128	169	174	155	142	109	118	107	104
Kandy	125	161	100	97	126	158	120	109	177	149	146	107
Matale	218	211	133	107	121	142	147	174	105	99	99	97
Nuwara Eliya	103	136	114	112	99	106	93	122	371	157	93	141
Galle	111	124	102	110	126	157	148	133	120	128	108	115
Matara	164	149	134	124	142	169	188	180	147	142	126	154
Hambantota	233	214	169	159	151	135	163	142	100	118	114	110

Continued over

Continued

Jaffna	142	114	119	110	94	94	85	83	79	80	78	103
Mannar	162	196	159	119	106	98	133	75	81	80	79	90
Mullaittivu	239	171	241	163	180	198	128	123	145	112	112	134
Batticaloa	256	256	225	163	182	241	237	135	93	95	100	108
Trincomalee	237	219	163	116	109	56	86	72	74	85	62	104
Kurunegala	472	574	518	315	160	171	161	117	100	97	95	102
Puttalam	243	346	260	175	139	102	153	77	93	89	113	117
Chilaw	251	403	444	239	148	170	130	98	120	99	131	140
Anuradhapura	308	402	364	243	226	147	154	131	123	106	94	107
Badulla	182	207	161	140	128	133	184	151	130	144	117	154
Ratnapura	204	133	118	109	117	167	130	136	132	115	128	90
Kegalle	133	120	137	93	103	167	165	138	134	139	92	119

Notes: The figures shown for Ceylon Tamils are in fact for all Tamils in Ceylon Tamil districts, that is, Batticaloa, Jaffna, Mannar, Mullaittivu and Trincomalee; the figures for Indian Tamils are for all Tamils in Indian Tamil districts, that is, Badulla, Kalutara, Kandy, Kegalle, Matale, Nuwara Eliya and Ratnapura. It was necessary to proceed in this fashion because no distinction was made between Ceylon and Indian Tamils in the published statistics from death registration. The distribution of Tamils by district in Sri Lanka was such, however, that this approach must give reasonable results.

The overwhelming majority of Moors (88% in 1911) were Ceylon Moors.

The district referred to as Mullaittivu district at this time later became Vavuniya district and has now been split into Mullaittivu and Vavuniya districts.

Table A2

Deaths from influenza or pneumonia in the last two quarters of 1918 and all quarters of 1919 for the whole of Sri Lanka and for the different administrative districts, together with indices showing the mortality of each quarter from these relative to the worst quarter 1900–1921

	1918		1919			
	3 ^a	4	1	2	3	4
	Deaths from influenza or pneumonia					
Sri Lanka	2,049	31,443	15,158	8,200	7,316	2,892
Colombo	509	2,810	1,124	1,426	654	553
Negombo	30	353	328	193	117	95
Kalutara	88	1,637	111	592	271	63
Kandy	404	4,228	897	730	916	461
Matale	60	2,080	216	131	276	74
Nuwara Eliya	131	2,678	436	362	1,246	196
Galle	45	1,448	159	455	428	128
Matara	7	734	166	138	247	42
Hambantota	12	1,103	444	153	83	33
Jaffna	218	2,558	653	289	265	246
Mannar	30	468	69	28	21	22
Mullaittivu	22	342	101	60	50	36
Batticaloa	38	338	1,306	1,115	787	178
Trincomalee	23	323	165	27	36	45
Kurunegala	70	2178	4,874	934	181	95
Puttalam	34	92	311	162	24	46
Chilaw	12	95	291	124	24	30
Anuradhapura	15	934	1,478	369	65	30
Badulla	124	3,762	1,465	490	1,171	358

Continued over**Continued**

Ratnapura	138	1,338	382	320	257	56
Kegalle	59	1,944	182	102	197	105

Relative to worst quarter 1900–1921 = 100

Sri Lanka	7	100	48	26	23	9
Colombo	18	100	40	51	23	20
Negombo	8	100	93	55	33	27
Kalutara	5	100	7	36	17	4
Kandy	10	100	21	17	22	11
Matale	3	100	10	6	13	4
Nuwara Eliya	5	100	16	14	47	7
Galle	3	100	11	31	30	9
Matara	1	100	23	19	34	6
Hambantota	1	100	40	14	8	3
Jaffna	9	100	26	11	10	10
Mannar	6	100	15	6	4	5
Mullaittivu	6	100	30	18	15	11
Batticaloa	3	26	100	85	60	14
Trincomalee	7	100	51	8	11	14
Kurunegala	1	45	100	19	4	2
Puttalam	11	30	100	52	8	15
Chilaw	4	33	100	43	8	10
Anuradhapura	1	63	100	25	4	2
Badulla	3	100	39	13	31	10
Ratnapura	10	100	29	24	19	4
Kegalle	3	100	9	5	10	5

^aThe district figures shown for this quarter in fact include the 20 influenza deaths which occurred in the previous quarter (with the effect that the district totals sum to 20 more than the Sri Lanka total).

References

- Beveridge, W.I.B. 1977. *Influenza: The Last Great Plague*. London: Heinemann.
- Burnet, F.M. 1979. Portraits of viruses: influenza virus A. *Intervirology* 11:201-214.
- Ceylon. 1919. *Administration Reports, 1918*. Colombo: Acting Government Printer.
- Ceylon. 1920. *Administration Reports, 1919*. Colombo: Government Printer.
- Cliff, A.D., P. Haggett and J.K. Ord. 1986. *Spatial Aspects of Influenza Epidemics*. London: Pion Ltd.
- Crosby, A.W. 1976. *Epidemic and Peace, 1918*. Westport, Connecticut: Greenwood Press.
- Davis, K. 1951. *The Population of India and Pakistan*. Princeton: Princeton University Press.
- French, H. 1920. The clinical features of the influenza epidemic of 1918-19. Pp. 66-109 in Ministry of Health, *Report on the Pandemic of Influenza, 1918-19*. Reports on Public Health and Medical Subjects No. 4. London: HMSO.
- Hoyle, F. and C. Wickramasinghe. 1977. Does epidemic disease come from space? *New Scientist* 76: 402-404.
- Hoyle, F. and C. Wickramasinghe. 1978. Influenza from space? *New Scientist* 79: 946-948.
- Mills, I.D. 1989. Influenza in India during 1918-1919. Pp. 222-260 in *India's Historical Demography*, ed. T. Dyson. London: Curzon Press.
- Ministry of Health. 1920. *Report on the Pandemic of Influenza, 1918-19*. Reports on Public Health and Medical Subjects No. 4. London: HMSO.
- Patterson, K.D. 1986. *Pandemic Influenza 1700-1900*. Totowa, New Jersey: Rowman and Littlefield.
- Stuart-Harris, C.H., G.C. Schild and J.S. Oxford. 1985. *Influenza: The Viruses and the Disease*. London: Edward Arnold.
- United Nations. 1976. *Population of Sri Lanka*. Bangkok: Economic and Social Commission for Asia and the Pacific.

The mines of Southern and Central Africa: an ecological framework



Bruce Fetter

*Department of History, University of Wisconsin Milwaukee, PO Box 413,
Milwaukee, WI 53201, USA*

Abstract

The mines of Southern and Central Africa can be seen as a congeries of microenvironments whose suitability for human habitation has been transformed as the result of capital investment as constrained by government policies. Mine managers sought improved living conditions to enhance productivity but minority governments imposed strict controls on African migration, which allowed only a unisex labour force in the gold mines of South Africa whilst permitting family settlement in the copper mines of what is today Zambia and Za•re.

My research over the past 25 years has been concerned with identifying and explaining the differential effects of colonial rule on the peoples of Central Africa. My first monograph was a history of Lubumbashi, the largest city in the Za•rian copperbelt, which I found to have provided benefits to its residents unavailable in villages in surrounding territories (Fetter 1976). Using terminology then current among historians of urbanization, I concluded that the city was parasitical on its hinterland. My next study addressed the disparities between the distribution of colonial resources and the emplacement of the rural African populations (Fetter 1983). I concluded that colonial rule in what is now Za•re, Zambia, and Malawi had fostered a spatial system in which the principal government resources went to those who lived near colonially created cities, a condition which I called regional imbalance.

My approach deviated from that of the majority of scholars in the field, whose work falls under the general rubric of political economy (Perrings 1979; Parpart 1983; Higginson 1989). Their major concern has been to show how European capital investment has caused the immiseration of African workers through the expropriation of the fruits of their labour. While I would never deny that Africans often suffered because of the labour demands imposed on them in the course of colonial rule, I feel that obsession with the evils of capital has obscured other important elements which also determined their experience.

In recent years, in search of a more complete description and explanation of the colonial experience, I have turned to demography. Statistically, the inhabitants of Central African mining camps are perhaps the best documented people in the region (Fetter 1990).¹ A preliminary examination of the available evidence showed that although the mines were initially very dangerous to the health of African workers, conditions soon improved and within a maximum of 30 years after their opening, the mines were healthier than the villages from which mineworkers were recruited (Fetter 1986, 1987). See Table 1.

¹ This study was based on a 1986 conference held at the University of Wisconsin Milwaukee funded by N.S.F. Award Number SES-8520051.

Table 1
Crude death rates for African mines, 1903–55

Year	Katanga ^a	N. Rhodesia ^b	S. Rhodesia ^c	Witwatersrand ^d
1903				71 (81)
1904				43
1905				46
1906			76	35
1907			61	35
1908			50	34
1909			47	36
1910			49	33
1911			33	32
1912			36	(25)
1913			28	(18)
1914	65 (118)		28	(19)
1915	52 (54)		26	(15)
1916	92 (94)		27	(13)
1917	106 (106)		22	(20)
1918	202 (202)		113	(17)
1919	51 (51)		19	(18)
1920	33 (33)		18	(14)
1921	27		21	(13)
1922	31		21	(15)
1923	33		16	(14)
1924	30		18	(13)
1925	50	16	15	(12)
1926	53	13	16	11 (13)
1927	45	21	16	14 (16)
1928	33	20	19	14 (16)
1929	23	14	21	13 (13)
1930	18	25	17	13 (14)
1931	12	25	15	12 (13)
1932	8	12	12	10 (11)
1933	7	13	11	9 (11)
1934	6	18	12	11
1935	6	10	14	10
1936	8	7	12	15
1937	9	10	11	15
1938	8	9	12	
1939	5	6	9	
1940	5	7	8	(6)
1941	6	7	9	
1942	6		8	
1943	6	7	8	
1944	4		9	
1945	4	5		
1946	4			
1947	3	4		
1948	4			
1949	3	4		(5)
1950	4	4		(5)
1951	3	4		(5)
1952	2	3		(4)

Table 1
Crude death rates for African mines, 1903–55 (continued)

1953	3	3	(5)
1954	4	3	(5)
1955	3	3	(5)

Source:

^aMouchet 1943; Union Mine 1955, 1957; Fetter 1976; Perrings 1979. Elisabethville only in parentheses.

^bKuczynski 1949; Northern Rhodesian Chamber 1957; Perrings 1979.

^cKuczynski 1949; van Onselen 1976.

^dAbrahams 1937; Rand Mines 1964; Perrings 1979; Richardson 1982; Turrell 1987. Materials relating to Chinese labour in 1903 and those relating to the Rand Mines group rather than the entire industry are in parentheses.

Whatever the disparity between the mines and their hinterlands, however, I concluded that national-level generalizations about the demography of Africans living under colonial rule were not possible in the current state of our knowledge. Instead, I turned to the study of limited microenvironments which had been created as the result of colonial activity. My first local study compared mortality patterns among schoolboys at an elite school in what is now Malawi in the early years of the twentieth century (Fetter 1989). The statistics derived from a missionary rollbook suggested that conditions for schoolboys improved as the result of a new water supply installed by the missionaries, while those for school leavers deteriorated as a result of colonial policies which forced them to leave their villages in search of salaried employment.

Central Africa is not, of course, the only region of the world where the demography of cities has differed from that of the surrounding countryside. Most industrialized countries of the world have passed from a time, often just before and during the industrial revolution, when cities had higher mortality rates than the surrounding countryside, to one in which the cities had lower rates (Preston and Van de Walle 1978; Szreter 1988). The explanation for that crossover, however, requires detailed knowledge of conditions both after the improvement of city health and before, when times were at their worst. The analysis of these earlier times has been extremely difficult, but in recent years, Landers (1990) has demonstrated that for London, at least, the work can be done.

Landers's work breaks new ground in that it presents eighteenth-century London not just as a place where humans died of many diseases but as a loosely bounded epidemiological microenvironment, where animals ranging from large mammals to single celled parasites interacted. Indeed, his analysis also extends to plant life, bacteria, and viruses. In short, his work demonstrates the feasibility of a broad approach to industrializing urban areas based on the entire spectrum of biological factors which affect human health.

Considering human interaction with the environment is by no means unique to Landers, but unlike the vast majority of his fellow anthropologists, he has chosen to study an industrial rather than a preindustrial or prehistoric society (e.g. Hardesty 1977). The approach of ecological anthropology has been a slippery one in that scholars have defined the field in a number of different ways. Most have considered it a branch of what used to be called man-environment studies, investigating the way in which individual groups have taken advantage of local resources and ways in which they have altered that environment. All too rarely, however, do anthropologists ask how humans fit into the total scheme of things. Two notable exceptions have been Ingold (1986) and Vansina (1990).

Ingold approaches the philosophical problems of differentiating human adaptive behaviour from that of other animals. In *The Appropriation of Nature* (1986), he establishes a series of polarities to distinguish human actions based on self-consciousness and creativity from seemingly similar behaviour in animals: hunting-collection versus foraging predation; sharing versus simple co-operation; land tenure versus territoriality; storage versus accumulation; and purposive hierarchy versus individualist inequality. His explanation of how hunters and collectors differ from other predators and social insects

demonstrates that an analyst with a good appreciation for non-human ecology can gain insights into the human condition.

In a more specifically African context, Vansina, in *Paths in the Rainforest* (1990), explores the relations between Western Bantu-speakers and their equatorial environments. Vansina adopts the term, 'biotope', which he defines as a 'host of habitats varying by soil, placed on a catena, and striving toward different biological climaxes.' He then analyses the adaptation of historical African populations to a broad variety of biotopes and to the ecotones, which lie on the frontiers between them. Within this framework, he reconstructs the history of peoples who lived for 5000 years in a region covering two million square kilometres of Equatorial Africa.

With the work of Landers, Ingold, and Vansina in mind I have undertaken the construction of an ecological framework for the recent history of the mines of Central and, indeed, southern Africa, which provided the technology on which the tropical mines depended. The framework is distinctive from those mentioned above in that it explains the effects of industrial technology in a specifically colonial setting. At this point it is not possible to present a full analysis, but in sketching an outline, I would like to propose a framework in which other scholars and I can work and, in the process, challenge each other.

The mines of industrial Africa constitute a set of discontinuous microenvironments defined by the presence of minerals which could be profitably exploited through Western technology of the late nineteenth and early twentieth centuries. These minerals include diamonds in South Africa, gold in South Africa, Zimbabwe, and Za•re, lead and zinc in Zambia, and copper in Za•re and Zambia. These deposits lie in a broad variety of ecotopes ranging from the Kalahari edge in the northwest Cape to the Highveld in the Transvaal and Zimbabwe and to the savannahs of north-central Zambia and southeastern and northeastern Za•re.

Just before the mine development, the human occupants of these territories consisted of rainfall and alluvial agriculturists who domesticated animals (cattle at higher elevations, sheep, goats, and poultry at lower ones), supplemented by hunting and collecting. Aggregation of human settlement was inhibited by poverty of the soils which could not, given existing technology, support large populations from local resources; by the near-absence of transport facilities capable of profitably importing high-bulk low-value commodities; and by the presence of epidemic diseases whose occurrence was exacerbated by urban population densities. Indeed, the minerals themselves are to be found in places seldom settled by humans, in the ridges between river systems rather than on the riverbanks which were preferred by African cultivators.

Although European prospectors sometimes identified mineral deposits before colonial conquest—deposits which had often been previously worked by Africans—the transformation of these environments took place under European minority rule. These minority governments all relied on their ability to coerce the earlier inhabitants into a variety of activities which they would not have otherwise chosen: migration away from lands desired by Europeans or toward work for those Europeans; and the payment of taxes in European money which frequently could be obtained only through the sale of the labour of able-bodied men.

Despite these commonalities, European governments took a variety of forms: those controlled by settler minorities in South Africa and in Southern Rhodesia after 1923; arbitrary regimes, including those run by the British military during and just after the Boer War and by the British South Africa Company for the Rhodesians between 1890 and 1924; and bureaucracies dependent on a colonial ministry in the European metropolis. These latter, although resembling one another in their hierarchical organization, showed enormous variation according to nationality: the British created legally distinct regimes for each of their colonies and consciously sought to incorporate African authorities into their

system in the policy of indirect rule; the Belgian government, subjected to threats from larger colonizers, centralized their authority in a single hierarchy and consciously delegated governmental functions to Catholic missions and Belgian companies.

Colonial governments therefore provided a legal framework and guaranteed civil order and a labour supply, which facilitated the transfer of European and American capital to Africa. This capital, in turn, made possible the importation of Western technology which transformed the microenvironments on which they were concentrated; along the railway lines, in urban areas, and at mine sites. These latter underwent the most profound changes due to the importation of mining equipment, energy (coal and electricity), food, water, and medical services (including public hygiene). Decision-making power resembled that of colonial bureaucracies as opposed to that of settler regimes. It lay in the hands of a hierarchically ordered bureaucracy with a line and staff organization whose membership was chosen from superordinate to subordinate by co-optation and which was ultimately accountable to managerial organs in South Africa, Europe, and the United States.

Local managements sought to optimize profits by establishing a delicate balance between output and unit production costs. This involved neither maximization of production—some deposits cost more to excavate than they could bring in as income—nor minimization of labour costs: in most cases, reducing labour costs to the lowest levels was more than matched by declines in productivity. Indeed, workers who were hungry, sick, or dead were of no use to the mining companies. It should be pointed out that this consideration is by no means limited to human behaviour in the mines, but is characteristic of animal communities generally. As one biologist puts it,

the productivity of a species in a particular habitat depends on the level of activity it can maintain and the relative cost of supporting that activity ... Costs include more than energy expenditure; they encompass all factors that increase mortality and reduce fecundity (Ricklefs 1973).²

The mines of colonial Southern and Central Africa in this respect can be considered a congeries of related social and biotic environments.

Socially they were ruled by colonial dominance hierarchies led by white company managers and their chosen African subordinates. Relations between managers and employees and between whites and blacks can be considered a form of intraspecific competition which was regulated by company authority, specialization of work functions, and, to a certain extent, co-operation among all members of the enterprise. African employees were constrained in their choice of adaptive responses by company policies and colonial laws. For example, Masters and Servants legislation restricted their ability to move to more suitable environments. Malingering, in this case the closest human analogue to dormancy, could be punished by beating and deprivation of salary. The principal forms of social adaptation therefore occurred between workers and their immediate superiors (black and white) and among workers (strong-weak, experienced-young, interethnic rivalry); practices of cultural adaptation could be easily learned and spread rapidly through the mining populations.

Collective adaptation to the mines as biotic environments, by contrast, was almost exclusively the purview of white managements. The early managers had little trouble clearing the mine sites of large mammals including carnivores. Smaller organisms, as an almost inverse function of their size, caused greater problems. Snakes, rodents, and insects were impossible to eliminate; fleas, ticks, lice, and mosquitoes communicated their own microscopic pathogens to their human hosts. In the early years of each mine, human populations suffered from epidemic outbreaks of infectious disease. Mine managers

² Vansina (1990:41) finds similar behaviour on the part of the Djue of Cameroon.

could not afford to wait for the development of a natural equilibrium between parasites and human hosts. Mortality and morbidity rates were high and workers eagerly quit the mines for their villages at the end of their work contracts. Even healthy workers were less productive when confronted with the combination of work demands and exposure to microscopic parasites.

Management could therefore not keep the mines running without spending money to diminish morbidity and mortality among workers. By the end of the nineteenth century, Western technology had already developed a number of devices for reducing epidemic illness in temperate climates: quarantines against bubonic plague, vaccination against smallpox, inoculation against diphtheria; treatment of water to reduce typhoid fever and cholera; and the beginnings of treatment of milk and other foods. In addition, tropical diseases such as malaria and yellow fever were proving amenable to vector control and medications such as quinine. Indeed, Colonel William Gorgas showed that 'seasoned' workers living with their families on the Panama Canal works were far less likely to die of pneumonia than a constantly renewed supply of unexposed workers. These measures all cost money and could therefore not be implemented simultaneously, but as they were, mortality and morbidity rates dropped substantially.

As microenvironments, the mines by the 1930s had become healthier than the villages from which miners were recruited. This marked a reversal from the mines' earlier days, when workers in the camps were at greater risk than people in the villages, owing to denser population and concomitant exposure to a broader pool of infectious disease, tainted food, and polluted water. Village economies had, moreover, deteriorated as a result of the periodic removal of able-bodied men from their labour forces. Recognizing the disparities between the villages and the European sector, rural Africans began to migrate voluntarily to urban areas. Since mine managements exercised strict control over the number, sex, age, and health of their workforce, not everyone attracted by the mines could find work there; migrants were diverted from the mines to urban areas where wages were lower and amenities fewer. Thus, alongside the mines developed related colonial microenvironments which lacked the elaborate facilities provided for the mineworkers but which were far better provided than the villages. The boundaries between the environments of the camps and those of the villages are thus not absolute but consisted of a series of thresholds determined in large part by colonial expenditures on public-health facilities for Africans.

Needless to say, the major determinant of where colonial government monies were spent was the location of Europeans, not the location of Africans. For politicians in settler societies, this choice was a matter of electoral survival; even in metropolitan-dominated bureaucracies, however, public amenities were geared to the needs of Europeans, who were deemed both more susceptible than Africans to tropical disease and more vital to the functioning of the colonial society. As a result, European mortality rates were lower than those of Africans.

Public-health facilities were not the only determinant of the Europeans' lower risk of death. Beginning just before World War I, cultural changes relating to personal hygiene began to disseminate in Europe, North America, Australia, and New Zealand. These changes included washing food and hands before meals and a general intolerance for dirt. By the 1930s, these practices coupled with inoculation and public hygiene reduced Western mortality rates for children and young adults to the lowest level known in human history, a phenomenon now known as the epidemiological transition (Ewbank and Preston 1990; Preston and Haines 1991).

Some of the same personal hygiene practices were adopted at this time by a small proportion of the African population of the mining camps although most Africans were not in a position to implement them until the 1970s. They lacked both the physical facilities to implement them and the predictability of daily life necessary to implement new routines. Without running water, ready access to treatment

centres, and money sufficient to guarantee subsistence, mothers could not be expected to take up new health practices such as washing vegetables and hands. Such changes in routine also required a certain residential stability which comes from security of urban tenure.

African adoption of European hygienic practices thus depended on the willingness of European authorities to allow them to live in cities. Paradoxically, the greatest objections came from ostensibly democratic settler governments where voting was limited to the European minority, which was particularly threatened by African urbanization. The 'self-governing' regime in South Africa, notably, did its best to limit African urbanward migration. For that minority of Africans deemed necessary for the maintenance of the urban economy, the South Africans imposed increasingly stringent policies of segregation. African urban residential areas, moreover, did not receive the same amenities as European ones: running water, electricity, and health services. Many urbanites, additionally, lived in cities illegally, renting cramped quarters from 'legitimate' householders. In sum, racial segregation inhibited the adoption of private hygienic practices which would undoubtedly have lowered African mortality rates.

The ultimate rationale for these restrictions was economic. Europeans were migrating to South African cities at the same time as Africans and were often competing for the same jobs. But the poor whites of South Africa enjoyed an important advantage over the poor blacks. The former could vote, giving them the power to remove elected officials who were insufficiently zealous in the protection of European economic privileges. This issue arose soon after the Boer War with regard to employment on the Witwatersrand gold mines. In the very first Transvaal elections under British auspices, General Louis Botha came to power in large part because of his opposition to the mining companies' use of indentured Chinese workers, who were seen as potential rivals for European jobs (Richardson 1982).

The situation arose again during the first major white strikes of 1913–14. The Chamber of Mines, alarmed that high mortality rates would discourage black workers from coming to the mines, asked for the advice of Colonel William Gorgas, who had so successfully reduced mortality for black workers on the construction of the Panama Canal. Gorgas observed that on the Witwatersrand, as in Panama, the principal causes of death were pulmonary diseases, whose incidence was increased in a constantly changing labour supply. He therefore recommended that the mines 'stabilize' the African labour force, that is, encouraging long-term residence at the mines by African men living with their families (Gorgas 1914). His recommendations, which were anathema to the striking miners, were flatly vetoed by the South African government, led by the same Louis Botha, on the grounds that stabilized black labourers might take away jobs from white workers (Meintjes 1970). In South African mines, then, as in the cities, European politics prevented African settlement in microenvironments where Africans would have survived longer than in rural areas.

The question of stabilization took a very different turn in the Belgian copper mines. The management of the Union Minière du Haut-Katanga, originally an Anglo-Belgian venture, was captured by Belgians only at the end of the First World War. The new executive in Africa saw labour problems in a very different light from the former Anglo-Saxon management. Black labour, owing to horrendous conditions during the first decade of operations, was in exceedingly short supply. White labour, by contrast, was relatively abundant and troublesome. White South African trade unionists in Katanga were demanding the privileges enjoyed by their comrades on the Witwatersrand. Unlike South African white miners, however, they had no hold on either the government or the management. They were aliens who could be replaced by Belgians from the Borinage who were eager for the work.

As early as 1921, Belgian managers began to consider stabilizing African labour, and within five years, the Union Minière was attempting to establish a permanently resident African labour force in the mining camps. The government had reasons of its own for encouraging this company initiative. The

original workforce had come primarily from British territories near the mines. This allowed British officials to register complaints against the mining company on behalf of the workers to Belgian authorities. Africans from British territories, moreover, had their salaries paid in sterling rather than in rapidly depreciating paper francs. Potential workers could be found within Belgian territory but at greater distance from the mines. Transportation costs would be minimized if these workers could be brought to the mines and kept there for three-year terms.

The government therefore welcomed the Union Minière's proposal to stabilize its work-force relying on African labour from exclusively Belgian sources. One manager referred to this policy as 'breeding' a labour supply. Despite changing sources of labour supply and the Great Depression, the company succeeded in attracting a labour force of mainly married men whose wives and children lived with them in the camps. Company doctors, social workers, and Benedictine missionaries taught personal hygiene practices, and camp mortality rates, according to company reports, dropped to levels as low as those of Western Europe between the wars (Union Minière 1955). See Figure 1.

Figure 1

Belgian managerial practices were so effective in creating a loyal and productive work-force that British and South African firms operating on the Northern Rhodesian Copperbelt were forced to adopt similar measures in order to prevent their best workers from moving to the Katanga mines (Parpart 1983). By 1960, these camps on both sides of the Copperbelt nearly equalled the health conditions enjoyed by white workers. Labour for the copper mines would never again be in short supply.

In the early 1960s, when Central African territories became politically independent, the ecological distinctiveness of the mining camps was breaking down. North of the Zambezi, the boundaries between camp and town were beginning to disintegrate and the two microenvironments were coming to have

common characteristics. South of the Zambezi, in the Republic of South Africa in particular, the distinction between town and camp remained. Either as the result of government pressure or because of perceived economic interest, the South African mines remained basically unisex operations dependent on short-term migration from rural areas in the Republic and its neighbours. In Central Africa access to urban areas, too, was limited but not to the degree practised in South Africa. The local dominance hierarchy had maintained its pre-eminence in these microenvironments and would continue to do so until the 1980s.

The ecological framework has proved particularly useful for the analysis of the recent history of the mines of Southern and Central Africa. It brings together the problems of the social environment with those of the biotic environment. It should also prove a powerful tool for understanding subnational differences in industrial societies.

References

- Abrahams, J.C. 1937. *Report on Nyasaland Natives in the Union of South Africa and in Southern Rhodesia*. Zomba: Nyasaland Government.
- Ewbank, D. and S. Preston. 1990. Personal health behaviour and the decline in infant and child mortality: the United States, 1900-1930. Pp. 116-149 in *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health* (2 vols.), ed. John Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid, Daphne Broers-Freeman. Canberra: The Australian National University.
- Fetter, B. 1976. *The Creation of Elisabethville, 1910-1940*. Stanford: Hoover.
- Fetter, B. 1983. *Colonial Rule and Regional Imbalance in Central Africa*. Boulder: Westview Press.
- Fetter, B. 1986. Relocating Central Africa's biological reproduction, 1923-1963. *International Journal of African Historical Studies* 19,3:463-478.
- Fetter, B. 1987. Decoding and interpreting African census data: vital evidence from an unsavory witness. *Cahiers d'Etudes Africaines* 27,1-2:83-105.
- Fetter, B. 1989. Colonial microenvironments and the mortality of educated young men in northern Malawi, 1897-1927. *Canadian Journal of African Studies* 23,3:399-415.
- Fetter, B. (ed.). 1990. *Demography from Scanty Evidence: Central Africa in the Colonial Era*. Boulder: Rienner.
- Gorgas, W.C. 1914. Recommendation as to sanitation concerning employees of the mines of the Rand made to the Transvaal Chamber of Mines. *Journal of the American Medical Association* 62,24:1855-1865.
- Hardesty, D. 1977. *Ecological Anthropology*. New York: Wiley.
- Higginson, J. 1989. *A Working Class in the Making: Belgian Colonial Labor Policy, Private Enterprise, and the African Mineworker*. Madison: University of Wisconsin Press.
- Ingold, T. 1986. *The Appropriation of Nature: Essays on Human Ecology and Social Relations*. Manchester: Manchester University Press.
- Kuczynski, R.R. 1949. *Demographic Survey of the British Colonial Empire*, Vol. 2. London: Oxford University Press.
- Landers, J. 1990. Age patterns of mortality in London during the 'long eighteenth century': a test for the 'high potential' model of metropolitan mortality. *Social History of Medicine* 3,1:27-59.
- Meintjes, J. 1970. *General Louis Botha: A Biography*. London: Cassell.
- Mouchet, R. 1943. Documents anatomo-pathologiques sur la nosologie de la main d'oeuvre indigène à Elisabethville de 1915 à 1921. *Bulletin I.R.C.B.* 14,2:422-452.
- Northern Rhodesian Chamber of Mines. 1957. *Year Book 1956*. Kitwe.
- Parpart, J. 1983. *Labor and Capital on the African Copperbelt*. Philadelphia: Temple.

- Perrings, C. 1979. *Black Mineworkers in Central Africa: Industrial Strategies and the Evolution of an African Proletariat in the Copperbelt, 1911-41*. New York: Africana.
- Preston, S. and M. Haines. 1991. *Fatal Years: Child Mortality in Late Nineteenth Century America*. Princeton: Princeton University Press.
- Preston, S. and E. van de Walle. 1978. Urban French mortality in the nineteenth century. *Population Studies* 32,2:275-297.
- Rand Mines. 1964. *Annual Report*. Johannesburg.
- Richardson, P. 1982. *Chinese Mine Labour in the Transvaal* London: Macmillan..
- Ricklefs, R. 1973. *Ecology*. Newton: Chiron.
- Szreter, S. 1988. The importance of social intervention in Britain's mortality decline, 1850-1914. *Social History of Medicine* 1,1:1-37.
- Turrell, R. 1987. *Capital and Labour on the Kimberley Diamond Fields, 1871-1890*. Cambridge: Cambridge University Press.
- Union Minière du Haut-Katanga. 1955. Service Médical, Rapport annuel.
- Union Minière du Haut-Katanga. 1957. *Evolution des Techniques et des Activités Sociales*. Brussels: Cuypers.
- van Onselen, C. 1976. *Chibaro: African Labour in Southern Rhodesia*, London: Pluto.
- Vansina, J. 1990. *Paths in the Rainforests: Toward a Political Tradition of Equatorial Africa*. Madison: University of Wisconsin Press.

Malaria: old infections, changing epidemiology



D.J. Bradley

Department of Epidemiology and Population Sciences, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

Abstract

The epidemiology of malaria has always varied between different parts of the world because of widely varying vectorial capacity. Mortality from malaria can be measured from clinical records or the rise of mortality during an epidemic, but better from observing the fall of mortality during control or from the population frequency of protective host genes. Holoendemic malaria may have doubled the infant and young-child mortality rate in Africa in the recent past, but death rates have fallen because of chemotherapy. The epidemiological pattern is changing. In the Sahel, water-resource developments tend to lengthen the transmission season, though less than might be expected, and urbanization tends to decrease transmission in Africa, not in Asia. The spread of multiple drug resistance of the parasites is making case management harder and deaths may rise. Malaria control has always been unsatisfactory in sub-Saharan Africa owing to the highly effective vector. The main current hopes for control are the use of the effective insecticide-impregnated bed nets and better case management. No simple concept of an epidemiological transition can reflect the very diverse changes occurring in human malaria worldwide.

Introduction

Malaria is, and has been, the most important parasitic disease of man in terms of mortality, morbidity and effects upon life generally, using 'parasitic' in its narrower meaning of the animal parasites: worms and protozoa. The malaria parasites are of great antiquity: they evolved along with man so that there are four malaria parasites specific to man, and other related parasites specific to the various great apes, monkeys, and other mammals. In spite of this long-term host-parasite relationship, there have been major epidemiological changes in human malaria over the last fifty years and they are taking place at an accelerating rate. This paper sets out the basic principles of quantitative malaria epidemiology, without which any attempts to understand changes would be meaningless; and describes the main epidemiological changes that are taking place.

An important genetic polymorphism in man, the sickle-cell gene, appears to be maintained by the protection it affords against human malaria and this is used to estimate the mortality due to human malaria, particularly in Africa, in the past. This is compared with the limited historical data on recorded mortality, and the possible causes of falling mortality are discussed. The results are related, though uncomfortably, to the concept of the health transition. The complex interaction of biology of the vector, present health-care, history of the use of medicines and insecticides, migration and environmental change does not fit simply into the idea of a single unidirectional process. There is considerable epidemiological material before these last issues are reached. The presentation focuses on the historical and health-transition issues.

Malaria: background to epidemiology

There are four malaria parasites of man. Three of them, *Plasmodium vivax*, *P. malariae* and *P. ovale*, are the cause of intermittent high fevers in man that can make a person very ill but are rarely fatal. The remaining species, *P. falciparum*, is the cause of malignant tertian, or *falciparum*, malaria which has a substantial mortality if it is untreated, especially in the first or an early attack. Following a single infection, there is intermittent presence of the parasites in the blood for several months. In people subject to frequent infective mosquito bites, the parasitaemia will continue throughout childhood.

Figure 1

The cycle of malaria transmission and course of disease, with determinants

The basic transmission cycle of the malaria parasites is seen in the central part of Figure 1. The sexual stages of the parasite develop in the blood of someone who has been infected for some time. Mosquitoes of the genus *Anopheles* become infected by feeding on that person. The parasites develop in the mosquito at a rate dependent on the ambient temperature but in the tropics usually take about twelve days before the maturing parasites have reached the mosquito's salivary glands ready to infect

the next person on whom the mosquito feeds. It will remain infective for the rest of its life. The interval between the mosquito becoming infected and being first able to pass on the infection is known as the duration of the extrinsic cycle. When someone is bitten by the infective mosquito the parasites pass to the liver, develop in the cells there, and then burst out to infect the red cells of the blood. The parasites multiply, usually synchronously, in the red cells and go through this cycle every 48 hours except for *P. malariae* in which the duration is 72 hours. With each red-cell cycle the parasites increase logarithmically and as the cells burst the classical attacks of fever occur, hence the intermittent fever pattern known as far back as Hippocrates. As the proportion of infected red cells increases, so do the fevers. In *falciparum* malaria the infected red cells become sticky and adhere to the lining of the capillaries of the brain. They use up the blood's oxygen and prevent it reaching the brain so that the patient may go into a coma and die: cerebral malaria.

People who become infected, and are not immune, suffer a classical unmodified attack of clinical malaria with a severe intermittent fever. Their attacks get worse, usually occurring every other day in *Plasmodium vivax* infections, and on alternate days but more irregularly in cases of *P. falciparum*, which gives rise to more severe disease and a mortality of five to ten per cent in those remaining untreated. Malaria morbidity in very highly endemic areas is mainly a feature of infancy and may account for 30–35 per cent of the cases seeking health care at rural dispensaries. It is evenly spread through the year in the forest form and may vary between ten per cent and 80 per cent of childhood fevers in the savannah zone. Some data suggests that mortality may reach its peak in the savannah, where malaria may be combined with malnutrition, maternal overwork, and lack of drugs in the season of maximal transmission.

The basic quantitative epidemiology of malaria

Malaria varies dramatically in its epidemiological pattern. In some places it appears as massive epidemics affecting people of all ages and causing temporary social disruption comparable to a very severe influenza epidemic. This pattern, alternating between large epidemics and scarce malaria, is called unstable malaria. At the other extreme is very stable malaria, where transmission is continuous, or regularly seasonal, and everyone is being infected and reinfected by mosquitoes. Unstable and stable malaria not only differ in their mortality and public-health effects, they are also very different in their responses to attempts at control. In particular, similarly energetic control measures may have completely different outcomes in different areas. These differences are biologically determined and must be understood if naive generalizations are to be avoided: for example, there was a tendency to blame the failure of malaria eradication in Africa on defective health services, whereas the problem was biologically intractable even with good health services.

The reasons for this epidemiological variation can be understood in terms of the basic-case reproduction rate (BCRR) of malaria, a convenient measure of transmission. The BCRR is the mean number of new cases of malaria to which one case will give rise directly after a single passage through the vector mosquitoes, under conditions of no immunity in the human population. Thus with a BCRR of five, one case will give rise to five cases in the next generation and 25 in the succeeding generation of cases; spread will continue potentially until all the people are infected. Conversely, should the BCRR fall below one, the disease will tend to gradually die out: the goal of malaria-transmission control may be formulated as keeping the BCRR below one. In an endemic malaria situation, the supply of uninfected susceptibles will run out and acquired immunity will play a key role. Three characteristics of the female anopheline mosquito largely determine the BCRR: the density of mosquitoes, man-biting habit and longevity.

Mosquito density is conveniently measured in terms of the number of female mosquitoes per human inhabitant of the area. Malaria transmission is proportional to mosquito density, as might be

expected. However, the effects of changes in the man-biting habit and longevity upon malaria transmission are greater. The man-biting habit is the chance that a given female mosquito will feed on man on any one day. Mosquitoes may feed as often as alternate days, and if all meals were upon man, the man-biting habit would be 0.5. For a mosquito that fed only every third day and with only 20 per cent of meals upon man (the rest upon domestic stock) the man-biting habit would be less than that of the other species by a factor of 7.5. However, since transmission is proportional to the square of the man-biting habit because it takes two bites to transmit malaria—one to infect the mosquito and one to infect man again—the second mosquito is 56 times less effective as a malaria vector. Mosquito longevity affects malaria transmission even more. It takes time for the parasite to develop within the female anopheline. The minimum time is temperature-dependent but even in the hot tropics at least ten days must elapse between the bite that infects the mosquito with malaria and the first time the mosquito is able to pass on the infection. Mosquitoes are often quite short-lived, with a steady daily mortality of five to 25 per cent of the population, so that very few mosquitoes that catch malaria survive to pass it on in many vector species. A really long-lived species can however be extremely effective at passing on the infection. In a sense the 'geriatric' mosquitoes are responsible for malaria transmission.

The preceding discussion shows that, for effective malaria transmission, a mosquito species needs to be long-lived and have a high man-biting habit. Both the vector-species complexes that are largely responsible for African malaria transmission feed preferentially on people, feed frequently, and are long-lived in nature. They include the world's most efficient malaria vectors. Any value for the basic-case reproduction rate of malaria (BCRR) in excess of one will allow spread of the infection. In parts of Asia such values as three and five are not uncommon in malarious areas. In sub-Saharan Africa values for the BCRR in excess of 1,000 are found. There is therefore in many parts of Africa a vast excess of capacity for malaria transmission above that required to maintain endemicity.

This very high BCRR is responsible for the key features of malaria in much of Africa: it is extremely hard to control transmission effectively since so great a proportionate reduction in transmission, perhaps a thousandfold, is needed before the infection dies out; secondly, everyone tends to become infected at an extremely early age and exposures to infection are usual in the first year of life; thirdly, because of the high level of transmission, the natural variations in the determinants of transmission from one year to another never bring the BCRR down close to one so that malaria is always highly endemic. This situation is called 'stable' malaria because it varies so little, and malaria epidemics are unknown in much of the continent: in other words it is 'too unhealthy for epidemics'! The chief determinant of the pattern of malaria seen in Africa is thus the pattern of acquired immunity to human malaria. If this did not occur, in many regions of Africa all the inhabitants would have malaria parasitaemia every day of their lives. The pattern of infection under the extreme of stable perennial transmission is called holoendemic malaria.

Malaria transmission and community impact

The relation between malarial transmission and its disease effects on the community has been the subject of some of the most vehement controversy in tropical health. This has centred on the problems of African holoendemic malaria—the effects of epidemics have been relatively unequivocal—and conflict was probably heightened by underlying differences of philosophy between the two schools of thought involved. One view was that an equilibrium is reached between parasites and the human community under the conditions of holoendemic malaria; this was most thoroughly propounded by Wilson, Garnham and Swellengrebel (1950). They were impressed by the limited morbidity and lack of mortality among indigenous adults, conscious of the difficulty of interrupting transmission with residual insecticides, and fearful of the increased lethality of the disease that might follow unsuccessful control attempts. Their philosophy could be represented as not interfering with a naturally-established

equilibrium and had conservationist implications. By contrast, the other group, led by Macdonald whose 1951 paper is a model of forensic logic, and including Russell and Soper, were convinced of the need to vanquish malaria, and were confirmed interventionists with a strong commitment to eradication.¹ Their philosophy saw nature as red in tooth and claw, and they were conscious of the price in infant deaths that had been paid to achieve equilibrium by adult life. Macdonald in particular had been immensely impressed by the fall in infant mortality in Freetown, Sierra Leone, that followed larviciding operations, as he had worked there (Macdonald and Chowdhuri 1926). The 'interventionists' can be said to have won the day in 1950, and their research on the effects of control measures has provided much of the best evidence for the impact of malaria upon mortality, discussed below.

Sources of information on consequences of malaria

The problems of diagnosis in rural clinics mean that routine reporting is not a source of reliable data. Nevertheless, information from numerous routine sources, particularly from West Africa, show 'malaria' to be one of the major clinical problems. Of eleven common diagnoses in Sierra Leone, malaria accounted for 18–35 per cent of the year-of-age specific diagnoses in young children at an outpatient department; it was the commonest cause both of hospital admission and of death in a hospital in north-west Uganda, though it occupied a much less prominent position in relation to adult patient diagnoses. In paediatric health-care facilities throughout tropical Africa, the three predominant causes of ill-health recorded are respiratory infections, diarrhoeal diseases and malaria, though the precise order in which they are placed depends on the region, the age-category under study and the type of health-care facility. Where more sophisticated approaches to the overall burden of ill-health are used, as has been done in Ghana (Ghana Health Assessment Project Team 1981), malaria emerged as the chief cause of loss of days of healthy life, an aggregate measure of the consequences of mortality, illness and disability. Malaria was responsible for 10.2 per cent of all healthy life lost from disease, amounting to 33 days per person per year, though of course much of this was due to deaths in infancy from malaria. Where attempts were made to transform this into economic loss in a fairly direct way, malaria remained the leading cause of loss with discount rates of zero and five per cent, falling to fifth place at a rate of ten per cent.

Scale of malarial mortality

There are three main ways of attempting to measure mortality due to a disease such as malaria: from clinical records as to the cause of death, from observing the rise in mortality during malaria epidemics and by determining the fall in mortality when malaria is brought under control. Molineaux (1985) has reviewed some examples of the consequences of malaria control and has also added a fourth method of assessing malaria mortality which depends on calculating the mortality necessary to maintain the observed level of the sickling gene in a balanced polymorphism.

The scarcity of post-mortem series in Africa and, more seriously, the extreme bias introduced because they only derive from tertiary-care facilities and very rarely include young children and infants, mean that what is usually viewed as the best possible source of accurate data is not useful for assessing malarial mortality in Africa or other developing countries. Clinical records of cause of death are equally unsatisfactory as most people die outside hospital; malaria is a very treatable condition unless it presents at hospital very late, and only those exceedingly ill are likely to be admitted. Add to this the scarcity of paediatric beds in many hospitals and it is clear that information on death certificates will be a poor guide to what is happening in the community. Nevertheless, malaria is the commonest cause

¹ In the case of Soper, well-founded in experience in Brazil and Egypt.

both of admission and of death in children under five years of age in a Ugandan community hospital (Williams, Hayes and Smith 1986). Epidemic malaria is rare in Africa, and relevant data come from outside the continent where the great epidemics of *falciparum* malaria in Punjab and Sri Lanka led to dramatic changes in mortality. In Sri Lanka the quarterly crude death rate at the height of the epidemic rose by a factor of 3.4, implying that two-and-a-half times more people died of malaria than of all other causes combined during that period. The Sri Lankan data on vital rates are sufficiently good to have been used in numerous, and controversial, analyses of the impact of malaria, discussed critically by Molineaux (1985). Perhaps the most informative picture comes by comparing the least and most malarious districts over a long period that includes the introduction of malaria control. This suggests that the crude death rate of the highly malarious district was consistently double that seen in the other, though the two had identical death rates after malaria control.

The best African data on malarial mortality derive from local field-research projects that have attempted to stop malaria transmission, usually by means of residual insecticides, though some recent projects have relied increasingly on chemotherapy and chemoprophylaxis. Of particular interest have been those where observation has continued after attempts at transmission control have been abandoned. All are research-scale because national-scale interruption of transmission has not been feasible in sub-Saharan Africa. The exceptions have been small islands, or areas at the edge of transmission due to temperature, in turn due to high altitude or location towards the southern limits of transmission; or at the northern limits from aridity. None of these situations is generally informative about the African situation. Three major field projects have used residual insecticides to lower drastically the transmission of malaria over substantial but research-scale patches of Africa.

The Pare-Taveta scheme was the earliest, with spraying from 1955, and in broad design the most elegant, since after several years the intervention ceased and the degree of reversion could be observed (Bradley 1991). Unfortunately, emphasis was originally placed on research areas other than infant mortality, there was a hiatus in demographic data collection, and the long-term demography was carried out subsequently in a slightly different but more detailed way. Striking findings on growth and morbidity were the apparent absence of malarial effect. Malaria was, however, clearly shown to be of importance in producing a substantial and widespread reduction of the haemoglobin level.

Mortality data, initially on a limited population, showed substantial year-to-year variation but a clear and large fall in the first year of full spraying, 1956; death rates remained down for the spraying years to 1958. After the spraying ceased and demographic-data collection had been resumed on a larger scale (1962–1966) there was a small mortality rise in the infant and over-40-years age groups, and a large reversion to the previous high death rates in the 1–4 years age group. It should be noted that malaria transmission only gradually resumed when spraying ceased and had not reached former levels even by 1966. Also, concern had been felt about stopping the spraying and there followed deliberate attempts to make antimalarials available to the population, so that not only was a complete reversion of vital rates highly undesirable, but efforts were made to prevent it. Other changes in socioeconomic progress were taking place, and drought also complicated the scene. Nevertheless, the immediate, precipitous fall in death rates, especially in infants and young children, as soon as spraying was complete, is very convincing. In particular, the infant mortality rate fell from a level of 165–260 before control to 78–132 after it had been instituted, a fall of 108 which amounted to a halving of the infant mortality rate. The young children aged 1–4 years also had a halved mortality, though it was of course much smaller than the IMR; falls in deaths in older people were less dramatic.

A second similar study, in the Kisumu area of Kenya, to the north-east of Victoria Nyanza, took place nearly 20 years later and used Fenitrothion to achieve a 96 per cent reduction in malaria

transmission. This was accompanied by a fall in the infant mortality rate of 40 per cent, from 157 to 93 per thousand. The malarial effect was greatest between three and ten months of age (Payne et al. 1976).

The third study, and that best documented and executed, was in the Garki area of northern Nigeria in 1971–1973, with extremely intense seasonal-malaria transmission. Intervention was by residual insecticide spraying with Propoxur, backed up by mass drug administration for part of the area, which had a very substantial effect in reducing transmission. Infants were not given drugs unless they were found to be infected (Molineaux and Grammiccia 1980; Molineaux 1985). The most dramatic findings were a large fall in the infant mortality rate from 255 to 55 and 102. Comparable unprotected villages had an IMR greater by 80 and 90 in the two intervention years. The death rate of children aged 1–4 years was less than half that in unprotected villages as was the crude death rate. Moreover, in the absence of protection there was a close seasonal parallel between the IMR and the rate of conversion of infants to parasite positivity, with the IMR about ten per cent of this rate-of-incidence measure. Under protection the IMR both fell and lost its seasonal peaks.

In Senegal, the first of three chemotherapy and chemoprophylaxis trials (Garenne, Cantrelle and Diop 1985) halved the mortality of those aged six months to 35 months, but had no effect on those older, or younger. Moreover, there was a large fall in the diagnosis ‘fever and malaria’ as the stated cause of death. As this was rather precisely balanced by a rise in the ‘miscellaneous or not known’ category for other causes of death, Carnevale and Vaugelade (1987) have cast some doubt on this interpretation.

Subsequent studies in the Congo at Kinkala, and in Bobo Dioulasso (Burkina Faso) as studied by Baudon et al. (1984), showed very little effect of intense chemoprophylaxis and chemotherapy upon mortality and are difficult to reconcile with the weight of other evidence except on some of the hypotheses considered below.

Historical assessment of malaria morbidity: inferences from genetics

The gene for sickle-cell anaemia, when present in homozygous SS form, is effectively lethal before reproduction. For this gene to be maintained in the population, the heterozygous AS form must have a selective advantage over the normal homozygous AA. It is believed that this advantage consists of a relative protection against the lethal effects of *falciparum* malaria. If the relatively conservative assumption be made that all AS individuals have a zero death rate from malaria, it is possible to calculate the minimum malaria mortality of the AA group, and thus of the whole population, needed to maintain the S gene at any observed equilibrium frequency. The sickling gene reaches a very high frequency in parts of sub-Saharan Africa and may reach a frequency of 40 per cent of sickle-cell trait carriers. At a level commonly encountered in East Africa of 18 per cent of sickle-cell traits, AS (or a gene frequency of 10% S), and assuming complete loss of sicklers before puberty, then an excess mortality of the order of 100 per thousand is required from malaria in the AA genotype people, or an excess of some 81 per thousand in the overall population. In other words, if the total cumulative mortality in the population by puberty is of the order of 300 per thousand, over a quarter of this mortality would be due to malaria to maintain the gene frequency in equilibrium. If any of the AS die of malaria, the proportion of malaria deaths would need to be greater.

An incidental consequence of the way in which the gene frequencies are maintained in balance by a ‘proportional mortality of genes’ is that where the non-specific infant mortality is high, a lower mortality from malaria per thousand live births will be needed to maintain the same gene frequency. To maintain the ten per cent sickle-cell gene frequency (18% sickle-cell trait) requires a mortality of 90 per thousand from malaria where this is the only cause of death, 81 at a non-malaria mortality of 100 per thousand before puberty and 72 at a non-malaria mortality of 200. To maintain 32 per cent heterozygosity in a population with 200 per thousand premature deaths from other than malaria,

requires a malarial mortality of 128 per thousand and 200 per thousand AA individuals, or doubles the mortality in the AA group.

Interpretations of malaria mortality

In reviews of differing aspects of malarial mortality, Molineaux (1985) and Carnevale and Vaugelade (1987) have considered the empirical evidence. Any attempt to put together the apparently conflicting evidence must take one of three broad approaches. One is to deny the validity of some of the data. This is not difficult as all the studies have methodological defects, some bristle with assumptions, and many have broad confidence limits even if the observations are correct. There is therefore much space for the author to choose the conclusions he prefers and to explain away discrepant results. This tends to be Carnevale's approach. The second way of dealing with the range of results on mortality is to assume the existence of confounding variables that affect the scale of the observed mortality that is ascribed to malaria. This view is adopted by Molineaux and developed by Cohen (1988). A third view, perhaps a variant of the second, is to assume heterogeneity of the consequences of equally endemic malaria due to interventions of other types, a position which can explain some of the problems confronting Carnevale.

Molineaux (1985), in an exceptionally careful and imaginative analysis of the data on malaria mortality, deals with two problems. The first is the much greater reduction in mortality observed after malaria control operations than the fall in deaths ascribed specifically to malaria would suggest. This is most apparent in non-African data, from Guyana and Sri Lanka, where an indirect effect of malaria control on mortality, 2–4 times the direct reduction of deaths from malaria, was observed.

A converse effect appears to prevail in two West African studies, where the removal or massive reduction in deaths from malaria has led to a much smaller fall than expected in infant and young child mortality. This was seen in Garki, Nigeria (Molineaux and Grammiccia 1980), where malaria control removed the seasonal peaks of malaria deaths but mortality remained high overall, and in The Gambia when a measles epidemic shifted the peak season of infant mortality from the malaria season without massive effects on total mortality. The most economical hypothesis to explain these results is that of competing risks: a certain number of children are postulated as likely to die anyway, possibly with low birth weights and for other ultimate reasons, and the immediate cause of death may be malaria if present and some other infection if malaria is absent. Both malaria and, say, measles may be sufficient causes and only one is necessary. In either situation, deaths averted by malaria control are not equal in number to deaths due to malaria. With the twin explanations of indirect malarial mortality and competing risks, it is of course dangerously easy to explain most sets of paradoxical results. In a long-term study many other variables also change; this fact limits the ability to refute some of these possible explanations of observed data.

Carnevale and Vaugelade (1987) are primarily concerned with African data that suggest a fall in adult as well as infantile mortality after malaria-control operations, and with the relatively small effects seen in some recent control projects, mainly using chemotherapy, in francophone Africa. Although they argue from the earlier campaigns forwards, it is reasonable to infer that the main starting points were their own and colleagues' recent work in Burkina Faso and in the Congo which showed very little reduction in mortality among children subject to a vigorous chemoprophylactic and chemotherapy regime. The earlier Senegalese study is open to criticism on diagnostic grounds. Carnevale and Vaugelade (1987) then attempt to undermine the conclusions of the three major control programs in Pare-Taveta, Garki and Kisumu, arguing that, first, the decline in mortality is general and progressive, owing to social and economic development along with a broad-based improvement of medical care, and secondly the insecticidal and chemotherapeutic programs are non-specific, contrary to what has been generally asserted.

It is certainly true that several of the communities studied have improved economically and in access to health care over the periods studied and this is likely to have been responsible for some of the changes observed, as discussed above, and the Kisumu changes are difficult to explain on any grounds. But the rapidity of the effects in Pare-Taveta and Garki argue strongly for a specific antimalarial intervention as the primary determinant of the main changes observed and described above.

The argument that residual insecticiding is non-specific in its effects is unconvincing. A more likely hypothesis is that the earlier studies did indeed demonstrate substantial malarial mortality; that the ones, intermediate in time, at Kisumu and at Ngayokhem in Senegal, have methodological problems that make them hard to interpret reliably; and that the recent francophone-country studies showed little benefit from organized chemoprophylaxis or chemotherapy because mortality from malaria had already been reduced by various forms of progress but chiefly by individuals purchasing chemotherapy for their families when febrile.

The studies are all compatible with heavy infant mortality in uncontrolled holoendemic malaria, greatly reduced by individual access to chemotherapy in recent years.

Epidemiological trends

In spite of its epidemiologically 'stable' nature, African malaria is undergoing substantial and rapid change at present in its basic epidemiology, in the pattern of mortality and in the pattern of resistance displayed by *P. falciparum* to antimalarial drugs. It is possible that in the medium-term immunization methods may become available to produce a fourth and massive type of change.

Several types of change in malaria transmission are taking place at present. In the Sahel and savannah areas where lack of surface water prevents transmission for several months of each year, the development of water-storage dams and of irrigation, particularly for rice production, is extending the breeding season for anophelines and increasing mosquito populations. Few new large dams are now being built, but small dams proliferate, often without any form of health assessment, and though their effect on malaria should not be exaggerated, they will have some effect in the more arid areas. Urbanization is proceeding rapidly and it has been predicted that up to 43 per cent of the population of Africa may live in urban areas by the year 2000. Usually the effect of urban development has been to decrease anopheline densities substantially. Pollution of urban surface waters has led to a marked increase in the 'dirty water'-breeding culicines such as *Culex quinquefasciatus* and a fall in the mosquitoes that flourish in clean water, such as the anophelines, and this has been well documented in Dar es Salaam. In South Asia, however, urban malaria is common because of the urban anopheline, *An. stephensi*, which is also a very good vector.

The widespread availability of some antimalarials, chiefly chloroquine, has undoubtedly affected the mortality from malaria in children. In some places, few fevers, whether due to malaria or not, escape partial if not complete courses of chloroquine and this has brought down the number of fatalities. Within the remaining areas of chloroquine sensitivity this trend will continue. It has already been argued above that this provides a better explanation of the apparently changing effects of malaria control on population mortality than do the suggestions of lack of specificity of the earlier insecticidal programs. Sir Ian McGregor (personal communication) has described the appearance of cerebral malaria in teenagers in Banjul, The Gambia, where it was previously a disease of infants. The combination of urbanization in reducing transmission, and frequent chemotherapy or even chemoprophylaxis in diminishing individual duration of parasitaemia, have probably been responsible for this.

If the malaria parasites had remained unchanging, the steadily increasing availability of antimalarials at health-care facilities and shops, and consequently in the home, would be expected to have led to a progressive decrease in malarial mortality so that it would have remained a major cause of

disease but become of limited importance as a cause of death. However, the spread of drug-resistant strains of *P. falciparum* is greatly altering this optimistic picture. Resistance to the folic-acid antagonists proguanil and pyrimethamine emerges rapidly after their widespread use, and pyrimethamine resistance is common in many parts of the tropics. But the crucial resistance has been to chloroquine which has dominated the treatment of malaria in the past because it is efficacious, cheap, relatively non-toxic, and easy to administer. The spread of chloroquine-resistant *P. falciparum* since 1950 has created a new situation. It spread from foci in Colombia and on the Thai-Vietnam border to envelop Southeast and South Asia, and then reached Africa a decade ago. The last five years have seen very rapid spread across Africa. It is common in East and Central Africa and has reached Cameroon, Ghana, and other West African states. It will soon have reached all the states of West Africa.

The second-line therapeutic drug is Fansidar (a mixture of pyrimethamine and a long-acting sulphonamide) which is more toxic and much more expensive, and to which resistance has been clearly described in East and part of Central-Southern Africa as well as in Asia. Alternatives for treatment are also expensive, resistance emerges quickly to mefloquine, and quinine is both toxic and more difficult to administer than other antimalarials. Chloroquine resistance and Fansidar resistance will surely continue to spread and the scanty number of new drugs to be available in the near future (halofantrine, artemether, etc.) suggests that mortality from malaria may cease to decline, and in areas where drugs are already used may begin to rise. The outlook is not good, in the absence of major innovations, among which a vaccine is the item on which most hopes are based.

Changes in approaches to control

The history of malaria was for many centuries about an increasing understanding of the circumstances under which it flourished, together with some preventive action, and a concurrent clinical understanding of the pattern of fever and association with splenomegaly that goes back to ancient China. Therapy improved substantially in the seventeenth century with the introduction of quinine, but otherwise the modern era began in the 1880s. As with all infectious and communicable diseases, in the late nineteenth century the field became dominated by microbiology and the causative *Plasmodium* was described in 1882. Although description of the stages of the life cycle in the mosquito by Ross soon followed, malaria was remarkable for concealing some of its parasitological secrets so that the whole life cycle, including the hypnozoite stages, took a complete century to unravel. However, malaria was equally remarkable for the very rapid translation of parasitological research to preventive action. This was mainly due to the orientation of Ross towards practical control. On demonstrating the mosquito transmission of *Plasmodium*, Ross turned from parasitological research to field control at once and the 'mosquito theory' was put to use within a few years, with great success, particularly by Watson in Malaysia and Gorgas in Panama. The subsequent rate of malariological research steadily increased, and while the frantic pace of modern molecular biological research is a very recent phenomenon, enormous progress was made in the first half of this century in many aspects of malaria, and also of control.

The nineteenth century saw great improvements in the control of communicable disease, chiefly as a consequence of environmental improvements. The provision of a safe piped water supply to every household was the most striking of a series of measures to improve the health of the population as a whole. This may be described as the public-health approach. The public-health approach (in this limited sense) to malaria is reduction in transmission, aiming to protect everyone. An alternative approach to disease control has been to seek to protect the individual. In malaria, this goes back to the use of quinine as a prophylactic and is most clearly seen today as the use of chemoprophylaxis by travellers to malarious areas and the provision of chemoprophylaxis to pregnant women in endemic areas. It is more concerned with preventing death or disease than with protecting the whole community.

It bears a closer relation to clinical medicine than does transmission control, and lends itself to distribution through a curative primary health-care system.

The broad public-health approach of our ancestors looked to environmental control leading to a man-made landscape under careful control and a fertile economy. Into this fits species sanitation: the selective alteration of the environment to prevent breeding of the local anopheline vectors of human malaria, which was so successful in areas of relatively lower vectorial capacity and especially organized plantation agriculture in the tropics. For the first two decades of this century, both environmental control of mosquito breeding and various patterns of chemotherapy were tried. Successes on the grand scale for the Panama Canal, Zambian copper belt, and South Asian tea plantations made the public-health approach dominant for areas where organized control of the environment was feasible, usually where there was a commercial opportunity and investment for development. The elaboration of a sophisticated species-sanitation approach provided a sound basis for this.

The discovery of the residual insecticidal properties of DDT changed the method of control completely and made large-scale area-wide control feasible from the 1940s (Table 1). It also shifted any previous focus on individual protection from bites by personal action towards centralized public-health efforts and strengthened the predominance of transmission control.

Table 1

An attempt to epitomize the decades since 1930 in malaria control

	Malaria
1940	Control
1950	Eradication: attack
1960	Eradication: consolidation
1970	Resurgence
1980	Chaos
1990	Hope

Eradication took this further still in the 1950s onwards. Not only was the interruption of transmission the central goal, but chemotherapy of individuals in the consolidation phase was viewed as an attack on transmission rather than as therapy for individual benefit. There was an extreme loss of interest in the malaria patient as such, and also in research. Eradication of malaria was treated rather as a military campaign; many of the workers were not from the health-care system and the program was usually kept separate from the remainder of the health services, at any rate until well into the consolidation phase.

Malaria became, in a unique sense, the World Health Organization's disease. It staked its reputation and half its staff on eradicating malaria. The initial successes of the eradication policy, the lasting benefits for many countries, and the subsequent difficulties in the consolidation phase over time, are well known and need no repetition here. The resurgence of malaria in India (Figure 2) illustrates what was seen elsewhere throughout Asia, while eradication was scarcely attempted in much of sub-Saharan Africa.

Figure 2

In the long and disorderly sorting out of policies during the decades of resurgence and chaos (see Table 2) there has been a shift towards a risk approach, in several steps. This approach of preventing deaths and shortening illness was conceded initially with reluctance to communities otherwise unable to cope with controlling malaria, and has gradually become more vigorously advocated by WHO. Much of this has been related to resolving tensions between malaria-control strategies and the primary-health-care concept which succeeded malaria eradication as a driving ideology at WHO, though both ideas are now moving to more empirical approaches, partly driven by economic pressures.

Table 2a

Recent estimates of the numbers of people exposed to or suffering from malaria within a year, expressed as an annual period prevalence

Exposed	2,073 million
Infected	270 million
Ill	110 million

Died

1 million

Table 2b

Current risk of malaria in relation to past eradication efforts in the world. There is considerable lumping of data by countries which exaggerates the 'reduced' category. Population in each of four categories

Never malarious	1,300 million
Eradicated	800 million
Reduced	2,120 million
No control	370 million

Nobody can look at the list of available antimalarials today with any complacency: the safe drugs are rapidly becoming ineffective, and the more effective drugs are either new or relatively toxic or both, and are expensive. New drugs may well be followed by parasite resistance, or previously inadequately perceived toxic effects may emerge, as was the case when amodiaquine was used for prophylaxis on a larger scale. And the list of candidate new drugs is tiny. As chloroquine resistance spreads to cover Africa, what remains that is both safe and cheap? The situation cannot change overnight as the gestation period for a new drug is long—nearer a decade than a year—and few join the queue to enter.

Hopefully, the present decade will see a transition from chaos to hope in malaria control. Vaccines could transform the scene, but even if not, there are advances taking place in many aspects of malaria control, such as the use of insecticide-impregnated bednets. As we learn the new, we must not forget the old. It is now necessary to combine these various accounts of epidemiological and control processes and of historical events into a coherent whole and in relation to the idea of a health transition.

In summary, there have been substantial changes in the mortality from malaria both in the African holoendemic situation where availability of personal chemotherapy has increased, and in Asia and Latin America where eradication attempts greatly reduced transmission and where mortality from resurgent malaria has again been handled by chemotherapy.

In finer detail, transmission has only fallen in those parts of Africa where urbanization has polluted the vector-breeding sites, where desertification has greatly curtailed the transmission season and where malaria control has been practised at the altitudinal or latitudinal limits of transmission. As transmission persists at a much higher level than is needed to maintain all the population infected (or subsequently partly immune), there continues to be a great threat to life of the young, comparable to that observed in the past in the Pare-Taveta study or inferred from the sickling frequency. This is only held in check by early chemotherapy of those ill from malaria. With the spread and intensification of chloroquine resistance, followed by multidrug resistance, in *P. falciparum*, malaria mortality is tending to rise again and this is likely to continue with the relative paucity of new drugs entering the market, and their high cost, unless an effective and cheap vaccine becomes available.

In the other malarious areas of the world, transmission was dramatically reduced, and almost halted in places by the eradication programs of the period 1950–1970. Both mortality and infection became very low. In Europe, the USA, and several small islands eradication was successfully achieved. In South Asia, Indonesia, and other large areas a massive resurgence of transmission took place following the collapse of the eradication programs, so that, for example, the annual incidence of malaria

in India, around 100 million cases a year in the 1930s, fell to 200,000 at the best year for eradication, and then rose again to reach levels of the order of 20–40 million annually after allowing for underreporting of cases. Urbanization did not reduce malaria as *An. stephensi* is an urban breeding mosquito and a good malaria vector. Mortality did not rise proportionately with the resurgence of transmission, presumably because of chemotherapy, but the spread of chloroquine resistance is making this more difficult. Nevertheless, the level of transmission is lower than in sub-Saharan Africa and this makes control an altogether more tractable proposition.

Changes in malaria 1930-1991

The course of malaria as a world problem has therefore been one of massive incidence and prevalence with great mortality up to the 1940s, with control taking place in limited areas of commercial agriculture or urban development where funds were available and the level of transmission was not too high. Some of these situations, such as parts of the USA and Europe, were in the demographic and health transitions, but others such as the tea-gardens of South India could not meaningfully be so described. No impression was made on sub-Saharan African lowland malaria where malaria mortality might approach 50 per cent of infant and young-child deaths, or a rate of over 60 per thousand births.

The decade following the Second World War saw the use of DDT on an increasing scale. Around the Mediterranean almost total control was achieved in countries of very varied standards of living, as also happened on various small islands in the Caribbean and elsewhere, regardless of where they were in relation to a demographic transition. Success in control, short of eradication, then followed in many relatively poor countries of Asia and Latin America through the use of residual insecticides, chiefly DDT, often paid for by foreign-aid funds. Good results needed a rudimentary health service capable of being improved and a low or medium BCRR. Temporary successes were recorded in countries as poor as Bangladesh, Sri Lanka and India.

From 1970, eradication had been achieved in European countries and the USA, places with relatively high-income, well-organized health services, and also less efficient malaria transmission or a short transmission season. Successes were also attained on some islands without those criteria. In Asia, difficulties in getting rid of the last cases through problems at the consolidation phase first prolonged that phase and eventually were followed by collapse of the eradication programs, which were rigid, and conspicuously lacking in research support. Failure was variously blamed on insecticide resistance of the mosquitoes, administrative failure of the consolidation case-finding program, resistance to drugs by the parasites, generally weak health services, and so on. The process was complex and often country-specific. Eradication is a high-risk activity that has to be oversold to have any hope of success and then must maintain activity and funding in the face of a disappearing health problem, and push eradication to completion in less than a decade, before political will is exhausted. In many poorer countries the remarkable fact is not that eradication failed but that it came so close to success. The outcome also depended on the BCRR as this determined the difficulty of the task. The same effort in Sri Lanka and Sierra Leone would eradicate malaria from the former while having a negligible effect on the latter. For practical purposes, sub-Saharan tropical Africa wisely did not participate in the eradication program and the holoendemic malaria there persisted.

During the 1970s and 1980s resurgence of malaria occurred in those countries where eradication had not succeeded. *P. vivax* preceded *P. falciparum*, as is usual in malaria epidemics. The level of endemicity reached was usually lower than in the 1930s and in particular the mortality appeared much lower, though official reluctance to acknowledge the failure of the program and the general collapse of morale and malaria control may have greatly impeded reporting. The overall impression gained is that malaria mortality is substantially reduced, probably from widespread self-medication with chloroquine.

With the massive spread of *P. falciparum* resistance to chloroquine, and often multi-drug resistance, it is likely that mortality will rise again, certainly in Africa.

Malaria and the demographic and health transition

Clearly, the relation between malaria and the so-called health transition and the demographic transition is complex. In African holoendemic malaria, significant control clearly affects the demographic transition because malaria is such a major cause of infant mortality. Even where this is less obvious, as in Sri Lanka, the near-eradication period was characterized by a great rise in fertility and population growth.

Apart from the direct contribution of malaria to mortality and fertility, where it is driving the demographic processes that lead to the transition, malaria control is relatively autonomous as a health variable. Its feasibility is determined by the biological BCRR which depends on local mosquito biology. Certainly, because control requires resources and a health-care delivery system, it is related to economic development. It does not, however, disappear 'spontaneously' as incomes rise, it requires a specific and well-organized effort to control transmission, though mortality does tend to fall once the appropriate chemotherapy is widely available.

But the malaria parasite is also capable of adaptation, and drug resistance may result in malarial increases that would not tidily fit a health transition that went one way. Conversely, moderate effort on islands with a relatively low BCRR can eradicate malaria at a stage long before the other aspects of the health transition would occur.

Thus, for malaria, as for other vector-borne infections, any simple concept of a demographic and health transition fits uncomfortably. It is truer for malarial mortality than for malarial infection. But biological factors which show immense areal variation play a major role, and both recent technical advances and ancient genetic mechanisms confuse the simple picture. Malarial mortality reduction may be part of the health transition, but the richness and diversity of malaria epidemiology and control cannot be contained within so simple and general an idea.

References

- Baudon, D., J. Roux, P. Carnevale, J. Vaugelade, C. Boudin, J. Chaize, J.L. Rey, M.B. Meyran and O. Brandicourt. 1984. Etude de deux stratégies de contrôle des paludismes, la chimiothérapie systématique des accès fébriles et la chimioprophylaxie hebdomadaire dans 12 villages de Haute-Volta, en zone de savanne et zone rizicole de 1980 à 1982. *Technical Document* No. 8450. Bobo Dioulasso: Organisation de Coordination et de Coopération pour la Lutte contre les Grandes Endémies.
- Bradley, D.J. 1991. Malaria. Pp. 190–202 in *Disease and Mortality in Sub-Saharan Africa*, ed. R.G. Feachem and D.T. Jamison. New York: Oxford University Press.
- Carnevale, P. and J. Vaugelade. 1987. Paludismes, morbidité palustre et mortalité infantile et juvénile en Afrique sub-saharienne. WHO/MAL/87.1036. Geneva: World Health Organization.
- Cohen, J.E. 1988. Estimating the effect of successful malaria programmes on mortality. *Population Bulletin of the United Nations* 25:6–26.
- Garenne, J., P. Cantrelle and I.L. Diop. 1985. Le cas du Sénégal (1960–80). Pp. 307–330 in *La Lutte Contre la Mort*, ed. J. Vallin and A. Lopez. Travaux et Documents no. 108. Paris: Presses Universitaires de France.
- Ghana Health Assessment Project Team. 1981. A quantitative method of assessing the health impact of different diseases in less developed countries. *International Journal of Epidemiology* 10:73–80.
- Macdonald, G. 1951. Community aspects of immunity to malaria. *British Medical Bulletin* 8:33–36.
- Macdonald, G. and K.L. Chowdhuri. 1926. Malaria in the children of Freetown, Sierra Leone. *Annals of Tropical Medicine and Parasitology* 20:239–262.

- Molineaux, L. 1985. La lutte contre les maladies parasitaires: le problème du paludisme, notamment en Afrique. Pp. 111–140 in *La Lutte Contre la Mort*, ed. J. Vallin and A. Lopez. Travaux et Documents no. 108. Paris: Presses Universitaires de France.
- Molineaux, L. and G. Grammiccia. 1980. *The Garki Project: Research on the Epidemiology and Control of Malaria in the Sudan Savanna of West Africa*. Geneva: World Health Organization.
- Payne, D., B. Grab, R.E. Fontaine and J.H.G. Hempel. 1976. Impact of control measures on malaria transmission and general mortality. *Bulletin of the World Health Organization* 54:369–377.
- Williams, E.H., R.J. Hayes and P.G. Smith. 1986. Admissions to a rural hospital in the West Nile District of Uganda over a 27 year period. *Journal of Tropical Medicine and Hygiene* 89:193–211.
- Wilson, D.B., P.C.C. Garnham and N.H. Swellengrebel. 1950. A review of hyperendemic malaria. *Tropical Diseases Bulletin* 47:677–698.

Lessons for the past: Third World evidence and the reinterpretation of developed world mortality declines



Peter Aaby

*Center for Prevention and Control of Infectious Diseases and Congenital Disorders,
Statens Seruminstitut, 5, Artillerivej. DK-2300 Copenhagen S. Denmark*

Abstract

Measles is the largest single killing infection of children in the world, and it is likely that its toll is heightened by the occurrence of a serious delayed effect of early infection. Variations in measles mortality, and in the seriousness of infection, have often been explained in terms of nutritional factors, but intercountry comparisons within the Third World fail to bear this out. In this paper an alternative interpretation is developed, based on the severity of the infective dose to which individuals are exposed. It is shown that this can account for a wide variety of observations which are inexplicable on the nutritional hypothesis, and can also explain the severity of virgin-soil epidemics without reference to the effects of genetic selection. The exposure hypothesis predicts that measles vaccination should have a marked effect on childhood mortality as a whole, and this prediction is born out in practice. The success of this interpretation has important implications for our understanding of historical mortality declines in the developed countries, particularly the decline in smallpox mortality.

Introduction

The interpretation we adopt of morbidity and mortality patterns in the past necessarily reflects our current understanding of the biological and social causes of severe infection and high mortality. Given the belief of the 'Second World War generation' in the importance of nutrition, hygiene and medical care, it is no surprise that declining mortality in the industrialized world should have been interpreted in such terms. The relative importance of these factors has been the subject of much discussion, and nutritional improvement (McKeown, Brown and Record 1972), sanitary reform and improved personal hygiene (Razzell 1974), and advances in treatment stemming from medical research, have all had their partisans. Other writers have argued, more modestly, that no good explanation yet exists for the dramatic fall in the severity of many infectious diseases in the industrialized world before the advent of antibiotics (Lancaster 1952).

Underlying these different interpretations is a rather limited set of factors which are, in principle, capable of explaining why infections that are nowadays relatively harmless should so often have proved fatal in the days of severe child mortality. These factors can be grouped under three headings: **'host factors'**, aggravating the infection: malnutrition, young age at infection, genetic susceptibility; **'transmission factors'**: greater exposure, more virulent strains of pathogen, synergism between infections; and **treatment and medical-care factors**: ineffective or aggravating treatment, non-use of effective treatment and preventive activities.

Most attempts to explain major historical or regional variations in child mortality patterns focus on one, or a few, of these factors, and it is usually the first group, the host factors, which are singled out (McKeown et al. 1972; Reves 1985). This choice has obvious implications for how we conceptualize the social production of health (Mosley and Chen 1984), and for our understanding of the historical process of health transition. Clearly, how we understand the mechanisms of severe and fatal infection

will affect the direction of sociocultural and historical investigations. If this understanding is not correct the latter are unlikely to generate information of any use in public-health planning or in development projects aiming at health improvements.

I shall argue, based on my work in Guinea-Bissau, that interpretations based exclusively on host factors are likely to be too limited in their scope; if not simply wrong. Research on the determinants of measles mortality suggests that a number of other factors are more capable of explaining the known regional and historical variations in its severity. These factors are probably implicated in other disease too, and they may be 'good to think with' when we try to understand the social production of health in general.

In the present paper the alternative mechanisms potentially responsible for severe forms of infection ('severe infection') will be briefly summarized. It would be interesting to apply the same principles to historical data on health, but my work in Guinea-Bissau has left me no time for primary research on such material. I shall therefore merely attempt to show the usefulness of this approach by reviewing current interpretations of some historical health issues: the causes of the decline in measles mortality in the industrialized world, sex differences in severity of infection, the causes of particularly high mortality in virgin-soil epidemics, and the role of smallpox control in the industrialized world's mortality decline. I hope that this will support my view that the mechanisms governing severe infection must be viewed more broadly if we are to understand the socioeconomic and cultural processes which produce health and disease.

Host or transmission factors?

Measles is the single largest killing infection in the world where children are concerned, with a death toll currently estimated at around 1.5 million each year (Aaby 1989, 1991a). Our understanding of this infection and variations in its severity thus has important health-policy implications. For measles, as probably for many other infections, the dominant approach has been to search for distinctive features among those individuals who die or suffer a severe infection, that is, the host factor approach (Aaby 1991a). It is assumed that there must be something wrong with these people since they die where most others survive the illness.

Those who die are seen as somehow weaker than other individuals, reconfirming our underlying cultural notion of 'survival of the fittest'. Thus, severe measles has been explained particularly with reference to malnutrition, the age at infection, genetic susceptibility and underlying diseases. There may be something to the emphasis on host factors in the developed countries since here very few individuals die, and these are mostly children previously affected by an underlying disease such as leukemia, AIDS or hereditary problems such as Down's Syndrome. But this perspective has little to offer where major regional and historical variations in case fatality rates (CFRs) are concerned.

Thus the nutritional hypothesis would lead us to expect much higher measles mortality in Bangladesh than in, for example, Guinea-Bissau as the state of nutrition is substantially better in the latter country. However, as Table 1 shows, the reverse is the case. The CFR is many times higher in Guinea-Bissau than it is in Bangladesh. Furthermore, there are no community studies to show unequivocally that nutritional state prior to infection has an effect on the latter's outcome, and there are several studies which have shown that it had no effect (Aaby 1988b).

Table 1
State of nutrition (weight/height) and measles case fatality ratio (CFR), children under three years of age only

Weight/height	Measles CFR
---------------	-------------

	%	%
Bangladesh	87	3
Guinea-Bissau	97	24

CFRs are highest among the youngest children, and this has led some writers to argue that age at infection is a major determinant of the level of severity of measles in the community, and hence that changes in age at infection could explain changes in severity (Reves 1985). It has thus been predicted that measles mortality will be higher in urban areas than in the countryside as the age at infection is generally lower in the former (Davis 1982; Foster 1984). These propositions are not supported by data from developing countries. All community studies suggest that the CFR is higher in rural areas than in urban areas (Aaby 1988b). Within a region, CFR may in fact be higher where mean age at infection in the community is highest. Thus we have shown that mortality risks from measles in rural Senegal are highest in the smallest villages where the interval between epidemics is longest, and the age of infection is highest (Pison, Aaby and Knudsen 1992).

Transmission factors

Major host factors do not seem to contribute much to a better understanding of variations in measles mortality, and it may be better to look for factors intrinsic to the infection itself rather than to inter-individual differences in levels of resistance.

Intensity of exposure and dose of infection

A common feature found in most circumstances where measles has a high acute CFR is that many individuals have contracted the infection simultaneously. In several outbreaks in Guinea-Bissau we found that age-specific mortality was considerably higher in houses with multiple cases than in those with only a single case (isolated cases) as is shown in Table 2. This association has been confirmed in all the other studies in which it has been tested (Aaby 1988b, 1989), and although in principle it might arise from underlying differences in the characteristics of households with different numbers of children, it has been possible to test empirically whether it reflects variations in the intensity of exposure to infection.

Table 2

Case fatality ratio in measles infection according to age, clustering and type of exposure, Bandim, Guinea-Bissau, 1979

Age (months)	Case fatality ratio %			
	(deaths/no. ill)			
	Isolated cases		Houses with multiple cases	
	%		Index cases	Secondary cases
			%	%
0-5	0	(0/1)	0	24 (4/17)
6-11	14	(1/7)	0 (0/15)	42 (11/26)
12-23	11	(2/19)	21 (3/14)	33 (14/43)
24-35	0	(0/10)	14 (2/14)	38 (14/37)
36-59	0	(0/10)	5 (2/38)	13 (5/39)

60+	33	(1/3)	6	(2/36)	0	(0/50)
Total	8	(4/50)	8	(9/117)	23	(48/212)

Source: Aaby (1989).

All of the isolated cases can be classified as 'index cases' infected from someone outside the home. Some of the multiple cases are secondary cases infected by an index case in the same house, and others are index cases who have contracted the infection elsewhere (Table 2). Where several cases occur in the same house the secondary cases will, presumably, have been more intensively exposed than was the index case. Difference in the severity of infection between index and secondary cases have now been examined in several studies from developing countries, as well as in reanalyses of historical data from Europe (Table 3). All these studies have found a significantly, or nearly significantly, higher mortality among secondary cases. Usually the latter's CFR is at least two to three times higher than that of the index cases.

In the studies from Guinea-Bissau, isolated cases display very similar mortality rates to those of index cases in multiple-case households (see Table 2) suggesting that the contrast between single and multiple cases is not due simply to differences in genetic constitution, sociocultural and therapeutic practices, or the risk of complications in families of different size. The difference between index and secondary cases could, however, be confounded by higher general mortality in the larger families which would also have a higher proportion of secondary cases. This possibility has been examined in studies from Kenya (Aaby and Leeuwenburg 1990) and Copenhagen (Aaby 1988a), and the relative risk (RR) between index and secondary cases was found to be the same regardless of family size.

Maternal fatigue could also be a confounding factor if mothers provided less care for secondary cases because they already had to care for an index case. However, an analysis of data from several epidemics in Guinea-Bissau showed that the mortality of secondary cases who had the same mother as the index case (25%, n=55) was no higher than that of other secondary cases (42% n=60) (Aaby et al. 1988c). Thus, differences in maternal care due to fatigue are unlikely to explain the excess mortality of secondary cases.

Table 3
Severity of measles infection according to type of exposure in different community studies

Country (reference)	Age	Case fatality ratio % (deaths/no.ill)				Relative risk (95% C.I.)	
Urban studies							
Guinea-Bissau (Aaby 1989)	0–4	8	(10/128)	30	(48/162)	3.8	(2.1–6.7)
Guinea-Bissau (Aaby et al. 1988a)	0–4	3	(1/37)	38	(10/26)	14.2	(3.4–59.5)
Guinea-Bissau (Aaby et al. 1988a)	0–4	5	(3/66)	17	(14/81)	3.8	(1.3–11.4)
England (Aaby et al. 1986a)	0–2	8	(4/48)	22	(8/36)	2.7	(0.9–7.8)
Copenhagen (Aaby (1988a)	0–2	11	(28/252)	27	(49/183)	2.4	(1.6–3.6)
Rural studies							
Guinea-Bissau (Aaby 1989)	0–4	7	(1/15)	38	(33/86)	5.8	(1.4–24.5)
Senegal	0–4	4	(8/198)	14	(37/226)	3.4	(1.7–6.7)

(Garenne and Aaby 1990)					
Kenya	0–4	3 (11/381)	7 (21/287)	2.5	(1.3–4.9)
(Aaby and Leeuwenburg 1990)					
Bangladesh	0–2	1 (1/134)	18 (4/22)	24.4	(5.6–106.1)
(Koster 1988)					
Bangladesh	0–4	15 (38/2551)	3 (17/630)	1.8	(1.0–3.1)
(Bhuiya et al. 1987)					
Germany	0–14	2 (2/93)	11 (10/95)	3.7	(1.1–12.1)
(Pfeilsticker 1863)					
Ratio of severe cases					
Gambia	0–23	7 (3/41)	31 (4/13)	4.2	(1.2–15.3)
(Lamb 1988)					

Table 4

Case fatality ratio among secondary cases of measles according to the number of index cases, Copenhagen, 1915–1925 and Kenya 1974–1981

Country (reference)	Age (years)	1 Index case	>2 Index case	Relative risk ^a
Denmark	0–4	18 (28/152)	35 (15/43)	1.8 (1.1–3.1)
Kenya	0–6	6 (18/303)	14 (5/73)	2.5 (0.9–6.6)

Source: Aaby and Leeuwenburg (1991)

^aMantel-Haenszel relative risk has been used to stratify for age.

If no confounding factor can be found to explain this excess (Aaby 1989), it is likely to be a biological phenomenon related to the difference in exposure. If so, a dose response effect should be expected. This was tested by examining whether the number of index cases affected the CFR of secondary cases in a study of hospital records from Copenhagen. Here children exposed to two or more index cases proved to have a significantly higher CFR than those exposed to a single case (RR=1.81; 95% CI:1.05 - 3.11, Aaby 1991a). A similar tendency was found in a reanalysis of data from the Machakos area in Kenya; secondary cases exposed to a single index case had a CFR of six per cent compared with 14 per cent for those exposed to more than one (RR=2.5; 95% CI: 0.9–6.6, Aaby and Leeuwenburg 1990).

Following the same line of reasoning we should expect severity to increase with the intensity of contact. We were able to test this expectation in a rural community study in Senegal (Garenne and Aaby 1990). In this area the Serer population live in large compounds divided into several households, and members of the same household may sleep in several different huts. In this setting we assume that the intensity of contact experienced by secondary cases increases gradually, from cases infected by a child from a different household in the same compound, through those infected by a household member sleeping in a different hut, to someone infected by another child sleeping in the same dwelling. As will be seen in Table 5, there was a corresponding increase in the CFR. Most studies have only analysed differences between index and secondary cases in the same house or household, but these results point to the importance of variations in the intensity of contact between infected children and the sources of infection.

Table 5

Measles case fatality ratio by age and intensity of exposure, Niakhar, 1983–1986

Age (months)	Case fatality ratio % (deaths/no.ill)		Secondary cases infected from					
			Same compound		Same Household			
			different household		Different hut		Same hut	
0-5	0	(0/4)	0	(0/2)	0	(0/6)	13	(2/15)
6-41	7	(8/115)	11	(10/89)	12	(18/152)	18	(31/174)
42-65	0	(0/79)	2	(1/50)	6	(4/66)	5	(4/77)
66+	0	(0/122)	0	(0/62)	0	(0/86)	2	(2/129)
Total	3	(8/320)	5	(11/203)	7	(22/310)	10	(39/395)
Odds ratio		1.0		1.9		2.3		3.8

Differences in the risk of becoming a secondary case apparently explain a very large part of observed variations in measles mortality, and mortality is correspondingly high where the proportion of secondary cases is large (Aaby 1988b). For example, the contrast in measles mortality between Guinea-Bissau and Bangladesh—incomprehensible from the nutritional point of view—is quite logical from the perspective of exposure. In Guinea-Bissau 61 per cent of the children under three years of age were secondary cases, and the CFR was 25 per cent whereas in Bangladesh the figures were only 14 per cent and three per cent respectively. Larger families and a high incidence of polygyny mean that children in West Africa have a much greater risk of becoming secondary cases (Aaby 1988b), and the level of measles mortality is particularly high. Consistent with the difference in CFRs, polygyny is found much more widely in West than in East Africa, married men having an average of 1.6–1.7 wives in the former compared with only 1.2 in the latter (Goody 1973).

The apparent contradiction between high mean age at infection and a high CFR may also be explained from the exposure perspective. In rural areas, epidemics occur infrequently, and the number of susceptible children per family thus increases. When an epidemic does arrive these are all likely to fall ill at once, and the risk of intensive exposure is therefore higher in such areas. A high mean age of infection indicates the relative isolation of the community and thus a larger mean number of susceptible children per family (Aaby et al. 1988a; Pison et al. 1992).

The simplest explanation of the higher CFR following intensive exposure would be that the absorbed dose of measles virus is greater (Aaby et al. 1985). In animal studies higher doses have been found to result in shorter periods of incubation. We have thus examined hospital records from Copenhagen in the period 1915–1925, so as to determine the interval between the occurrence of a rash in the index case and those in the secondary cases (Aaby 1992) which should serve as an approximation for the latter's incubation period. Fatal secondary cases displayed a much shorter interval than did survivors ($p < 0.05$, Wilcoxon two sample test) indicating a correspondingly shorter period of incubation. In fact this difference may be understated because the prodromal period has been found to be prolonged in severe cases (Aaby et al. 1986a), and so our method will tend to exaggerate the length of their period of incubation.

These observations are thus all compatible with the hypothesis that the occurrence of severe infection is related to the absorption of a high dose of virus (Aaby et al. 1986a).

Amplification of severity or mildness in the community

Most studies on the severity and transmission of infection have restricted themselves to differences between index and secondary cases (Aaby 1988b). However, since severe cases excrete more measles

virus (Scheifele and Forbes 1972)—and presumably other pathogens as well—the severity of the secondary cases might depend on that of the index case. This possibility has been tested among children in the Copenhagen infectious diseases hospital, where both the index and secondary cases were hospitalized and their clinical severity assessed. Secondary cases infected by a severe index case (pneumonia or death) had significantly more pneumonia, and a significantly higher CFR, than those infected by milder index cases, and their relative risk of dying was 4.6 times greater (95% CI: 2.1 - 10.1, Aaby 1991a).

A similar observation was made in Kenya where secondary cases infected by an index case who died experienced significantly higher mortality than those where the index case survived (RR=4.7; 95% CI:1.6 - 13.4; Aaby and Leeuwenburg 1990). Although poor treatment of both the index case and the secondary case may contribute to this difference, the data do suggest that severe index cases generate secondary cases of greater severity. Depending on the pattern of transmission of measles in a community either positive or negative feedback processes may occur. Where most index cases are mild the cases to which they give rise will also be relatively mild, and the net result will be a lowered mortality. However, where the index cases themselves are severe they will transmit an even more severe infection to the new cases, further aggravating the overall level of mortality.

It is probably this mechanism which underlies the increasing severity sometimes observed in institutional outbreaks. For example, in a outbreak among soldiers of the Highland Regiment at Bedford in 1914–1915, the CFR was only two per cent (n=87) in the first month but rose to 14 per cent (n=442) in the two subsequent months (RR=6.2, 1.6 - 24.9, $p=0.002$; Kinnear 1923). Among children in a German refugee camp in 1915, during the First World War, the CFR increased from ten per cent (n=50) in the first two months of the epidemic, to 45 per cent (n=628) during the remaining four months (RR=4.6, 2.0-10.5, $p<0.0001$; Reder 1918).

A special instance of this ‘amplification’ phenomenon was found in our Senegal study where in the very large compounds outbreaks could take the form of a series of waves of infection, the CFR increasing exponentially with each new generation of cases (Table 6) (Garenne and Aaby 1990). This pattern of amplification presumably reflects increases in the infective dose, since severe cases apparently excrete more measles virus than do the milder ones. However, it is possible that severe cases are also more likely to transmit complications, and this too may contribute to the phenomenon.

Table 6
Measles case fatality ratio by age and generation of cases (wave), Niakhar, 1983–1986

Age (months)	Index cases %		Secondary cases according to generation of cases							
			2nd wave		3rd wave		4th wave		5th wave	
0–5	0	(0/4)	0	(0/13)	22	(2/9)	0	(0/1)		
6–41	7	(8/115)	12	(36/290)	12	(11/90)	29	(7/24)	45	(5/11)
42–65	0	(0/79)	4	(5/142)	5	(2/38)	0	(0/9)	50	(2/4)
66+	0	(0/122)	1	(1/177)	1	(1/81)	0	(0/15)	0	(0/4)
Total	3	(8/320)	7	(42/622)	7	(16/218)	14	(7/49)	37	(7/19)
Odds ratio	1.0		2.3		3.6		5.5		16.1	

Cross-sex transmission

As an unexpected extension of the transmission perspective, we found in several studies in Guinea-Bissau that secondary cases infected by someone of the opposite sex had a higher CFR than those infected by someone of the same sex (Table 7) (Aaby et al. 1986b). This tendency did not depend on the classification of who infected whom, for it could be shown that mortality for children aged 6-59 months was higher in houses where one boy and one girl had measles together (26%) than in those with two boys or two girls (11%) (RR=2.65; 95% CI: 1.20 - 5.84). Something similar seems to have occurred in Copenhagen at the beginning of this century (Aaby 1991b). Mortality was significantly higher ($p < 0.05$, Mantel-Haenszel, $\chi^2 = 4.36$) in families where one boy and one girl caught measles during the same outbreak than in those where two boys or two girls fell ill (RR=1.85; 95% CI: 1.04-3.30).

Table 7

Case fatality ratio among secondary cases of measles by age and by sex of infecting child, Guinea-Bissau 1979-84

Age (Months)	Case Fatality Ratio %							
	(deaths/no.ill)							
	<u>Same-sex transmission</u>		<u>Cross-sex transmission</u>		<u>Same-sex transmission</u>		<u>Cross-sex transmission</u>	
	Male to Male	Female to Female	Male to Female	Female to Male	Male to Male	Female to Female	Male to Female	Female to Male
6-35	26 (9/35)	16 (5/31)	49 (22/45)	36 (16/45)	26 (9/35)	16 (5/31)	49 (22/45)	36 (16/45)
36-59	10 (1/10)	6 (1/16)	20 (3/15)	18 (3/17)	10 (1/10)	6 (1/16)	20 (3/15)	18 (3/17)
Total	22 (10/45)	13 (6/47)	42 (25/60)	31 (19/62)	22 (10/45)	13 (6/47)	42 (25/60)	31 (19/62)

Source: Aaby et al. (1986b)

This tendency has been examined in a number of studies. In a small outbreak in Keneba, in The Gambia, where the children were under close medical observation, and there was no mortality, individuals infected by someone of the opposite sex were more likely to suffer pulmonary complications than those infected by their own sex (RR=2.82, 95% CI: 0.9-9.7; Aaby and Lamb 1991). Since this result could be influenced by the severity of the index case, the effect of sex differences on the increase in severity from index to secondary case was also examined. Secondary cases infected by the opposite sex were significantly more likely to show increased severity, relative to the index case, than were those infected by someone of their own sex ($p = 0.026$).

In Niokholonko, a rural area of Senegal under demographic surveillance since 1970, 243 deaths were registered during three separate outbreaks of measles. Unfortunately the total number of cases was not recorded, and so it was not possible to document the immediate source of infection where a death had occurred. However, on the assumption that this was most likely to be a maternal sibling, we examined the mortality risk, by sex composition of sibships, in families with only two maternal siblings under ten years of age. When significant background factors were taken into consideration, the relative risk of dying of measles during an outbreak was 1.81 times greater (95% CI: 1.17-2.82) in sibships consisting of one boy and one girl than in those containing two children of the same sex (Pison et al. 1992).

A similar pattern has been detected in several other studies in Kenya (Aaby and Leeuwenburg 1991), Senegal, Greenland and Germany (author's unpublished observations). Published case reports of

measles fatalities, giving the sex of both index and secondary cases, also reveal higher mortality from cross-sex infection (Aaby et al. 1986b). There are indications that this pattern may also occur where other infections are concerned since we have found that twin pairs of unlike sex suffer higher post-neonatal mortality than do same-sex twins (Aaby and Mølbaek 1990).

There is no simple explanation for the higher mortality associated with cross-sex infection. It is not obvious that children of opposite sexes have more intensive contact with each other. In fact, the available studies indicate that the reverse is the case (Aaby et al. 1986b). It thus seems probable that a biological mechanism is involved. Since viruses take material from the cells they have grown in, it may be that those grown in male or female cells have a different potential for infecting cells of the opposite sex (Aaby et al. 1986b).

Delayed impact of measles and intensive exposure

The importance of the transmission approach may transcend the immediate outcome of the infection. Most studies have only dealt with acute measles mortality (within one month of the rash), but some recent work has suggested that measles may have a profound effect on morbidity and mortality beyond this period (Aaby, Clements and Cohen 1987). Though there are few studies of the determinants of such 'delayed impact', it is likely that it arises from intensive exposure during the period of acute infection. For example, among the children hospitalized in Copenhagen, secondary cases had a higher risk of dying after the first 30 days of measles infection (3/152) than did index cases (3/472) (RR=3.11; 95% CI: 0.7-14.0). In rural Senegal, the secondary cases had a significantly higher risk of dying in the year following measles infection than did the index cases (RR=3.5) (Table 8) (Garenne and Aaby 1990).

Table 8

Post-measles mortality rate (6–52 weeks) by age and intensity of exposure, Niakhar 1983–1986

Age (months)	Case fatality ratio % (deaths/no. of survivors)	
	Index cases in compound	Secondary cases in compound
0–5	0 (0/4)	5 (1/210)
6–41	2 (2/107)	7 (25/356)
42–65	1 (1/79)	2 (3/184)
66+	0 (0/122)	1 (2/275)
Total	1 (3/312)	4 (31/836)
Odds ratio	1.0	3.5

Source: Garenne and Aaby (1990)

In Guinea-Bissau, we also found that children who had been exposed to measles at home during the first six months of their lives had a mortality between ages six months and five years which was three times higher than community controls (34% vs 11%) (Aaby et al. 1990a). In a Cox regression analysis, taking the known background factors into consideration, the mortality hazards ratio was 5.7 times higher (95% CI: 2.7 - 12.0) among the exposed children than the controls. The difference in mortality was significant in each of the age intervals 6–11, 12–23 and 24–35 months. This delayed excess mortality was found both among children who had had measles (despite being below six months of age) and among those who had been exposed without showing clinical symptoms. The latter were particularly likely to die of diarrhoea.

These tendencies have been revealed systematically in several outbreaks in Guinea-Bissau (Aaby et al. 1990a) but have not been examined elsewhere. The observation of delayed mortality among children exposed to measles early in life—without showing clinical symptoms—suggests that measles may be responsible, indirectly, for a much larger share of childhood mortality than is usually assumed. The mechanisms underlying this longer-term relationship require further study, but some form of persistent infection and immunosuppression may be involved. Its effect may go even further. In two outbreaks in Guinea-Bissau, it was found that children of mothers exposed to measles during pregnancy had a significantly higher perinatal mortality (Table 9) (Aaby et al. 1988b) as well as an increased post-perinatal mortality (Aaby et al. 1990b).

Table 9

Perinatal mortality among children of mothers exposed to measles during pregnancy, Bandim, Guinea-Bissau, 1979

Type of mortality	Perinatal mortality risk % (deaths/no. at risk)		
	Exposed	Controls	O.R. (95% C.I.) ^a
Stillbirths	6.5 (7/107)	1.4 (5/346)	4.8 (1.7–13.8)
Died in first week	9.0 (9/100)	2.6 (9/341)	3.6 (2.3–5.6)
Perinatal	15.0 (16/107)	4.0 (14/346)	4.2 (2.1–8.5)

^aOdds ratio (95% confidence interval)

While these tendencies have not been studied elsewhere, there are indications that early exposure may have consequences for health in later life. The delayed fatal form of measles known as subacute sclerosing panencephalitis (SSPE) is most frequent among children who had measles early in life, suggesting that the victims are more likely to have been intensively exposed as secondary cases (Aaby et al. 1984d). A study from Denmark found that the frequency of cancers and immunoreactive diseases among adults with no history of childhood measles was four times higher than that of controls with such a history (Rønne 1985). Since all the controls had measles antibodies, it is most likely that they had been exposed to the virus at a very early age when still partly protected by maternal antibodies or immunoglobulin. It thus seems that early intensive exposure is connected with excess morbidity in later life.

Testing the contrasting hypotheses: the impact of measles vaccination

In seeking to explain any given problem the hypothesis we adopt must be able to account for all observed variations in our data and to resolve the contradictions inherent in competing interpretations. Measured against these criteria there is little to indicate that host factors such as malnutrition and young age at infection play a major role in determining the severity of measles infection. Host-factor hypotheses cannot explain major contrasts in measles mortality, and some of their predictions are contradicted by the known epidemiology of severe measles.

On the other hand, transmission factors appear to explain much more of the variation in mortality; differences in mortality between historical periods, according to the degree of crowding, between regions, and between rural and urban areas, all correspond to differences in the intensity and type of exposure. Patterns which conflict with the host factor hypotheses, for example the differences between

Africa and Asia, and between rural and urban areas, are compatible with the transmission factor approach.

Table 10
Efficacy of measles vaccine against death: different studies

Country	Age at vaccination	Period of follow-up	Mortality %	Vaccinated	Vaccine efficacy against death (95% C.I.)
Nigeria (Hartfield and Morley 1963) ^c		18 months	12 ^b (25)	0 (23)	100
Zaire (Kasongo Project Team 1981) ^d	7–9 months	30 months	7–9.5 ^a	3.8	46–60
Guinea-Bissau (Aaby et al. 1984c) ^e	6–35 months	12 months	14.3 ^a (70)	1.9 (361)	87 (70–94)
Guinea-Bissau (Aaby et al. 1989) ^f	7–24 months	24 months	13.2 ^b (53)	4.8 (124)	63 (2–86)
Guinea-Bissau (Aaby, Knudsen et al. 1990) ^g	9–23 months	24 months			66 (32–83)
Senegal (Garenne and Cantrelle 1986) ^h	6–35 months	30 months	33.6 ^a	23.2	31
Haiti (Holt et al. 1990) ⁱ	6–13 months	9–39 months	6.6 ^a (1056)	1.3 (235)	85 (36–96)
Bangladesh (Clemens et al. 1988) ^j	9 months	9–60 months		(8135)	36 (21–48)
Bangladesh (Koenig et al. 1990) ^k		9–60 months	(8135) ^a		46 (35–54)

Note: ^a Unvaccinated

^b Placebo

^c A small placebo study carried out at the beginning of the 1960s.

^d The study in Zaire was carried out in the mid-1970s. Vaccination was introduced in one area and mortality compared with a comparable neighbouring area and with data from the same two areas prior to vaccination being introduced.

^e The study was carried out in 1980 in the capital of Guinea-Bissau. Mortality has been compared for children who attended a child examination and were vaccinated against measles and children who did not attend, mostly because they were temporarily absent.

^f The study carried out in 1984–86 represents a ‘natural experiment’. When blood samples taken in connection with vaccination were analysed two years later, it was found that, during a short period, none of the children had seroconverted. They can thus be considered to have received a ‘placebo’.

^g The study carried out between 1984 and 1987 compared mortality for children who received vaccination and those who did not in two districts in the capital of Guinea-Bissau.

^h Two systematic measles vaccination campaigns were carried out in a rural area of Senegal, in 1965 and 1967. Mortality was compared with an area where vaccination was not available.

ⁱ The study compares mortality for children who had participated in a serological study in 1982. The estimation of vaccine efficiency against death takes account of background factors with a significant effect on mortality (socioeconomic status, literacy, knowledge and use of oral rehydration therapy, and birth intervals).

^j The study from the Matlab area in 1982–84 compares vaccination status for 536 children who died with 1072 controls.

^k This study of 8135 vaccinated children and controls from the Matlab area covers 1982–85. The study overlaps partly with the previous study.

In health-related matters, the ultimate test of a hypothesis lies in its implications for prevention, and our two competing hypotheses imply very different assessments of the value of specific disease interventions (vaccination, immunoprophylaxis). If measles is primarily a killer of weak and malnourished children, then immunization might have only a limited effect on survival because those whose lives are thereby saved would have a high risk of dying from other diseases (Hendrickse 1975; Kasongo Project Team 1981; Mosley 1985). However, if intensity of exposure is the major determinant of measles' severity then there is no reason why weaker children should be particularly at risk (Aaby et al. 1984c), and in this case immunization should have a significant impact on overall survival.

Community studies available from Nigeria, Guinea-Bissau, Senegal, Zaire, Bangladesh and Haiti all indicate at least a 30 per cent reduction in child mortality after the age of vaccination, and seven of the studies found a reduction of at least 45–50 per cent (see Table 10). None have failed to find a marked effect on mortality, although it is true that most suffer from potential methodological problems because they are not randomized double-blind trials. However, one study found that mortality among children who seroconverted was only a third of that of children failing to seroconvert because they had received an ineffective vaccine as a placebo (Aaby et al. 1989).

In all studies the reduction in mortality was greater than expected from the direct contribution of measles deaths to overall child mortality. For example, in Bangladesh, the reduction in mortality between ten and 60 months of age was 36 per cent although measles accounted only for four per cent of all deaths among the controls (Clemens et al. 1988). Thus measles immunization seems to be highly effective in preventing both acute and delayed mortality from the disease.

Lessons for the past

Although no other infection has been studied systematically from the transmission-factor perspective, there are indications that the mechanisms emphasized here may apply more widely (Aaby et al. 1985), and I shall argue, using four examples, that the transmission perspective has a general importance for the understanding of historical data on mortality and morbidity.

Measles mortality decline in the industrialized world

The contrasting hypotheses outlined above provide quite different interpretations of the process of measles mortality decline. In England, for which the best data and analyses are available thanks to Reves (1985), a very important fall in measles mortality took place between 1900 and 1930 (Table 11) which was attributed to improvements in nutrition (Mayer 1969). The main evidence for this attribution is a correlation between low social status, as measured by the number of rooms occupied by the family, and measles CFR (Morley 1973). However, this correlation does not necessarily indicate that poor children had a higher case fatality rate because they were malnourished. There are no specific data linking nutritional changes with improved survival rates among those infected, and it could equally well be that children living in small apartments had a higher risk of intensive exposure to the infection.

It has also been hypothesized that the decline in the industrialized world's measles mortality at the beginning of this century reflected a rise in the age at infection (Reves 1985). English fertility fell dramatically—from 6.7 children per woman in 1875 to 2.6 in 1925—and it has been assumed that this increased the average age at infection because children in smaller families were older when they first encountered the virus (*The Lancet* 1985; Reves 1985). However, this effect may have been offset by a contemporary increase in the risk of early infection outside the family due to greater urbanization, improvements in transport, and the spread of schooling and public child care.

Table 11
Age-specific measles mortality rate (per million), England and Wales 1911–1930

Age (years)	Period			
	1911–15	1916–20	1921–25	1926–30
<1	3,130	1,931	1,337	1,284
1–4	2,905	1,861	1,182	999
5–9	205	190	95.9	95.9
10–14	17.1	12.7	6.5	10.3
15–19	9.4	12.8	2.4	3.6
20–24	7.1	10.0	1.7	1.6

Source: Reves 1985

There are no data for England as a whole to show any increase in the age at infection in this period (Reves 1985). The few available community studies and reports of notifications suggest no major change, and the proportion of cases under three years of age has been remarkably constant (Aaby et al. 1988c). At a time when measles in the United Kingdom was severe, Picken (1921) noted that variations in mortality could not be explained by differences in age distribution, and it is known that the mean age at infection in England and Wales fell from 5.5 to 4.4 years between 1944 and 1968 (Anderson and May 1982).

Thus it remains true, as Lancaster (1952) noted, that there is no adequate explanation for the decline in CFR in the industrialized world before the advent of antibiotics. However, the decline may be better understood—from a transmission perspective—in terms of reductions in the frequency of intensive exposure and a feedback process. Data from England and Germany, together with hospital studies from Copenhagen (Aaby 1988a), dating from the period of severe measles mortality show that this was higher for secondary cases than for index cases whilst community studies from Sunderland in 1885 (Aaby et al. 1986a) and rural Germany in 1861 (Pfeilsticker 1863) suggest that the proportion of secondary cases in Europe was as high as that found in present-day Africa.

Although there are no proper epidemiological studies of measles transmission in contemporary industrialized countries, it seems likely that smaller families and changes in schooling and public child care have significantly reduced the proportion of secondary cases since the beginning of this century. Factors such as increased urbanization, universal primary education and the spread of public child care will have raised the likelihood of early infection by taking the child out of the family and increasing its contact with other children. The earlier immunity thus acquired will also have reduced the risk of children becoming infected at home as secondary cases should school-age siblings bring the virus into the home.

Thus the social processes stemming from industrialization contributed to breaking the older pattern of transmission in which children experienced their most numerous contacts with susceptible peers on entering school, and school-aged children brought the virus home intensively infecting their susceptible younger siblings. The importance of this pattern seems to have been recognized by Brownlee (1920), the leading contemporary epidemiologist, who proposed that schools maintain registers of all their pupils' siblings. Thus, when measles broke out in the school, teachers would be able to advise parents to board out any small siblings with relatives until the epidemic had passed. This proposal constitutes a dramatic recognition of the importance of preventing intensive exposure at home. I do not know if it

was ever implemented but it seems likely that social processes were more effective in this respect than any school-based public-health measures over the first decades of this century.

It may also have been significant that doctors in the 1920s and 1930s began using convalescence serum (and later gamma-globulin) to prevent, or modify, measles infection. This treatment would have been aimed precisely at the high-risk group for measles mortality, namely small children exposed to a sibling at home, and its impact has not been properly assessed. Gamma-globulin has had a marked effect on the CFR in virgin-soil epidemics among indigenous peoples who all contract measles at the same time (Aaby et al. 1987).

Undoubtedly the most important contribution was made by reduced family size, leading to fewer secondary cases and thus to lower mortality, but this cannot explain the entire reduction in mortality. Secondary cases still occur in the developed countries, but their mortality is virtually nil. However, as the proportion of secondary cases falls, the index cases will themselves become milder, having been infected by mild index cases rather than severe secondary cases. Thus, secondary cases too may have become milder, and the general severity of measles in the community been gradually reduced, through a process of positive feedback.

In order to examine this question further we have used records from the infectious diseases hospital in Copenhagen. Table 12 shows that children infected at day care institutions had a significantly lower mortality than did index cases contracting measles from another child in the same apartment building. Presumably exposure was less intense in the kindergarten because children sick with measles were kept at home, whereas children at home may have maintained contact with infected playmates. Since children infected in kindergartens also had much lower mortality, the establishment of such institutions may have contributed to the fall in measles mortality by reducing the proportion of secondary cases.

Table 12
Measles case fatality ratio according to source of infection and age, Copenhagen 1915–1925

Age (years)	Day care %	Neighbours, playmates %	Secondary cases %
0	13 (5/40)	40 (4/10)	29 (16/55)
1	5 (3/64)	22 (5/23)	29 (24/84)
2	6 (2/31)	22 (4/19)	20 (9/44)
0–2	7 (10/135)	25 (13/52)	27 (49/183)

Source: Aaby (1988a)

It may also have promoted the transmission of measles in a milder form. Table 13 shows marked annual variations in both measles mortality and the proportion of children infected in day-care institutions between 1915 and 1925. When this proportion was high the overall CFR tended to be low. This was only partly due to the lower mortality of children infected at kindergarten, because as the table shows, children under three, infected elsewhere, also had a reduced CFR in years when the proportion of kindergarten infections was high (logistic regression, $p=0.001$ $\chi^2=10.88$, 1 df). It is possible that in some years the epidemic was dominated by transmission among kindergarten children, whilst in others transmission at schools and apartment buildings was more important. These observations support the suggestion that an increase in mild cases in the community reduces the absolute number of severe cases. It may be a process of this kind which generated the considerable annual variation in CFR, a variation explained by contemporary epidemiologists in terms of changing pathogenic virulence.

Table 13

Case fatality ratio according to age, year and frequency of children infected in day-care institutions, Copenhagen, 1915–1925

Year	Case fatality rate (no. of cases)							
	All cases		Children <3 years		Children >3 years not in day care		Children <3 years infected in day care	
	%		%		%		%	
1915	10	(114)	16	(49)	20	(35)	29	(49)
1916	20	(295)	29	(123)	31	(116)	6	(123)
1917	10	(100)	21	(42)	20	(38)	10	(42)
1918	8	(153)	18	(66)	19	(58)	12	(68)
1919	14	(172)	23	(84)	25	(72)	14	(84)
1920	18	(244)	23	(125)	24	(115)	8	(125)
1921	8	(72)	10	(30)	11	(28)	7	(30)
1922	8	(14)	14	(69)	16	(57)	17	(69)
1923	16	(304)	29	(96)	30	(90)	6	(96)
1924	9	(275)	14	(108)	16	(83)	23	(108)
1925	10	(312)	12	(160)	14	(125)	22	(160)
Total	13	(2208)	20	(952)	22	(817)	14	(952)

There are few well-documented accounts of fundamental changes in CFR from developing countries. The three series we have, from Kenya, Gambia and Guinea-Bissau, all indicate that neither nutritional changes nor variations in age at infection were responsible, but that fundamental changes occurred in the transmission patterns and reduced the proportion of secondary cases (Aaby 1990). Furthermore, the extension of vaccination coverage proved to be essential in each study. Experience from Africa suggests that vaccinated children often have milder infections with lower CFR (Aaby et al. 1986c). It is therefore likely that they contribute to the transmission of milder infection in the community.

To sum up, it seems likely that the most important causes of measles mortality decline were social changes which diminished the risk of intensive exposure within the family. Chief among these were the fall in family size and greater social contact among young children which increased the risk of infection outside the home. Furthermore, the continual reduction in the numbers of fatal cases has reduced the risk of transmission of measles in a severe form.

The strength of the weaker sex: the social production of ‘biology’

‘Men are more often colour blind than women,’ said Joanna. ‘It’s one of those sex-linked things,’ she added, with an air of erudition. ‘You know, it passes through the female and comes out in the male.’ ‘You make it sound as though it was measles,’ said Emlyn Price. (Agatha Christie, *Nemesis* 1974).

Western popular beliefs have been characterized by a complementary view of the sexes; men are seen as being endowed with greater physical strength, while women are more resistant biologically. More boys are born, but in the end it is the women who survive (Mims 1976). Scientific studies of infectious diseases have supported this belief by showing many infections, including measles (see Table 14), to be more severe in males (Karzon et al. 1961; McGlashan 1969; Denny et al. 1977). Such findings are usually explained in terms of ‘sex determined immune capacity’ and ‘hormonal differences influencing immunity’ (Mims 1976).

The pattern observed in the Western world has come to be seen as the 'biological' or 'natural' one, and it causes little surprise if boys have higher mortality. On the other hand, where girls have higher mortality we are in need of an explanation. Usually the reference is to differential treatment (Bhuiya et al. 1987). However, for a biological constant there are some surprising irregularities in the pattern even within the Western world. For example, as also illustrated in Table 14, beyond a certain age females experience higher risks of measles mortality than do males. The reason that fewer girls die overall is that they have lower mortality in the youngest years, where the CFR is particularly high, and not because they have an advantage at all ages. This pattern of early female advantage, being reversed at older ages, has also been found in other infections such as polio (Hall, Nathanson and Langmuir 1957).

Table 14
Measles case fatality ratio and relative risk for females by sex and age, registered cases 1883–1902, Aberdeen.

Age (years)	Case fatality ratio %				Relative risk for females %
	Cases	Males	Females	Total	
0	3034	14.5	13.6	14.0	0.94
1	5222	10.9	9.2	10.1	0.85
2	5195	3.6	3.2	3.4	0.89
3	5053	1.5	1.7	1.5	1.13
4	4836	0.9	0.9	0.9	0.92
5	5352	0.7	0.7	0.7	1.01
6	4628	0.4	0.5	0.5	1.42
7	2818	0.3	0.7	0.5	2.27
8–14	3155	0.3	0.4	0.4	1.27
15+	1081	0.6	1.0	0.8	1.76

Source: Wilson (1905)

The lower mortality among females may be due less to sex-determined immunological strength than to the fact that infection 'passes through the female and comes out in the male', as pointed out by Agatha Christie. Since severity is associated with intensive exposure and cross-transmission of infection, sex differences in these respects might explain the sex difference in mortality. In fact these two mechanisms would tend to work together. If members of one sex are more likely to become secondary cases, they will also be more likely to be infected by a member of the other sex, since it is these who will comprise the bulk of index cases. Thus small inter-sex variations in the risk of becoming a secondary case may result in much larger variations in severity and CFR.

In European culture there are several indications that girls catch measles and other infections at a younger age than males (Aaby et al. 1984d). One dramatic illustration of this is the age-distribution of viral infections in Denmark, where females outnumber males by more than two to one in the first year of life (Table 15). Since the reporting rate has been high (50%), and the total number of males and females is the same, biased reporting is unlikely to be responsible for this pattern. Since there is no indication that girls are more likely to be infected if exposed at home, it seems likely that they have a higher risk of infection elsewhere. This may be due in part to differences in behaviour, as girls appear to have closer face-to-face contact with their playmates as early as kindergarten age, and this could facilitate the transmission of infection. Furthermore, we have found in Guinea-Bissau that girls tend to have fewer maternal antibodies against measles which would render them susceptible at any earlier age than boys (author's unpublished observations). Both of these tendencies would lead girls to contract

infection at a younger age than boys, predisposing the latter to a higher risk of domestic infection as a secondary case which may be one reason for the higher mortality of boys in the industrialized countries (Aaby et al. 1983).

Table 15
Reported cases of measles by sex and age, Denmark 1963–1969

Age (years)	Males %	Females %
0	3.09	6.35
1–6	86.05	95.09
7–14	48.61	39.75
15+	3.57	2.80
Total	141.32	143.99

Source: Hertz, Sørensen and Vejerslev (1977)

Higher measles mortality for girls has been reported from several communities, usually with a Muslim culture (Monastiri 1961; McGregor 1964; Bhuiya et al. 1987; Fargues and Nassour 1988). It seems likely that girls are more bound to the home in these communities and that boys have a greater chance of getting out and thus of being infected outside the home. As a consequence, girls suffer a dual disadvantage: a higher risk of becoming a secondary case as well as a higher risk of being infected by someone of the opposite sex. In a study in rural Senegal, girls were found to have 30 per cent higher CFR than boys, and virtually all of this difference could be explained with reference to the transmission pattern. There were slightly more secondary cases among the girls, and they were significantly more likely to be infected by the opposite sex. When these differences were taken into consideration, there was only a five per cent higher CFR among the girls (author's unpublished observations).

Thus, before assuming that sex differences are based on biology, or differential treatment, it may well pay to examine the mode of transmission in that particular environment. There are very few studies where the sex of both the infected and the infecting child has been documented, but they have shown that girls were more likely to be the index case, or that boys were more likely to be infected by a member of the opposite sex (Pfeilsticker 1863; Asano et al. 1977; Aaby and Lamb 1991). The higher risk among older girls could be related to a lower degree of schooling, or to the way in which their role in child care exposes them more intensively to infection from siblings.

Virgin-soil epidemics

It seems that the Mandans were surrounded by several war parties of their more powerful enemies the Sioux, at that unlucky time, and they could not therefore disperse upon the plains... they were necessarily enclosed within the piquets of their village, where the disease [smallpox] in a few days became so very malignant that death ensued in a few hours after its attacks; and so slight were their hopes when they were attacked, that nearly half of them destroyed themselves with their knives, with their guns, and by dashing their brains out by leaping head-foremost from a thirty foot ledge of rocks in front of their village. The first symptoms of the disease was a rapid swelling of the body, and so very virulent had it become, that very many died in two or three hours after their attack, and that in many cases without the appearance of the disease upon the skin (Catlin 1841).

Five hundred years ago, the Amerindian population may have numbered as many as 100 million, or one fifth of the world's population (Crosby 1976; McNeill 1976). Where there are reliable data, they imply that the indigenous groups diminished by 90–95 per cent, or vanished completely, during the first 100–150 years of contact. Virgin-soil epidemics of diseases like smallpox, measles, influenza, whooping cough and typhoid fever—occurring where no one had immunity and thus everyone succumbed at once—were the major cause of this reduction and the prime means by which Europe acquired the Americas. They were also important on the islands of the Pacific basin.

There is no indication that the Indians were malnourished (Black et al. 1971), and, as in Catlin's account of the Mandan, the epidemics' dramatic impact has usually been ascribed to the Indians' lack of immunity (Black et al. 1971), or to the breakdown of nursing when everybody became sick simultaneously (Crosby 1976). But several features of these epidemics are inconsistent with such hypotheses, and the possibility of an unusually severe form of infection has been little considered.

The excessive epidemic mortality has fuelled speculation concerning the Indians' lack of immunity, or 'constitutional weakness' on their part. In support of such views, it was claimed in some older literature that mestizos often fared better than pure Indians (Smellie 1846; Bridges 1950). In modern terms, the suggestion is that Amerindians' susceptibility was increased because they 'lack certain genetic traits which have become common in other populations through prolonged selective pressure' (Black et al. 1971). Although the specific traits making non-Indians more resistant have not been demonstrated, it has been emphasized that in most studies of measles vaccination, Indians and Innuits displayed a stronger febrile response (0.4–0.5° higher fever) than did populations with prior measles experience (Black et al. 1971).

A further argument in favour of increased genetic susceptibility is provided by the observation that death in the prodromal phase (before the appearance of the rash) seems to be a distinctive feature of virgin-soil epidemics, as is also implied by Catlin's account. Normally measles deaths are thought to arise from complications due to secondary infections, most of which occur after the appearance of the rash. In the first measles epidemic on Greenland in 1951, 32 per cent of all deaths (25/77) occurred in the prodromal phase (Christensen et al. 1953).

The genetic hypothesis is difficult to test because the relevant traits and immunological mechanisms remain unknown, but such a hypothesis may not be necessary. In many epidemics among Europeans and Africans with long experience of measles the CFR has been as high as those documented for virgin-soil outbreaks among Indians and Innuits (Table 16). The epidemics have shared other features with the latter, occurring in institutions or closed communities where transmission was rapid and incidence universal. For example, during the Paris Commune the Garde Mobile had a measles CFR of 40 per cent (86/215) (Squire 1877), and outbreaks in institutions and refugee camps have often had CFRs of 25–50 per cent (Gannelon 1892; Godfrey 1928).

Table 16
Measles mortality in virgin-soil epidemics among Indians and Innuits

Locality	Year	Medical care	Gamma-globulin	Cases	Case fatality rate %
Fort York, Canada (Smellie 1846)	1846	Yes	No	145	27.6
Tierra Del Fuego (Bridges 1950)	1884		No	1000	60.0
S.W. Greenland (Christensen et al. 1953)	1951	Yes	Yes	4247	1.8

Ungava Bay	1952	Yes	No	900	7.4
Baffin Island	1952	Yes	Yes	900	2.0
(Peart and Nagler 1954)					
Xingu, Brazil	1954	Yes/No	No	654	17.4
Xingu, Brazil		Yes/No			9.6
Xingu, Brazil		No	No		26.8
(Nutels 1968)					
Yanomamo, Brazil	1968	Yes	No	480	8.8
(Neel et al. 1970)					

Prodromal deaths are also found elsewhere, but are much less likely to be diagnosed as measles in the absence of the rash. The higher fever response among vaccinated Amerindians, not previously exposed to the disease, could well be due to the controls having a low level of antibodies which displayed a modified fever reaction to the vaccine. It has been a common feature of the few well described virgin-soil epidemics that the disease was usually relatively mild to begin with, becoming much more severe with the progress of time (Squire 1877). This does not suggest that genetic susceptibility was particularly important—and the same feature has characterized measles outbreaks in institutions (Picken 1921)—whilst subsequent epidemics in the same isolated communities were often much milder. This again fails to support the genetic hypothesis since too little time had elapsed for genetic selection of ‘certain traits’ to have taken place.

The most commonly accepted explanation for heightened severity is the social breakdown which can occur during virgin-soil epidemics (Morley 1973; Crosby 1976). In support of the view that lack of nursing was the primary contributor to surplus mortality, several studies have suggested that even minimal provisioning for sick people could reduce mortality (Peart and Nagler 1954; Neel et al. 1970). For example, Table 16 shows that the CFR was lower in those epidemics with some medical care.

It is quite likely that fear of imminent death, fatalism and neglect of sick relatives may have contributed to the high CFRs observed in virgin-soil epidemics. However, the CFR has been equally high in situations where complete social breakdown did not occur. For example, in the Scottish Highland Division at Bedford in 1914–1915, 12 per cent (65/529) of the adult males stricken with measles died. In child institutions in New York, mortality was as high as 24 per cent in the period 1916–1918 (Godfrey 1928). While these institutions may well have been under stress due to the simultaneous occurrence of many cases of measles, there is no indication that basic necessities like food and water were not provided.

Furthermore, there are a number of features of virgin-soil epidemics which are difficult to relate to social breakdown. Lack of nursing and medical care can probably not produce prodromal deaths, and several accounts of virgin-soil epidemics have stressed that the disease was particularly severe for adult females (Corney 1913; Christensen et al. 1953). Though this could reflect womens’ greater risk of being inadequately nursed while sick, it seems also to hold in Greenland where care was sufficient to secure a relatively low level of overall mortality (see Table 16). Adult women aged 15–54 had significantly worse mortality, and a higher rate of complications, than did their male age-mates (Christensen et al. 1953).

All of these features, which fail to fit the genetic or social breakdown hypotheses, appear to be compatible with the transmission factors which we emphasized hitherto:

(a) In a virgin-soil epidemic, where there is no prior immunity, the proportion of individuals infected within the home will be much higher. The Mandans, for example, had large huts with as many as 40 inhabitants. In a reanalysis of data from the first measles epidemic on Greenland, secondary cases displayed a higher CFR than index cases (author’s unpublished data).

(b) A progressive increase in mortality, from an initially low level, would be predicted from the mechanism of amplification of severity. As the proportion of severe secondary cases rises, subsequent cases will become even more severe.

(c) The higher mortality of adult females is hard to reconcile with the notion that they have a greater immunological capacity. Pregnancy, which is immuno-suppressive and connected with more severe infections, might explain some of the difference, but in the Greenland epidemic non-pregnant women also had a higher CFR than their male age-mates (Christensen et al. 1953). It seems simpler to explain this in terms of women's role at home, taking care of domestic tasks and sick family members, which is likely to have increased their risk of intensive exposure.

(d) Prodromal deaths and a shorter period of incubation have been noted in several virgin-soil epidemics. These phenomena are compatible with lack of genetic resistance to the infection, but also with the higher dose of infection which is likely to result from more intensive exposure and is known to reduce the period of incubation.

(e) The impact of medical care and gamma-globulin. Table 16 suggested that medical care could reduce mortality, but this seems to have been the case particularly where gamma-globulin was used prophylactically. When it is given just before exposure, or in the early phase of incubation, some of the invading virus will be neutralized, the period of incubation prolonged, and the severity reduced. For example, when measles attacked the Innuits and Indians around Hudson Strait in Canada, the CFR was as high as 7.4 per cent (67/900) of the total population on the southern side (Ungava Bay). Here, no one received gamma-globulin, while other kinds of treatment and care were provided by the Indian health service. When measles reached the other side of the strait (Baffin Island) medical supplies were dropped by parachute. These included gamma-globulin, and there were no deaths among the 269 persons who received it ($p < 0.001$).

The key factor in these cases seems to have been a reduction in infective dose, rather than either basic nursing care or genetic susceptibility. By implication, excessive doses of infection are likely to have been a major factor in producing the excess mortality that has characterized virgin-soil epidemics.

European mortality decline: the role of smallpox vaccination

The children whose death might be prevented by measles vaccine are at risk of dying not because of the severity of measles *per se*, but because they are on 'the road to death' and their nutritional status is so poor that they are likely to die of any infectious disease. Thus, preventing a death among these children may not necessarily save a life, but only change the cause of death (Mosley 1985).

McKeown's explanation of the decline of mortality throughout Europe since the eighteenth century has been widely accepted, even in medical circles, despite his minimization of medicine's contribution before the advent of antibiotics in the 1930s (McKeown and Brown 1955; McKeown and Record 1962; McKeown et al. 1972; McKeown, Record and Turner 1975). McKeown argued that, theoretically, the reduction in mortality from the late eighteenth century and throughout the nineteenth century could be due to one or more of the following factors: medical measures for preventing and treating disease in the individual; a spontaneous decline of mortality due either to decreased virulence or increased resistance; improvements in environmental conditions.

He discarded the contribution of medical measures because the nature of infectious disease was not understood until 1850, and specific prevention or treatment was unlikely to have had much influence on the national death rate before the introduction of chemotherapy. Spontaneous mortality decline was also ruled out as being unlikely to have contributed much to the decline. Thus McKeown concluded that it was changes in environment, particularly improvements in diet and the results of sanitary

reforms, which were responsible. To a large extent this conclusion was reached negatively, by excluding all the (perceived) alternatives. The data in favour of improved diet as the main cause of reduced mortality are not very convincing, and as has been pointed out by others (Razzell 1974), it is actually questionable to what extent diet did improve over the period. Furthermore, there are no data to indicate that improvements of the kind that may have occurred would have affected the causes of death most implicated in the decline.

However, what is more critical is the apparent ease with which McKeown et al. (1972) dismissed the eighteenth-century role of inoculation against smallpox followed by that of vaccination after 1800. The former role was dismissed on the grounds that it required a much greater effectiveness for the technique than would be accepted by virologists, and that it attributed to this crude measure an influence on total mortality which would not be expected from any modern immunization procedure. The argument against vaccination was that control of smallpox contributed little to the decline of mortality between 1838 and 1900. However, what happened after Civil Registration in 1838 is a poor guide to events in the first decades of the century.

The available data suggest a very dramatic drop in mortality, related to vaccination, in the first decades after the practice became widespread, or even compulsory (Razzell 1974; Mercer 1985). The arguments against the effectiveness of inoculation seem to be erroneous in the light of data from England, Copenhagen and Boston which suggest that it was highly effective when used during epidemics. Furthermore, the time trend of mortality also suggests that inoculation contributed to a decline in smallpox deaths (Mercer 1985). These data have been presented more fully by others and will not be discussed in detail here.

What is more interesting in the present context is McKeown et al.'s (1972) claim that the control of smallpox has been erroneously credited with 'an influence on total mortality which would not be expected from any modern immunization procedure'. This seems to reflect *a priori* beliefs, rather than specific investigations, and to be based on the assumption that no single infection can ever have had such an influence because its victims would always have been drawn from the weakest and most vulnerable sections of the population.

However, if experiences with measles and measles vaccination is any guide to what we can expect in relation to smallpox and smallpox vaccination, McKeown was certainly wrong. Measles immunization may help to reduce total mortality in several distinct ways:

- (a) Acute deaths are prevented, and there is no indication that those saved from measles are more likely to die from other causes afterwards.
- (b) Vaccinated children who subsequently contract measles have a much lower CFR, probably because the vaccine has induced some partial immunity.
- (c) Measles gives rise to some delayed mortality, and this mortality is also prevented by immunization.
- (d) Unvaccinated children who catch measles have a lower CFR than they would otherwise do because the degree of clustering and the risk of intensive exposure have been reduced; such cases also contribute to the transmission of milder measles thus lowering the CFR still more.
- (e) Early exposure to measles is connected with much higher mortality later on in childhood; immunization should reduce the risk of this kind of exposure.
- (f) Exposure to measles during pregnancy is connected with much higher perinatal and post-perinatal mortality. Immunization should reduce this kind of exposure.

Of these possible mechanisms, only the first three relate to the effect of immunization on individual survival. All the available studies suggest that the effect of measles vaccination on overall mortality, a reduction of some 40–50 per cent after the age of vaccination, was much larger and could

be explained by the proportion of all deaths attributable directly to the disease. To these gains should be added those obtained by eliminating the long-term effects of early exposure and exposure during pregnancy. Experience from Guinea-Bissau suggests that these may produce excess deaths equal in total to 50 per cent of those from acute measles.

I am not aware of the extent to which measles and smallpox can be compared in terms of their determinants and effects, but they appear to be similar in at least some respects. Smallpox was as fatal to well-nourished individuals as to the poorly-nourished (Mercer 1985). Severe cases transmitted a more severe form of infection. Immunized individuals had a much lower CFR if they contracted the infection. Contemporary opinion connected smallpox to many later complications, for example consumption (Mercer 1985), and it may thus have functioned like measles. However, I am not aware of any study which has actually measured the size of the delayed mortality after acute smallpox, nor is there, to my knowledge, any study which has examined the impact of early exposure or exposure during pregnancy.

There seem to be rather consistent data from several European countries, at the turn of the eighteenth century, to show that smallpox deaths constituted 10–15 per cent of all deaths (Mercer 1985). This is likely to be an underestimate since many smallpox deaths may not have been properly registered as such. If those who died of smallpox did not also display higher risks of dying from other infections, then immunization should have reduced overall mortality by at least 10–15 per cent, and maybe by much more if smallpox was also associated with delayed mortality.

It should be worthwhile to examine whether smallpox gave rise to other than acute deaths. Given the level of mortality attributed to smallpox, and the parallel with the impact of measles control, the disease may well have been a prime mover in the European mortality decline from the eighteenth century. It is likely that McKeown's critiques (Razzell 1974; Mercer 1985) may have underestimated the impact of smallpox control. If there is anything to the contemporary notion of an association between smallpox and consumption (Mercer 1985), the much improved control of smallpox in the early nineteenth century could partly explain the major drop in tuberculosis mortality in its central decades.

Most epidemiological and historical research has searched for external factors to explain severe infection, assuming in Mosley's words that infection was not severe *per se*. Thus severe infection is seen as a reflection of other problems, which has consequences for our explanations of mortality change as well as for public-health policy. However, experience with measles suggests that, under certain conditions of exposure, some infections may indeed be a problem *per se* and may induce long-term consequences. Smallpox may belong to this category. If so, the main point may be, not to explain why smallpox was severe under certain conditions, but that the absence of smallpox may itself explain many other health phenomena, in particular the mortality decline which started 200 years ago.

Conclusion

There are undoubtedly many other aspects of past morbidity and mortality which may be fruitfully examined or reanalysed from the perspective of the transmission process. For example, Reves (1985) noted as a contradiction that diarrhoea mortality among children in England started falling, not at the time of the sanitary reforms, but at the same time as other forms of child mortality related to pneumonia, measles and whooping cough. Some of the important diarrhoea pathogens may also be influenced by the crowding together of many susceptibles, in which case they would have displayed the same reduction in severity as did measles and whooping cough following the fall in family size and the associated changes in transmission patterns.

However, diarrhoea death is probably often a symptom, rather than a result of some specific pathogen. For example, among children in Guinea-Bissau who were exposed to measles under six months of age, most of the excess deaths in later childhood were related to diarrhoea. Such children

had a risk of diarrhoea death which was eight times higher than that of controls. It is likely that diarrhoea death is related to previous measles infection, though this has not actually been demonstrated, and that intensive exposure to measles induces an immunodeficiency leading to elevated risk later in childhood. It is therefore not at all contradictory that diarrhoea mortality should decline at the same time as measles mortality.

Some other aspects of measles mortality are harder to understand. Contemporary accounts of measles in developing countries consistently report higher CFRs in rural than in urban areas, but British records usually indicated the reverse. This contrast could usefully be explored further to see whether it might be explained by differences in number of offspring, spacing and degree of crowding, or in the number and size of rooms, between urban and rural areas in Britain.

In the attempt to combine biological and social realities to obtain some kind of understanding of the variations in morbidity and mortality, discussion has been dominated by a rather limited set of host, disease and treatment factors believed to determine the severity of infectious diseases. The underlying assumption has usually been that there is something distinctive about those who die, that they are somehow the weak ones. I have argued here that a better insight might be obtained by asking whether the infection itself could vary in severity such that the essential question is not the distinction between children who die and those who survive, but that between a fatal infection and a less severe one.

In this paper I have mainly emphasized factors relating to the intensity of exposure, but synergistic relations between several infections occurring at the same time may also have had substantial effects on the severity of each of them. This is an aspect that I have not been able to study in any detail, but it seems worth pursuing in the future. It is likely that there was an important interaction between tuberculosis and measles in the past, such that measles paved the way for tuberculosis, and tuberculosis aggravated the severity of acute measles infection. The point has not been to suggest that host factors have no importance; they have. However, exploring the implications of the transmission-factor approach for historical data may provide many more lessons for both the past and the present.

References

- Aaby, P. 1988a. Severe measles in Copenhagen, 1915-1925. *Reviews of Infectious Diseases* 10:452-456.
- Aaby, P. 1988b. Malnutrition and overcrowding exposure in severe measles infection. A review of community studies. *Reviews of Infectious Diseases* 10:478-491.
- Aaby, P. 1989. Malnourished or overinfected? An analysis of the determinants of acute measles mortality. *Danish Medical Bulletin* 36:93-113.
- Aaby, P. 1990. Social and behavioural factors affecting transmission and severity of measles infection. Pp.826-842 in *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health*, ed. John C. Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid, Daphne Broers-Freeman. Canberra: Health Transition Centre, The Australian National University.
- Aaby, P. 1991a. Determinants of measles mortality: host or transmission factors? Pp. 83-116 in *Medical Virology* 10, ed. L.M. de la Maza and E.M. Peterson. New York: Plenum Press.
- Aaby, P. 1991b. Severity of measles and cross-sex transmission of infection in Copenhagen, 1915-1925. *International Journal of Epidemiology* 20:504-507.
- Aaby, P. 1992. Patterns of exposure and severity of measles infection, Copenhagen 1915-1925. *Annals of Epidemiology* 2:257-262.
- Aaby, P., J. Bukh, I.M. Lisse and A.J. Smits. 1983. Les hommes sont-ils plus faibles ou leurs soeurs parlent-elles trop? Essai sur la transmission des maladies infectieuses. *Anthropologie et Societ * 7,2:47-59.
- Aaby, P., J. Bukh, I.M. Lisse and A.J. Smits. 1984a. Overcrowding and intensive exposure as determinants of measles mortality. *American Journal of Epidemiology* 120:49-63.

- Aaby, P., J. Bukh, I.M. Lisse, A.J. Smits, J. Gomes, M.A. Fernandes, F. Indi and H. Soares. 1984b. Determinants of measles mortality in a rural area of Guinea-Bissau: crowding, age and malnutrition, *Journal of Tropical Paediatrics* 30:164-168.
- Aaby, P., J. Bukh, I.M. Lisse and A.J. Smits. 1984c. Measles vaccination and reduction in child mortality: a community study from Guinea-Bissau. *Journal of Infection* 8:13-21.
- Aaby, P., J. Bukh, I.M. Lisse and A.J. Smits. 1984d. Risk factors in subacute sclerosing panencephalitis (SSPE): age- and sex-dependent host reactions or intensive exposure? *Reviews of Infectious Diseases* 6:239-250.
- Aaby, P., J. Bukh, I.M. Lisse and A.J. Smits. 1986a. Severe measles in Sunderland, 1885: a European-African comparison of causes of severe infection. *International Journal of Epidemiology* 15:10-17.
- Aaby, P., J. Bukh, I.M. Lisse and A.J. Smits. 1986b. Cross-sex transmission of infection and increased mortality due to measles. *Reviews of Infectious Diseases* 8:138-143.
- Aaby, P., J. Bukh, J. Leerhoy, I.M. Lisse, C.H. Hordhorst and I.R. Pedersen. 1986c. Vaccinated children get milder measles infection: a community study from Guinea-Bissau. *Journal of Infectious Diseases* 154:858-63.
- Aaby, P., J. Bukh, G. Hoff, I.M. Lisse and A.J. Smits. 1987. Humoral immunity in measles infection: a critical factor? *Medical Hypotheses* 23:287-302.
- Aaby, P., J. Bukh, I.M. Lisse and C.M. da Silva. 1988a. Measles mortality decline: nutrition, age at infection or exposure? *British Medical Journal* 296:1225-1228.
- Aaby, P., J. Bukh, I.M. Lisse, E. Seim and M.C. da Silva. 1988b. Increased perinatal mortality among children of mothers exposed to measles during pregnancy. *Lancet* 1:517-552.
- Aaby, P., J. Bukh, I.M. Lisse and C.M. da Silva. 1988c. Measles mortality: further community studies on the role of overcrowding and intensive exposure. *Reviews of Infectious Diseases* 10:474-477.
- Aaby, P., J. Bukh, D. Kronborg, I.M. Lisse and M.C. da Silva. 1990a. Delayed excess mortality after exposure to measles during the first six months of life. *American Journal of Epidemiology* 132:211-219.
- Aaby, P., J. Bukh, I.M. Lisse, E. Seim and M.C. da Silva. 1990b. Increased post-perinatal mortality among children of mothers exposed to measles during pregnancy. *American Journal of Epidemiology* 132:531-539.
- Aaby, P., J. Clements and N. Cohen. 1987. *Key Issues in Measles Immunization Research: A Review of the Literature*, Geneva: World Health Organization.
- Aaby, P., H. Coovadia, J. Bukh, I.M. Lisse, A.J. Smits, A. Wesley and P. Kiepiela. 1985. Severe measles: a reappraisal of the role of nutrition, overcrowding and virus dose. *Medical Hypotheses* 18:93-112.
- Aaby, P., K. Knudsen, T.G. Jensen, J. Thørup, A. Poulsen, M. Sodemann, M.C. da Silva and H. Whittle. 1990. Measles incidence, vaccine efficacy and mortality in two urban African areas with high vaccination coverage. *Journal of Infectious Diseases* 162:1043-1048.
- Aaby, P. and W.H. Lamb. 1991. Sex and transmission of measles in a Gambian village. *Journal of Infection* 22:287-292.
- Aaby, P. and J. Leeuwenburg. 1990. Patterns of transmission and severity of measles infection. A reanalysis of data from the Machakos area, Kenya. *Journal of Infectious Diseases* 161:171-174.
- Aaby, P. and J. Leeuwenburg. 1991. Sex and patterns of transmission of measles infection. A reanalysis of data from the Machakos area, Kenya. *Annals of Tropical Paediatrics* 11:397-402.
- Aaby, P. and K. Mølbaek. 1990. Siblings of opposite sex as a risk factor for child mortality. *British Medical Journal* 301:143-145.
- Aaby, P., I.R. Pedersen, K. Knudsen, M.C. da Silva, C.H. Hordhorst, N.C. Helm-Petersen, B.S. Hansen, B.S., J. Thørup, A. Poulsen, M. Sodemann and M. Jakobsen. 1989. Child mortality related to seroconversion or lack of seroconversion after measles vaccination. *Pediatric Infectious Diseases Journal* 8:197-200.
- Anderson, R.M. and R.H. May. 1982. Directly transmitted infectious diseases: control by vaccination. *Science* 215:1053-1060.

- Asano, Y., N. Nakayama, T. Yazaki, R. Kato, S. Hirose, K. Tsuzuki, S. Ito, S. Isomura and M. Takahashi. 1977. Protection against varicella in family contacts by immediate inoculation with live varicella vaccine. *Pediatrics* 59:37.
- Bhuiya, A., B. Wojtyniak, S. D'Souza, L. Nahar and K. Shaikh. 1987. Measles case fatality among under-fives: a multivariate analysis of risk factors in a rural area of Bangladesh. *Social Science and Medicine* 24:439-443.
- Black, F.L., W. Hierholzer, J.P. Wodall and F. Pinheiro. 1971. Intensified reactions to measles vaccine in unexposed populations of American Indians. *Journal of Infectious Diseases* 124:306-317.
- Bridges, E.L. 1950. *Uttermost Part of the Earth*. London: Hodder and Stoughton.
- Brownlee, J. 1920. Public health administration in epidemics of measles. *British Medical Journal* 1:534.
- Catlin, G. 1841. *Letters and Notes on the Manners, Customs, and Condition of the American Indian*. London: Author.
- Christensen, P.E., H. Schmidt, H.O. Bang, V. Andersen, B. Jordal and O. Jensen. 1953. An epidemic of measles in Southern Greenland, 1951. *Acta Medica Scandinavica* 144:313-322, 430-541.
- Clemens, J.D., B.F. Stanton, J. Chakraborty, A. Chowdhury, M.R. Rao, M. Ali, S. Zimicki and B. Wojtyniak. 1988. Measles vaccination and childhood mortality in rural Bangladesh. *American Journal of Epidemiology* 128:1330-1339.
- Corney, B.G. 1913. A note on an epidemic at Rotuma, 1911. *Proceedings of the Royal Society of Medicine* 6:138.
- Crosby, A.W. 1976. Virgin-soil epidemics as a factor in the original depopulation in America. *William and Mary Quarterly* 33:289-299.
- Davis, R. 1982. Measles in the tropics and public health practice. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 76:268-275.
- Denny F.W., A.M. Collier, F. Henderson and W.A. Clyde. 1977. Infectious agents of importance in airways and parenchymal diseases in infants and children with particular emphasis on bronchiolitis. *Pediatric Research* 11:234-236.
- Fargues, P. and O. Nassour. 1988. *Douze Ans de Mortalit  Urbaine au Sahel*. Paris: Presses Universitaires de France.
- Foster, S.O. 1984. Immunizable and respiratory diseases and child mortality. *Population and Development Review* 10 (supplement):119-140.
- Gannelon, C. 1892. *La Rougeole   l'Hospice des Enfants-Assist s*. Paris: G. Steinheil.
- Garenne, M. and P. Aaby, P. 1990. Pattern of exposure and mortality in Senegal. *Journal of Infectious Diseases* 161:1088-1094.
- Garenne, H. and P. Cantrelle. 1986. Rougeole et mortalit  au Senegal:  tude de l'impact de la vaccination effectu e a Khombole 1965-1968 sur la survie des enfants. Pp. 515-532. in *Estimation de la Mortalit  du Jeune Enfant (0-5 Ans), pour Guider les Actions de Sant  dans les Pays en D veloppement*, ed. P. Cantrelle, S. Dormont, P. Farques, J. Goujard, J. Guignard, C. Rumeau-Rouquette. Paris: INSERM.
- Godfrey, E.S. 1928. Measles in institutions for children. *Journal of Preventive Medicine* 2:1-33, 251-272.
- Goody, J. 1973. *The Character of Kinship*. Cambridge: Cambridge University Press.
- Hall, W.J., N. Nathanson and A.D. Langmuir. 1957. The age distribution of poliomyelitis in the United States in 1955. *American Journal of Hygiene* 66:214-234.
- Hartfield, J. and D. Morley. 1963. Efficacy of measles vaccine. *Journal of Hygiene (Cambridge)*, 61:143-147.
- Hendrickse, R.G. 1975. Problems of future measles vaccination in developing countries. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 69:31-34.
- Hertz, J.B., T.B. S rensen and L. Vejerslev. 1977. Morbillid dsfald i Danmark 1958-1969, *Ugeskrift L ge* 138:589-593.

- Holt, E.A., R. Boulous, N.A. Halsey, I.M. Boulous and C. Boulous. 1990. Childhood survival in Haiti: protective effect of measles vaccination. *Pediatrics* 85:188-194.
- Karzon, D.T., G.L. Eckert, A.L. Barron, et al. 1961. Aseptic meningitis epidemic due to Echo 4 virus. *American Journal of Diseases of Childhood* 101:610-622.
- Kasongo Project Team, 1981. Influence of measles vaccination on survival pattern of 7-35-month-old children in Kasongo, Zaire. *Lancet* 1:764-767.
- Kinnear, W. 1923. Epidemic of measles in the Highland Division at Bedford, 1914-1915. *Edinburgh Medical Journal* 30:593-599.
- Koenig, M.A., M.A. Khan, B. Wojtyniak, J.D. Clemens, J. Chakraborty, V. Fauveau, J.F. Phillips, J. Akbar and U.S. Barua. 1990. Impact of measles vaccination on childhood mortality in rural Bangladesh. *Bulletin of the World Health Organization* 68:441-447.
- Koster, F.T. 1988. A review of measles in Bangladesh with respect to mortality rates among primary versus secondary cases. *Reviews of Infectious Diseases* 10:471-473.
- Lamb, W.H. 1988. Epidemic measles in a highly immunized rural West African (Gambian) village. *Reviews of Infectious Diseases* 10:457-462.
- Lancaster, H.O. 1952. The mortality in Australia from measles, scarlatina and diphtheria. *Medical Journal of Australia* 2:272-276.
- Lancet* 1985. Infant mortality and family structure, 2:650.
- Mayer, J. 1969. Measles and the state of nutrition. *Post Graduate Medicine* 45:253-253.
- McGlashan, N.D. 1969. Measles, malnutrition in blindness in Luapula Province, Zambia. *Tropical and Geographical Medicine* 21:157-162.
- McGregor, I.A. 1964. Measles and child mortality in the Gambia. *West African Medical Journal* 14:251-257.
- McKeown, T. and R.G. Brown. 1955. Medical evidence related to English population changes in the eighteenth century. *Population Studies* 9:119-141.
- McKeown, T., R.G. Brown and R.G. Record. 1972. An interpretation of the modern rise of population in Europe. *Population Studies* 26:345-382.
- McKeown, T. and R.G. Record. 1962. Reasons for the decline of mortality in England and Wales during the nineteenth century. *Population Studies* 16:94-122.
- McKeown, T., R.G. Record and R.D. Turner. 1975. An interpretation of the decline of mortality in England and Wales during the 20th Century. *Population Studies* 16:94-122.
- McNeill, W.H. 1976. *Plagues and Peoples*. London: Blackwell.
- Mercer, A.J. 1985. Smallpox and epidemiological-demographic change in Europe: the role of vaccination. *Population Studies* 39:287-308.
- Mims, C.A. 1976. *The Pathogenesis of Infectious Diseases*. London: Academic Press.
- Monastiri, H. 1961. Quelques donnŽes statistiques relatives   la mortalit  par rougeole dans la Commune de Tunis. *La Tunisie M dical* 39:179-187.
- Morley, D. 1973. *Paediatric Priorities in the Developing World*. London: Butterworth.
- Mosley, W.H. 1985. Will primary health care reduce infant and child mortality? A critique of some current strategies. With special reference to Africa and Asia. Pp. 103-137 in *Health Policy, Social Policy and Mortality Prospects*, ed. J. Vallin and A.D. Lopez. Li ge: Ordina.
- Mosley, W.H. and L.C. Chen. 1984. An analytical framework for the study of child survival in developing countries. *Population and Development Review* 10 (Supplement):25-45.
- Neel, J.V., W.R. Centerwall, N.A. Chagnon and H.L. Casey. 1970. Notes on the effect of measles and measles vaccine in a virgin-soil population of South American Indians. *American Journal of Epidemiology* 91:418-429.

- Nutels, N. 1968. Medical problems of newly contacted Indian groups. Pp 68-76, in *Biomedical Challenges Presented by the American Indian*, Scientific Publication no 165. Washington DC: PAHO.
- Peart, A.T.W. and F.P. Nagler. 1954. Measles in the Canadian Arctic. *Canadian Journal of Public Health* 45:146-156.
- Pfeilsticker, A. 1863. Beitrage zur Pathologie der Masern mit besonderer Berÿcksichtigung der statistischen Verhÿltnisse, Tübingen, L. Fr. Fues. Thesis.
- Picken, R.M.F. 1921. The epidemiology of measles in rural and residential areas. *Lancet* 1:134-953.
- Pison, G., P. Aaby and K. Knudsen. 1992. Increased risk of measles mortality for children with a sibling of the opposite sex among the Fula Bande and Niokholonko, Senegal. *British Medical Journal* 304:284-287.
- Razzell, P.E. 1974. 'An interpretation of the modern rise of population in Europe' – a critique. *Population Studies* 28:5-17.
- Reder, J. 1918. Die Bedeutung der Masern in Sammelniederlassungen nach den im k.k. Flÿchtlingslager Gmÿnd gemachten Wahrnehmungen. *Zeitschrift fÿr Kinderheilkunde* 18:355-370.
- Reves, R. 1985. Declining fertility in England and Wales as a major cause of the twentieth century decline in mortality. The role of changing family size and age structure in infectious disease mortality in infancy. *American Journal of Epidemiology* 122:112-126.
- Rÿnne, T. 1985. Measles virus infection without rash in childhood is related to disease in adult life. *Lancet* 1:15.
- Scheifele, D.W. and C.E. Forbes. 1972. Prolonged giant cell excretion in severe African measles. *Pediatrics* 50:867-873.
- Smellie, L. 1846. Rubeola among the North American Cree. *Monthly Journal of Medical Science* 6:413-418.
- Squire, D. 1877. Report on the very fatal epidemic of measles in Fiji. *Medical Times Gazette* 1:323-324.
- Wilson, G.N., 1905. Measles: its prevalence and mortality in Aberdeen. *Public Health* 18:65.

Sexual behaviour in the face of risk: preliminary results from first AIDS-related surveys*



John Cleland¹, M. Cara^{1,2}, J-C. Deheneffe³ and B. Ferry²

¹London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK

²Global Programme on AIDS, WHO

³SONECOM, Belgium

Abstract

Preliminary results are presented from nationally representative surveys of the adult populations of five African countries, conducted in 1989 and 1990. General awareness of AIDS was high, as was knowledge of sexual transmission. In four of the five surveys, large proportions, from 25 to 64 per cent, of both men and women perceived themselves to have a high or moderate risk of HIV infection. High proportions also reported that they had modified their behaviour typically by more care in selecting partners or greater faithfulness. Greater use of condoms was mentioned rarely. The results, particularly on behavioural change, should not be interpreted literally. But the fact that so many report modification of behaviour at least suggests a willingness to contemplate the need for change. The prognosis would have been much worse, had these surveys indicated widespread denial of risk and unwillingness to consider changes in behaviour.

Introduction

It is widely accepted that, pending the development of an effective vaccine or therapy for HIV/AIDS, behavioural change is the only means of averting the continued spread of the disease. The statement could go further. The advent of effective biomedical prevention or treatment is unlikely to bring a complete solution to the problem, unless accompanied by changes in sexual behaviour. After all, penicillin did not eradicate syphilis or gonorrhoea.

The central theme of this paper is the behavioural response to the advent of the HIV pandemic as reported in representative surveys of the general population of five African countries. Single cross-sectional surveys have severe limitations for analysing change as well as surveys of the general population. Evaluation of specific interventions among specific target groups would perhaps give a more accurate picture. Repeated surveys, or better still prospective studies, are required for any confident assessment of behavioural change. Nevertheless, some insights can be gained from single surveys that enquire into knowledge of HIV transmission, sexual behaviour, perceived risk and intended or reported change.

* The survey data used in this paper were processed by WHO/GPA Social and Behavioural Research Unit which coordinated these surveys. The Principal Investigators should be acknowledged for authorizing the use of these data in a comparative approach: Central African Republic: Dr Pierre Somse, Chief of the National AIDS Programme; CTMte d'Ivoire: Professor Seri Dedy, Universit^Z d'Abidjan and Professor Goz^Z Tape, Ecole Normale Sup^Zrieure d'Abidjan; Lesotho: Dr Agatha Lawson, WHO Consultant for the Ministry of Health National AIDS Control Programme; Mauritius: Dr Clement Chan Kam, National AIDS Coordinator; Togo: Dr Genevieve Awissi, National AIDS Programme.

The surveys are part of a collaborative research program, conducted under the auspices of the Global Programme on AIDS (GPA) at the World Health Organization. Field work was done in 1989 and 1990, during the first year of attempts to check the spread of HIV. It is likely that the situation has changed in recent years and thus it would be inappropriate to assume that these results portray the current situation.

The choice of geographical focus for this paper is partly pragmatic. There is more relevant information from this region than for other developing regions. The epidemiology of the disease is also quite different in sub-Saharan Africa from its epidemiology in Europe or North America. Moreover, the demographic impact of AIDS is likely to be most severe in Africa first. We consider briefly this issue of impact before proceeding to the main topic of response.

Demographic impact of AIDS in Africa

A vivid illustration of the uncertainties associated with attempts to project the spread of AIDS and its demographic consequences is provided by a recent report of a modelling workshop, held in December 1989 under the auspices of the United Nations Population Division and the World Health Organization (UN, WHO 1991). In essence the UN provided standard input values and asked a number of mathematical modellers to project trends in HIV infections, AIDS mortality and standard demographic indicators over a 25-year period from an initial infection level of 3.5 per cent. The fixed values included the age distribution, life tables (in the absence of AIDS), marriage and fertility rates, probabilities of progression from infection to AIDS and to death, the proportions of the adult population in specified sexual-behaviour strata, and the probability of vertical transmission (0.4). Heterosexual infectivity (the probability of transmission of the disease per sexual contact) was allowed to vary to yield best, worst and intermediate scenarios, as was the proportion of sexual contacts with condom use (2%–20%).

Key results from six models are shown in Table 1. Intermediate scenarios have been taken, which assume a male-to-female infectivity of 0.03 and a female-to-male value of 0.01 per sexual contact. These probabilities are thought to be realistic though much uncertainty about infectivity remains. The huge diversity of predictions is immediately apparent: the projected level of HIV infection ranges from three to 40 per cent; life expectancy from 26 to 58 years; and the rate of natural increase from 0.5 to 2.8 per cent. Moreover, there is no apparent consistency between the level of infection at the end of the 25-year period and demographic consequences. For instance, the Bulatao model shows an HIV prevalence of nearly 40 per cent but a higher life expectancy than the Palloni model, which projects a level of HIV infection of only 2.8 per cent.

No attempt was made by the workshop participants to reconcile these differences though there was general agreement that, apart from the assumed value for infectivity, the handling of sexual behaviour in the models was the major source of divergence in results. The work of Anderson and colleagues is particularly relevant here (e.g. Anderson et al. 1989). They have shown convincingly that the degree of sexual contact between different sexual-behaviour strata has important implications for the spread of HIV. To the extent that individuals with high-risk behaviours seek similar partners, the initial spread of the infection will be rapid but it will stabilize at a relatively low level. Conversely, less polarized patterns of sexual behaviour imply a slower rate of diffusion but a higher eventual prevalence. Age segregation of sexual partnerships is a further consideration. The greater the sexual contact between older men (among whom HIV infection is likely to be high) and younger women, the greater is the potential demographic impact of AIDS, because of enhanced probabilities of vertical transmission from mothers to children.

Table 1
Model estimates of HIV prevalence, life expectancy and rate of natural increase at end of 25-year projection period

Author of model	HIV prevalence (%)	Life expectancy (eo)	Annual natural increase
Auvert	31.0	26	0.5
Brouard	15.0	NA	2.4
Bulatao	39.5	47	2.5
Dietz	21.2	36	1.5
IAG	3.5	58	2.6
Palloni	2.8	42	2.8

Though subsequent and perhaps more realistic models, such as IWGAIDS and SIMULAIDS, have been developed, the intermediate scenario of the Bulatao-World Bank model received wide endorsement from international agencies for a number of years (e.g. Chin 1991). It is therefore worth examining both the results and the behavioural inputs to this model and comparing them to the admittedly sparse and fragile empirical evidence.

Detailed projections are shown in Table 2. In the absence of AIDS, very large, and indeed optimistic, falls in mortality were expected over the 25-year period. These gains are eliminated and even slightly reversed by AIDS. Thus life expectancy, instead of rising by ten years, falls by four years. However, the rate of natural increase remains strongly positive at an annual rate of 2.4 per cent compared to a projected 3.0 per cent in the absence of AIDS.

Behavioural inputs to the model were based on a handful of small-scale studies and are shown in Table 3. The adult male population is composed of three main groups: single (13%); married monogamous (26%) and married non-monogamous (61%). Among the latter group, 26% are assumed to have contact with prostitutes (with 24 partners on average per annum) while the remainder experience 'casual' sex and have eight partners on average per year. The females are more likely than males to be classified as monogamously married (50%) but over one third (36%) are assumed to be married but non-monogamous (with 12.6 partners per year on average). Nearly five per cent of females (all single) are defined as prostitutes. Summing across all strata, these model values imply about 12 partners per year for an average adult person.

How do these model values relate to more recent empirical evidence? The only body of data based on nationally representative samples in Africa comes from the surveys sponsored by the Global Programme on AIDS (GPA) at WHO. Key results from the earlier surveys may be found in Cara'l et al. (1990). A few indicators reveal the huge gulf between model assumptions and self-reported behaviour. The model assumes that three-quarters of adult men will experience commercial or casual sex in a 12-month period. The average value from four surveys (Central African Republic, CTMte d'Ivoire, Lesotho and Togo) is only 30 per cent. As mentioned earlier, the model takes an overall average of 12 non-marital partners per year. The results from the GPA surveys vary between 0.6 and 2.6 partners per annum for men. And finally, the model assumes that 39 per cent of men have commercial sex in a year, whereas the results of four GPA surveys give a range of eight to 13 per cent.

Table 2
Results of the Bulatao-World Bank model (intermediate scenario)

--	--

	Period 1985–1990	Period 2005–2010	
		Without AIDS	With AIDS
Rate of natural increase	3.4	3.0	2.4
Crude death rate	14.8	8.4	16.4
Infant mortality	101.2	61.3	104.7
Childhood mortality (5q0)	180.0	90.0	215.0
Life expectancy	51.5	61.4	47.0

Table 3
Behavioural values used in the Bulatao-World Bank model (intermediate scenario)

	Male		Female	
	% of adult population	Mean partners p.a. (excl. spouse)	% of adult population	Mean partners p.a. (excl. spouse)
Single	13.0	24.0	14.0	52.6
(prostitute)	(-)	(-)	(4.3)	(143.0)
(other)	(13.0)	(24.0)	(9.7)	(12.6)
Married monogamous	26.0	0	50.0	0
(with monogamous spouse)	(15.0)	(0)	(15.0)	(0)
(with non-monogamous spouse)	(11.0)	(0)	(35.0)	(0)
Married non-monogamous	61.0	14.8	36.0	12.6
(contact with prostitutes)	(26.0)	(24.0)	(-)	(-)
(other)	(35.0)	(8.0)	(36.0)	(12.6)

It is quite possible that surveys of sexual behaviour in Africa suffer from severe underreporting. Nevertheless, it is difficult to accept that this factor alone accounts for the discrepancies noted above. It seems likely that the initial modelling of the disease and its impact was based on exaggerated assumptions of sexual mobility in Africa. The advent of more empirical data on sexual behaviour has already led to important modifications but it remains true that the sophistication of mathematical models of HIV transmission has far outstripped their empirical foundations. The medium to long-term prospects for the spread of HIV infection still remain essentially unknown.

This opinion is not intended to engender complacency about the future. Highly infected areas face a devastating prospect in the short term, and, even in modestly infected countries, the harsh consequences of the disease for already meagre health services can hardly be exaggerated.

The response of AIDS: introduction

Any comprehensive account of the response to AIDS, in Africa or elsewhere, would have to take into account international, national, institutional, community and individual responses. We cannot attempt such an ambitious task in this paper but will confine ourselves to a few observations.

The international response to AIDS has been orchestrated to a large extent by the Global Programme on AIDS (GPA) at WHO. The GPA has made two highly significant contributions. First it has encouraged governments to recognize officially the threats, or potential threats, of the disease. All African countries now have a National AIDS Program with the mandate of planning and coordinating priority activities. The latter are usually formalized in five-year strategic plans.

The proposed budgets for the five countries that have conducted surveys and published results (to be discussed in the next section) are shown in Table 4.¹ The pattern of proposed expenditure differs considerably between countries. Large budget items may include management and administration, blood-transfusion services, laboratory equipment, training of medical and paramedical staff, counselling, and epidemiological surveillance. Most important for the purpose of preventing the spread of infection is the amount allocated directly to public information and education campaigns. In Mauritius and Togo, about one-third of the total proposed budget is devoted directly to IEC. In the Central African Republic, the proportion is higher (56%) but in Lesotho it is much lower.

Table 4
Proposed budgets for anti-HIV/AIDS programs, in selected African countries

Country	Plan period	Sum proposed per head population (US\$)	Percentage allocated to IEC
Central African Republic	1987-92	1.9	56
C TM te d'Ivoire	1989-93	0.3	39
Lesotho	1989-93	0.7	17
Mauritius	1988-92	1.0	32
Togo	1990-92	0.7	37

Source: National AIDS medium-term plans

It is never possible to gauge from government plans and associated budgets the extent of real commitment and effort; thus the substantial sums shown in Table 4 are no guarantee of concerted or effective action. What is needed in the longer term is an equivalent for AIDS of the well known index of family-planning program strength, developed by Berelson, Mauldin and Lapham (Lapham and Mauldin 1972). Some such development is surely essential for the evaluation of anti-AIDS activities, which is certain to become a high priority in the future.

The second significant achievement of GPA has been to avert, or at least diminish, coercive and punitive responses to the HIV pandemic. Brandt (1988) reminds us that the hysteria over syphilis during the First World War led to compulsory testing, and incarceration of infected prostitutes, though neither of these coercive measures proved effective. The possibility of similar reactions to AIDS were foreseen early and appropriate countermeasures taken.

Indeed, it could be argued that concern for the welfare of HIV-positive persons has gone too far when it takes the form of not informing those at high risk of infecting others of their status when this information is available from screening. This brings us to one of the central policy and moral dilemmas. The following scenario is an increasingly common one in the larger African cities. There are, say, 10,000 prostitutes: women who identify themselves as such and have sex with anonymous clients for money. As a group they have perhaps several hundred clients per year. The level of HIV infection among clients, as indicated by male patients attending STD clinics, may be 25 per cent, while 70 per cent of prostitutes themselves are infected. Some ten per cent of the general adult population of the city is HIV positive (as measured by testing of pregnant women). Condom use is rare in commercial sex encounters and even less common between spouses.

¹ Of course, it would be preferable to have information on actual expenditure but this is not yet available.

One policy option would be to inform HIV-positive prostitutes, counsel them and find alternative employment. But job opportunities for women are likely to be scarce and it is not feasible politically to give positive discrimination to former prostitutes in the job market. Moreover, new entrants to the profession would soon become infected by clients. It seems that all options are foreclosed except vigorous promotion of condoms, both by intensive education of prostitutes themselves and public information campaigns to influence clients and potential clients. There is some evidence that educational interventions for prostitutes can be effective, not least perhaps because they are relatively accessible and have a group identity. For instance, use of condoms rose among Nairobi prostitutes from ten to 70 per cent over a relatively short period of time, though they were used only in one-quarter to one-third of all sexual encounters (Ngugi et al. 1988).

If the relative success of interventions among prostitutes rests to some extent on the group dimension with mutual reinforcement of behavioural change, the importance of the response to AIDS at community and institutional level is underscored. Groupings based on common activity (e.g. the army, police force, large private companies, institutions of secondary and higher education), on a common religious or political belief, and on gender may prove to be the key mediators of change.

It is thus unfortunate that the dominant theories of risk behaviour come from social psychology and emphasize the individual. As Weinstein (1987) puts it, the central image of most theories is of an individual weighing the costs of taking some precaution against the perceived benefits. The key concepts are typically: knowledge, or beliefs, about cause and prevention, and about the likelihood and severity of harm; perceived effectiveness of possible preventive measures and the costs or barriers associated with them. These ingredients are most apparent in Becker's Health Belief Model (Becker 1974). A more sociological dimension is apparent in Fishbein and Ajzen's theory of reasoned action, in the form of the concept of subjective norms: the expectations by significant others and the strength of the motivation to conform to them (Fishbein and Ajzen 1975). A final important element of many risk-behaviour theories derives from Rotter's work on perceived behavioural control, that is the individual sense of ability to determine or influence outcome (Rotter 1966, Ajzen and Madden 1986). One feature that is consciously absent from most theories is values related to risk itself. Such values are typically ambivalent. Thus risk taking in some spheres of life is strongly endorsed while in others it is regarded as irresponsible.

These and other related models have been used to predict, with varying degrees of success, the propensity of individuals to initiate and sustain a range of risk-reduction behaviours. They have also been used, particularly in the USA, to examine sexual risk behaviour in relation to HIV. (For a major review see Becker and Joseph, 1988.) While they have often discriminated powerfully between individuals who modify their behaviour in response to an intervention and those who do not, it is doubtful whether they can predict or explain the broad societal change in sexual-behaviour norms that is required to avert the continued spread of the HIV virus in Africa. Moreover sexual risk behaviour differs from many other forms of risk behaviour. Unlike smoking for instance, there is less diffuse social pressure to change because it is a private activity. Unlike smoking also, change may involve very delicate negotiation between partners. Nevertheless, the theories and models of social psychologists may continue to dominate research, partly because of lack of alternatives from the disciplines of sociology or social anthropology, and partly because they are amenable to quantification in social surveys. It is to the contribution of surveys in the elucidation of the behavioural response to AIDS that we now turn.

Behavioural response to AIDS: evidence from surveys

The Social and Behavioural Research Unit (SBR) within the GPA designed three main questionnaires for use among the general population (GPA 1988; Carballo et al. 1989). The most commonly used

instrument, the KABP questionnaire, concentrates on knowledge, attitudes and beliefs about HIV/AIDS but also contains a limited number of questions on sexual behaviour as an optional module. The second instrument, the Partner Relations (PR) questionnaire, is focused more explicitly on sexual behaviour, though it also includes sections on knowledge and beliefs about AIDS. A third instrument was designed, a combined KABP-PR questionnaire, in response to the demand for a survey that would encompass both sexual behaviour and knowledge and beliefs about HIV/AIDS.

This central material on AIDS-related variables is supplemented by background information on education, literacy, media exposure, ethnicity, religion, occupation and characteristics of the community. There are also sections on condoms (knowledge, perceived access and attitudes) and on symptoms of sexually-transmitted diseases (STD) in the last 12 months. The rationale for these contents is descriptive rather than theoretical, and epidemiological rather than sociological. They are intended primarily to document the climate of opinion regarding AIDS and incidence of sexual behaviours that carry a risk of HIV infection, namely intercourse with casual or commercial partners without regular use of condoms. Though some insight into the indirect determinants of these behaviours, and changes thereof, may be gained by examining variations in relation to social, economic, marital and community characteristics, there is little attempt to measure underlying values and subjective meanings and thereby provide a more comprehensive understanding. An overriding need is to provide simple descriptions of knowledge, perceived risk, and sexual behaviours, and to monitor changes over time. Surveys have the potential to meet this need. Conversely they have limited potential to elucidate the cultural context of sexual behaviour (Abramson and Herdt 1990). For this, complementary anthropological and other forms of more intensive study are required (GPA 1989).

Shown below are the countries in Africa that have participated in the three types of survey under the guidance and coordination of GPA.

KABP	PR	Combined
Burundi Central African Republic Guinea-Bissau Kenya Mauritius Rwanda Tanzania Togo	C TM te d'Ivoire Tanzania	Lesotho Nigeria S.W. Nigeria (2 states) Uganda Zambia

Not all these surveys have been completed. There are however a large number of additional surveys, some funded by GPA, that were not part of the collaborative program. The quality of these surveys varies widely and few are based on representative samples. Their results are not considered in this paper.

KABP and PR surveys faced many problems of definition and measurement. Overshadowing these concerns is the question of validity. There are many who doubt that standardized surveys can yield data of reasonable quality on such a topic as sexual behaviour (Muhondwa 1988; Ankrah 1989). Similar doubts were raised about family-planning surveys 30 years ago but were later shown to be largely unfounded. Whether a similar verdict on sexual-behaviour surveys will be eventually reached remains uncertain. Detailed analysis of the internal consistency of data will provide some indications but assessments of reliability and validity by means of special field studies are a high priority. At this

early juncture, a degree of scepticism is appropriate and the results to be presented in this paper should be interpreted very cautiously.

This paper summarizes preliminary survey findings on the reported behavioural response to AIDS based on nationally representative samples of the general population in CTMte d'Ivoire, Lesotho, Togo, Central African Republic and Mauritius. These countries are those in which social researchers were willing and able to participate. Sample sizes range from 1600 to 2500 and the age range covered varies from 15 years and over to 15–44. All these surveys have been conducted by teams of national investigators in 1988–90 in collaboration with SBR/GPA.

These five countries are not among the worst affected by AIDS in Africa. Recently published data derived from the US Bureau of the Census HIV/AIDS surveillance data base indicate that the prevalence of HIV-1 among the general sexually-active urban population is about seven per cent in the Central African Republic, about five per cent in CTMte d'Ivoire but only 0.1 per cent in Lesotho (Anderson et al. 1991). No parallel data are available for Togo or Mauritius, though the level of infection is thought to be very low. In Togo, the prevalence of HIV in 1987 among the general population was found to be 0.1 per cent, though higher among blood donors in the capital (R^Žpublique du Togo 1989). In Mauritius, a few cases of infection have been detected. All five countries have national AIDS program and public information campaigns.

One undoubted ability of mass-media campaigns is to disseminate rapidly increased awareness of a topic. This ability is reflected in the relatively high proportion of all adults who have heard of AIDS and who are aware of sexual transmission (Table 5). The results suggest that awareness is almost synonymous with knowledge of sexual transmission. However, knowledge of sexual transmission coexists with beliefs or uncertainties about casual transmission. Even in highly-educated countries, it has proved difficult to dispel fears of infection from superficial contact with public facilities such as toilets (Becker and Joseph 1980).

Table 5
Knowledge of AIDS

	C.A.R.	C TM te d'Ivoire	Lesotho	Mauritius	Togo
Percentage aware of AIDS	84	83	94	92	64
Of those aware, percentage who know about sexual transmission	96	90	97	94	95
Among those aware of AIDS and sexually active, percentage who know someone with AIDS	21	7	10	4	5

Personal knowledge of someone with AIDS, a potential predictor of behavioural change, is understandably much lower, ranging from 21 per cent in the Central African Republic to four per cent in Mauritius (Table 5).

Details of reported sexual behaviour are given in Table 6 and summarized in Figure 1. Classification of sexual relationships, particularly in Africa, is extremely complex and any standardized survey has to use a vastly oversimplified approach. The system adopted in the GPA surveys was to classify all sexual relationships that had lasted a year or more, or were expected to continue for at least one year, as regular partnerships. Marriages fall into this category. At the other extreme, sexual

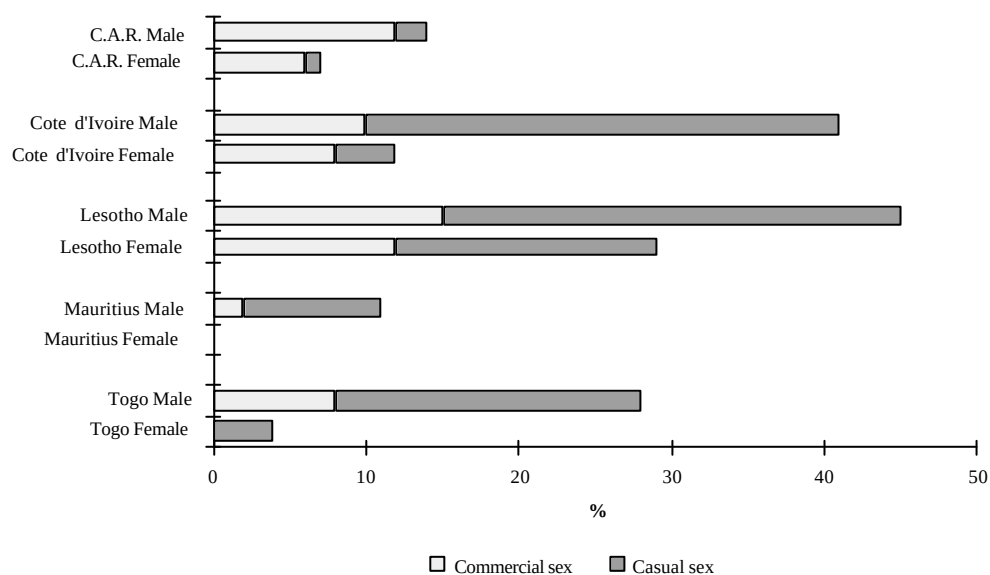
encounters involving the giving or receipt of money or gifts were defined as commercial sex. The intermediate category was termed 'casual sex' and included all sexual encounters or relationships that were neither commercial nor regular, in the senses described above.

We do not know exactly how respondents interpreted the questions that underlay this classification and it is easy to imagine confusion, particularly between commercial and casual sex. There is nevertheless a certain consistency in the proportions of males reporting commercial sex in the preceding 12 months, which range from eight to 15 per cent in four of the five surveys. It is perhaps not surprising that culturally distinct Mauritius is the outlier in this regard, with only two per cent of men reporting commercial sex. Table 6 indicates that regular use of condoms was rare in these commercial encounters.

The reporting by men of casual, but not commercial, sex is much more variable between countries. At face value, the results imply that casual sex is much more common than commercial sex in CTMte d'Ivoire, Lesotho and Togo but less common in the Central African Republic.

The testimony of women regarding their sexual behaviour is usually regarded with greater scepticism than that of men. Figure 1 shows clearly the now familiar difference, with higher reporting of non-marital sex by men than by women. Indeed in Mauritius, no female respondent reported sexual contact outside of a regular partnership while in Togo only four per cent did so.

Figure 1
Per cent reporting commercial and casual sex in the last 12 months, among those who have heard of AIDS



A large difference between men and women in the period prevalence of commercial sex is to be expected, because of the unbalanced ratio of prostitutes to clients. For casual sex, a greater equality might be anticipated. However, the opposite tendency is apparent in both CTMte d'Ivoire and Lesotho; in these surveys, the disparity between men and women is much greater for casual than for commercial

sex. This imbalance suggests that male and female respondents may interpret questions differently and that women may underreport casual sex.

Perceived personal risk of harm is a central element of risk-behaviour theories. The standard question on this topic reads: 'What are the chances that you yourself might get AIDS?' When interpreted literally, answers define the segment of the population that may be prepared to modify behaviour. It is thus analogous to stated desire to limit or postpone children in family-planning research. It is of interest, therefore, to assess the general population in terms of perceived risk of HIV/AIDS and to examine how these perceptions relate to reported behaviour.

The distribution of the population in terms of perceived risk shows substantial proportions stating that they have a moderate or high risk, in all countries except Mauritius (Table 7 and Figure 2). Little difference is apparent between the sexes except in CTMte d'Ivoire. There is no evidence here of widespread denial. Indeed, perception of risk is high in relation to self-reported non-marital sex, particularly in the Central African Republic. In addition, sizeable minorities of both men and women were not prepared to label themselves, giving a 'don't know' response.

Table 6

Per cent distribution according to reported sexual behaviour in last 12 months, among sexually experienced reported respondents who have heard of AIDS

	Commercial sex		Casual sex	No casual sex	No sex	Total
	without condoms	always with condoms				
C.A.R.						
Males	10	2	2	79	6	100
Females	5	1	1	85	9	100
C TM te d'Ivoire						
Males	9	1	31	52	7	100
Females	8	0	4	81	6	100
Lesotho						
Males	13	2	30	33	22	100
Females	11	1	17	49	22	100
Mauritius						
Males	2		9	63	26	100
Females	0		0	71	28	100
Togo						
Males	7	1	20	57	16	100
Females	0	0	4	81	15	100

Do feelings of susceptibility relate in the expected manner to reported behaviour at the individual level? The relationship between behaviour and risk perception is explored in Table 8 and Figure 3 for

male respondents only.² As may be seen there is only a modest relationship between 'objective' and perceived risk.

Except in Mauritius, appreciable proportions of men who reported no sex at all in the last 12 months, or no casual sex, nevertheless perceive themselves to have a moderate or high risk of getting AIDS. The proportions are somewhat higher among those reporting commercial or casual sex (particularly in Lesotho) but the relationship with behaviour is modest.

² The numbers of women reporting commercial or casual sexual are too small in several surveys to sustain further analysis. Moreover, it is probable that women perceive risk from their husbands' behaviour rather than their own.

Table 7
Per cent distribution according to perceived personal risk of getting AIDS among sexually experienced respondents who are aware of AIDS

	Not likely	Low	Don't know	Perceived risk Moderate	High	Total
C.A.R.						
Males	22	21	NA	NA	56	100
Females	20	15	NA	NA	64	100
C ^M te d'Ivoire						
Males	22	20	23	39	6	100
Females	36	8	27	21	8	100
Lesotho						
Males	34	19	20	18	9	100
Females	43	13	22	14	9	100
Mauritius						
Males	70	20	7	NA	3	100
Females	62	17	17	NA	4	100
Togo						
Males	37	26	11	NA	27	100
Females	26	26	17	NA	31	100

Figure 2
Per cent reporting high (moderate) risk of getting AIDS, among those who have heard of AIDS

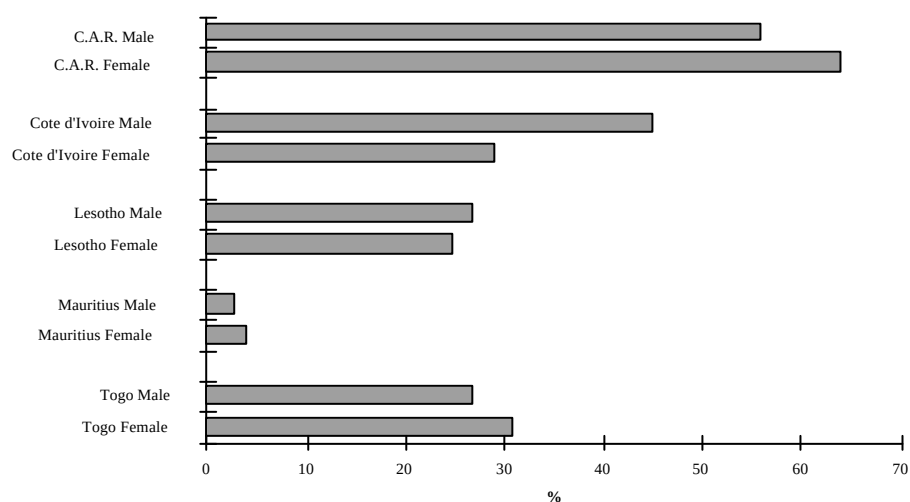


Table 8

Per cent distribution according to perceived risk of HIV infection, by sexual behaviour in last 12 months, among sexually experienced men aware of AIDS

	Commercial sex without condoms (%)	Casual sex (%)	No casual sex (%)	No sex (%)
Central African Republic				
High	57	73	56	47
Don't know	NA	NA	NA	NA
Low	28	8	20	30
Not likely	15	19	24	23
(n)	(108)	(26)	(852)	(70)
Côte d'Ivoire				
High	44	52	40	25
Don't know	19	20	26	25
Low	12	11	10	11
Not likely	25	17	24	39
(n)	(122)	(435)	(743)	(94)
Lesotho				
High	44	29	14	16
Don't know	13	23	24	16
Low	27	14	18	15
Not likely	16	34	44	53
(n)	(70)	(160)	(180)	(118)
Mauritius				
High	0	5	2	3
Don't know	8	5	8	11
Low	31	30	18	19
Not likely	61	60	72	67
(n)	(26)	(110)	(744)	(306)
Togo				
High	42	33	36	18
Don't know	6	5	13	17
Low	30	23	28	17
Not likely	22	39	23	48
(n)	(50)	(143)	(409)	(114)

The same impression is given when perceived risk is classified by the number of casual or commercial partners in the last 12 months (Table 9). In the three surveys for which this information is available, risk rises according to number of partners but not as steeply as might be expected.

There are a number of possible explanations for the weak relationship between reported behaviour and risk perception among men: underreporting of high-risk sexual behaviour; beliefs in casual or non-sexual transmission; changes in behaviour by those who perceive risk; or incomprehension of the question on perceived risk. Clearly, any interpretation of these survey data must be tentative, but it does appear that perception of risk is widespread and is not concentrated among those men whose behaviour puts them at higher than average risk.

Figure 3

Perceived risk of HIV infection by sexual behaviour in the last 12 months among those who have heard of AIDS

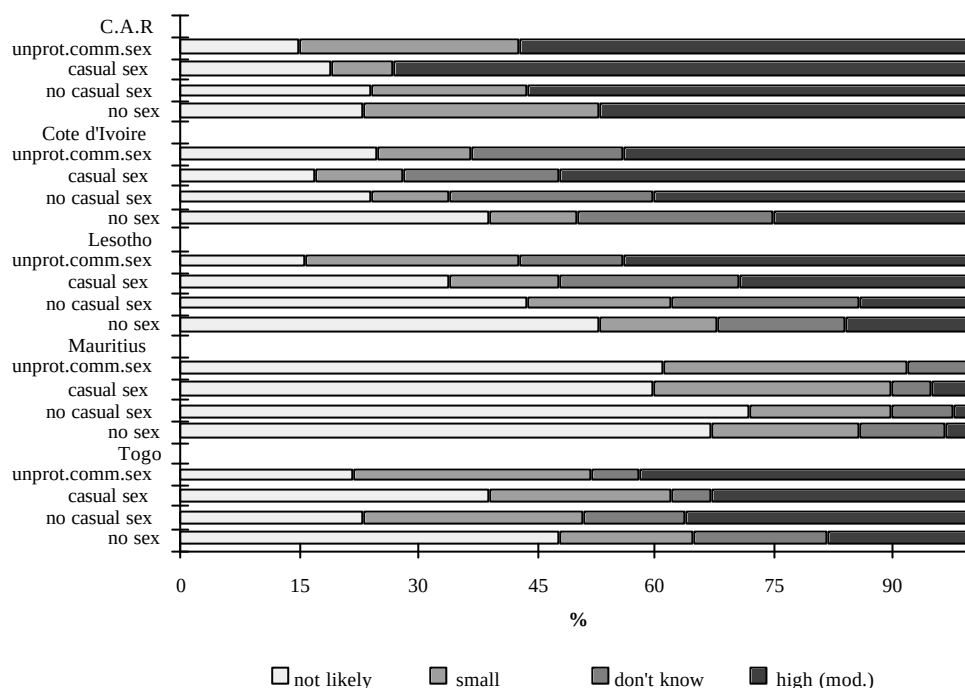


Table 9

Percentage who reported that they have a high (or moderate) risk of getting AIDS, by number of sexual partners, among sexually experienced men who are aware of AIDS

	Number of casual/commercial partners			
	0	1	2-4	5+
C TM te d'Ivoire	38	45	53	56
Mauritius	3	3	4	5
Togo	23	23	46	47

We come now to the most important issue of all: behavioural change. Respondents were asked whether they had changed their behaviour in response to the threat of AIDS, what changes they had made, and (for those reporting no change) whether they intended to change. Perhaps the wording or location of the question in the interview invites a positive response, because the proportions reporting change are surprisingly large (Figure 4). Over 60 per cent claim to have changed, or modified, their behaviour, except in Mauritius and among females in CTMte d'Ivoire and Togo. The dominant change reported (except in Mauritius) implies sexual behaviour (Table 10). Typical answers included in this category are 'greater faithfulness' and 'more care in selecting partners'. Use of condoms, however, is mentioned very rarely. The relationship between reported behavioural change and actual behaviour is complex, because no information was gathered on the timing of change. Thus the extent to which reported sexual behaviour in the last 12 months reflects changes or not is unclear. Nevertheless, Figure

5 reveals a consistent pattern in all surveys but the Mauritian one. The period prevalence of reported risk behaviour is highest among men who intend to modify behaviour but have not yet done so (admittedly a small group). This result enhances the credibility of data. The incidence of casual or commercial sex tends to be much higher (except in Lesotho) among those who report change than among those who report no change; this is consistent with the common finding in risk-behaviour research that change in response to a threat is more typically partial than absolute.

Figure 4

Per cent who report, or intend, behavioural change among those who have heard of AIDS

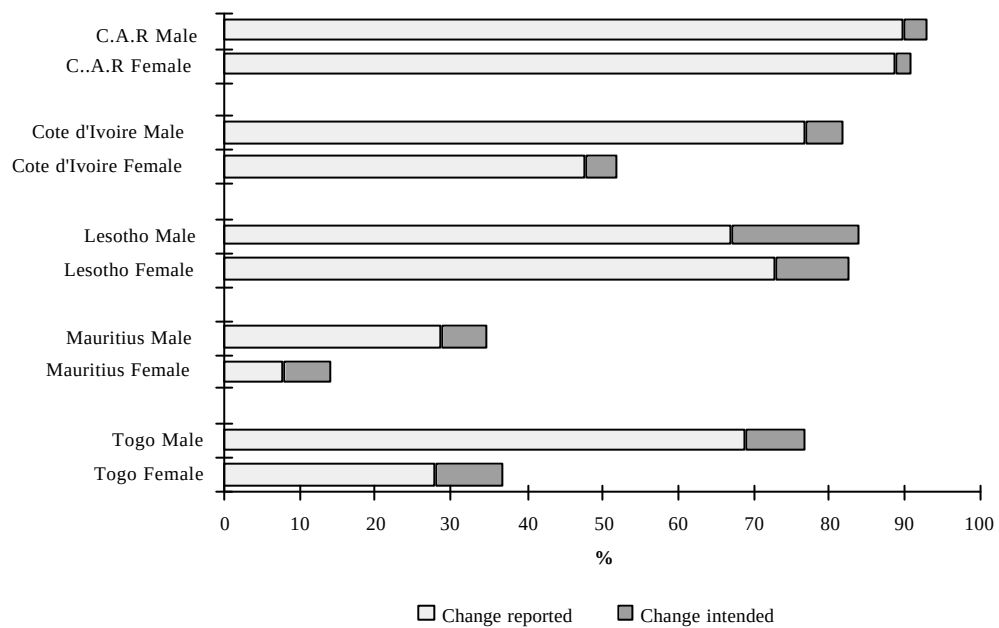


Table 10
Percentages mentioning specified changes, among those who report behavioural change

	Less casual/commercial sex	Greater use of condoms	Change Ineffective change implying beliefs in casual transmission	Other/ambiguous
C.A.R.				
Males	92	4	0	4
Females	98	0	0	2
C TM te d'Ivoire				
Males	59	6	1	33
Females	56	3	5	34
Lesotho				
Males	60	7	1	32
Females	60	2	1	36
Mauritius				
Males	47	1	14	38
Females	14	0	32	53
Togo				
Males	84	0	6	6
Females	73	1	11	15

Finally, Table 11 shows demographic and socioeconomic differentials in reported behavioural change. These are summarized in the form of relative risks: ratios of proportions. For instance, the top left figure of 0.97 is the proportion reporting change among men under 25 years of age, divided by the proportion among older men. Younger, more educated respondents and those reporting high exposure to mass media are more likely to report change in most countries, but the relative risks are not strikingly large. Perhaps the most important, and disturbing, feature is the lack of an appreciable urban-rural difference. In view of the greater concentration of HIV infection in urban areas, it is surprising that reported modification of behaviour is not more common among the urban than the rural population.

Figure 5
Per cent males reporting commercial and casual sex in the last twelve months, by behavioural change

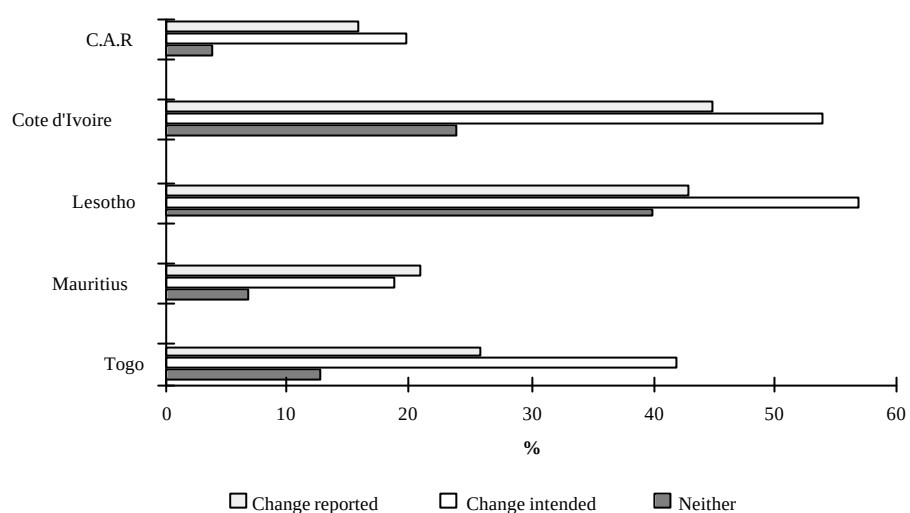


Table 11
Relative risks of reported behavioural change, among sexually-experienced respondents aware of AIDS

	Age less than 25 vs 25+	Never married vs married	Urban vs rural	Higher vs lower media exposure	Secondary school + vs primary/none
C.A.R.					
Males	0.97	0.94	0.99	0.99	1.01
Females	1.08	1.10	1.04	1.02	1.06
C TM te d'Ivoire					
Males	1.10	0.97	1.01	1.22	1.31
Females	1.39	1.11	0.97	1.27	1.42
Lesotho					
Males	1.13	0.90	0.84	1.14	1.06
Females	0.96	0.89	0.98	1.06	1.04
Mauritius					
Males	1.32	1.68	0.86	1.30	0.99
Females	0.91	1.32	0.98	0.52	0.82
Togo					
Males	1.03	1.08	1.07	1.29	1.22
Females	1.79	1.58	1.26	1.40	1.16

Discussion

The results presented in this paper raise both methodological and substantive issues. The central methodological issue is whether or not surveys of the general populations in Africa can provide trustworthy descriptions of risk behaviour, behavioural change and associated beliefs and attitudes. The verdict is of great practical importance, because it will greatly influence the design of an appropriate research strategy to measure progress towards curbing the spread of HIV. Once again an analogy from family planning is appropriate. Over the last 20 years, surveys have emerged as the predominant method of evaluating birth-control programs. Their role has been crucial in convincing government officials and politicians, both in developing countries and even more so in donor countries, that investment in family planning is worthwhile, and, to some extent at least, successful. The pressure will inevitably increase on WHO and donor agencies involved in AIDS-related activities to demonstrate some impact from the considerable sums that are being spent. Surveys of the general population have a crucially important potential role here. If survey data prove to be of low reliability, the prospects for useful evaluation of AIDS programs are severely diminished.

It is premature to reach a verdict on the credibility of survey data based on this preliminary analysis of the first five GPA-sponsored surveys in Africa. Both rigorous checks of internal consistency and field tests of response reliability are under way and GPA intends to sponsor several field tests in 1992. Meanwhile results from other major surveys will become available. Thus by the end of 1992 we should be in a much stronger position to assess the contribution and limitations of general population surveys to AIDS evaluation.

The substantive messages carried by these surveys, of course, depend on their credibility. Taken at face value the main results can be summarized as follows: high general awareness of HIV/AIDS (except in Togo); among those aware, almost universal knowledge of sexual transmission; a widespread perception of personal risk (except Mauritius); a weak relationship between risk perception and reported risk behaviour; a very high level of reported behavioural change in response to AIDS; the main reported change implying fewer, or more circumspect, sexual contacts; little mention of condoms in terms of either actual or intended use.

The evidence of widespread behavioural change should be regarded sceptically, because of the absence of corroboration. While there is supporting evidence of behavioural change in Kigali (Cara'l and Piot 1988; Linden et al. 1991), the more general impression is of the absence of change despite wide knowledge (e.g. Konde-Lule, Berkley and Downing 1989; Adamchak, Mbizvo and Tawanda 1990). Similarly there are reports of highly successful condom marketing schemes (Liskin, Wheston and Blackburn 1990), but these achievements have to be balanced by the major body of survey evidence, from both Demographic and Health Surveys and from GPA surveys, that use of condoms remains at a very low level. Moreover, HIV infection continues to rise and there is no evidence of a decline in other sexually-transmitted diseases.

We should also be sceptical of the impression given in surveys that the major behavioural response will take the form of a reduction in sexual partners rather than greater use of condoms. It may be difficult to make sex unpopular and the huge efforts, now under way, to promote condoms in Africa are likely to improve familiarity with, and decrease hostility towards, this device. A closer coordination between family-planning and AIDS programs may further facilitate these efforts. Particularly for single persons in high prevalence areas, barrier methods should receive top priority at family-planning clinics, rather than the lowest priority to which they are more typically assigned.

Several commentators have warned that African marriage systems may present particular obstacles to the acceptance of condoms (Bledsoe 1989; Larson 1989). As Bledsoe puts it, marriage is a process

that involves the testing of relationships. Condoms not only obstruct fertility but, in more subtle ways, introduce an element of distance and suspicion into the delicate process of marriage formation.

Despite this warning, the general message of the GPA surveys provides some grounds for optimism. Of course, survey respondents, as any other group, do not always do what they say. But the fact that so many report behavioural modification does at least suggest a readiness to do so. The prognosis would have been much worse, had these surveys shown widespread denial of risk and unwillingness to contemplate the need for change.

In another regard, these preliminary results have already proved useful. They suggest that model projections of the AIDS pandemic are based on very weak behavioural assumptions. It is to be hoped that those involved in the modelling of the disease will take account of this new evidence and thereby provide more credible insights into what the future might hold and what the likely effects of interventions will be.

References

- Abramson, P.R. and G. Herdt. 1990. The assessment of sexual practices relevant to the transmission of AIDS: a global perspective. *Journal of Sex Research* 27:215-232.
- Adamchak, D., M. Mbizvo and M. Tawanda. 1990. Male knowledge of and attitudes and practices towards AIDS in Zimbabwe. *AIDS* 4:245-250.
- Ajzen, I. and T.J. Madden. 1986. Prediction of goal-directed behaviour: attitudes, intentions, and perceived behavioural control. *Journal of Experimental Social Psychology* 22:453-474.
- Anderson, R.M., R.M. May, M.C. Boily, G.P. Garnett and J.T. Rowley. 1991. The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS. *Nature* 352:581-589.
- Anderson, R.M., T.W. Ng, M.C. Boily and R.M. May. 1989. The influence of different sexual-contact patterns between age classes on the predicted demographic impact of AIDS in developing countries. *Annals of the New York Academy of Sciences* 569:240-274.
- Ankrah, M.E. 1989. AIDS: methodological problems in studying its prevention and spread. *Social Science and Medicine* 3:265-276.
- Becker, M.H. ed. 1974. *The Health Belief Model and Personal Health Behaviour*. Thorofare, NJ: Charles B. Slack.
- Becker, M.H. and J.G. Joseph. 1988. AIDS and behavioural change to reduce risk: a review. *American Journal of Public Health* 78:394-410.
- Bledsoe, C. 1989. The cultural meaning of AIDS and condoms for stable heterosexual relations in Africa: recent evidence from the local print media. Paper presented at IUSSP Seminar on Population Policy in Sub-Saharan Africa: Drawing on International Experience, Kinshasa, Zaire, 27 February-2 March.
- Brandt, A.M. 1988. AIDS in historical perspective: four lessons from the history of sexually transmitted diseases. *American Journal of Public Health* 78:367-371.
- Cara'l, M., J. Cleland, J.C. Deheneffe and L. Adeokun. 1990. Research on sexual behaviour that transmits HIV: the GPA/WHO collaborative surveys. Preliminary findings. Paper presented at IUSSP Seminar on Anthropological Studies Relevant to the Sexual Transmission of HIV, Soderborg, Denmark, 19-22 November.
- Cara'l, M. and P. Piot. 1988. HIV infection in developing countries. *Journal of Biosocial Sciences* (Supplement, ed. A.G. Hill and D.P. Roberts) 10:35-50.
- Carballo, M., J. Cleland, M. Cara'l and G. Albrecht. 1989. A cross national study of patterns of sexual behaviour. *Journal of Sex Research* 26:287-299.
- Chin, J. 1991. Present and future dimensions of the HIV/AIDS pandemic. Paper presented at Seventh International Conference on AIDS, Florence, June.
- Fishbein, M. and I. Ajzen. 1975. *Belief, Attitude, Intention and Behavior*. Reading, MA: Addison-Welsey.

- Global Programme on AIDS. 1988. *Knowledge, Attitudes, Beliefs and Practices. Research Package, Phase 1 and 2*. Geneva: World Health Organization.
- Global Programme on AIDS. 1989. *Guidelines on Qualitative Approaches to Study KAP and Partner Relations in the Context of HIV/AIDS*. Geneva: World Health Organization.
- Konde-Lule, J., S. Berkley and R. Downing. 1989. Knowledge, attitudes and practices concerning AIDS in Ugandans. *AIDS* 3:513-519.
- Lapham, R.J. and W.P. Mauldin. 1972. National family planning programmes: review and evaluation. *Studies in Family Planning* 3,3:29-52.
- Larson, A. 1989. Social context of Human Immunodeficiency Virus transmission in Africa: historical and cultural bases of east and central African sexual relations. *Reviews of Infectious Diseases* 11,5:716-731.
- Linden, C., S. Allen, M. Cara'l, F. Nsenguremyi, P. van de Perre, A. Serufilira, J. Tice, D. Black, T. Coates and S. Hulley. 1991. Knowledge, attitudes and perceived risk of AIDS among urban Rwandan women: relationship to HIV infection and behaviour change. *AIDS* 5:993-1002.
- Liskin, L., C. Whetton and R. Blackburn. 1990. *Condoms - Now More than Ever*. Population Report H8. Baltimore: Johns Hopkins University.
- Muhondwa, E.P.Y. 1988. Impertinent interviews and lying respondents: a critique of the survey research method for heterosexual HIV transmission research in Africa. Mukimbili Medical Centre, Dar-es-Salaam, Abstract No 5902, Fourth International Conference on AIDS, Stockholm.
- Ngugi, E.N., J.N. Simonson, M. Bosire, A.R. Ronald, F.A. Plummer, D.W. Cameron, P. Waiyaki and J.U. Ndinya-Adola. 1988. Prevention of transmission of Human Immunodeficiency Virus in Africa: effectiveness of condom promotion and health education among prostitutes. *Lancet*:870-887.
- République du Togo. 1989. *Programme National de Lutte contre le SIDA: Plan à Moyen Terme*. Lomé and Geneva.
- Rotter, J. 1966. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs* 80.
- United Nations and World Health Organization. 1991. *The AIDS Epidemic and its Demographic Consequences*. New York: Sales No. E.91.xiii 5.
- Weinstein, N.D. 1987. Cross-hazard consistencies: conclusions about self-protective behaviour. Pp. 325-334 in *Taking Care: Understanding and Encouraging Self-Protective Behaviour*, ed. N.D. Weinstein. Cambridge: Cambridge University Press.

Old and new factors in health transitions



John C. Caldwell

Health Transition Centre, NCEPH, ANU, Canberra ACT 0200, Australia

Abstract

The introductory section of the paper notes that the health transition literature suggests a greater range of cultural, social and behavioural influences on health, especially child survival, than has attracted the attention of most social science researchers. They concentrate disproportionately on the impact of parental education, especially maternal education, perhaps because these are measures that are easily quantified and readily available in census and surveys. The major part of the paper discusses the implications of the finding by Preston and Haines that there is little evidence that child survival in the United States a century ago was much affected by mother's literacy, ethnicity or English-speaking ability. This review draws on that evidence to argue that Third World mortality has in contrast been reduced over recent decades by two imports: modern medical technology and the Western scientific attitude that induces a successful collaboration with the former. This attitude is largely a product of modern education and it is this symbiosis in reducing mortality between modern medical technology and the scientific outlook that explains why steep mortality declines in the contemporary Third World depend both upon providing an easily accessible modern health service (with a significant curative component) and the development of mass schooling (particularly for girls). It also explains the steep differentials in child survival by mother's education.

Introduction

I will start by referring to my article (Caldwell 1986), 'Routes to low mortality in poor countries', to identify themes in it with a view to showing what has and has not featured in recent debates. I shall concentrate on five themes in that article: the role in the mortality transition of (1) the provision of modern health services, (2) education, especially parental education, and above all maternal education, (3) the use of family planning, (4) the nature of the society in terms of such measures as women's independence and an egalitarian or radical tradition, and (5) the communist route to low mortality. Much of the recent discussion has been on the role of parental education, so it might be worthwhile glancing at the other themes, not necessarily in the listed order, to see why they have been ignored or rejected, before focusing on new evidence on parental education. Some of these preliminary considerations are necessary for the discussion of the role of education in the second part of the paper. It might be noted in passing that the concentration on education is at least partly explained by its easy quantification, at least as measured by duration, and hence its almost universal inclusion in censuses and surveys.

Most of this paper is devoted to issues regarding the interaction between social factors and health in both the West and the contemporary Third World which have been raised by the publication of new material on the American historical experience.

The role of modern health services

'Routes to low mortality...' placed a good deal of emphasis on breakthrough periods when much more of a country's population was covered by easily accessible health services. The most successful

services were simple, free or low-cost, and universal. Yet most comments on that article relate to its evidence of the impact on child mortality of parental education and largely ignore the references to a symbiosis between educational and social change on the one hand and increased modern medical provision on the other.

One reason is undoubtedly the appeal to the social scientist of explanations involving largely social causation. There is an even greater appeal to those medical scientists who have rebelled against technology and wish to demonstrate its impotence.

Yet the evidence that the health transition could have proceeded far without technology is slim. Two of the great health successes, where the shackles of income level seem to have been almost completely cast aside, have been Kerala and Sri Lanka. It is worthwhile examining the situation in those two societies in the early decades of the present century. Both were, by South Asian standards, individualistic societies even then, and in both, women had a great deal of independence. Indeed, in Kerala the raising of children and the care for their health was reported to have been almost entirely the concern of women, with men unable to play a role at least in the predominating matrilineal castes (Sushama 1990: 785; Mateer 1883: 209-210). In both societies there was great sensitivity to illness. In Kerala Ayurvedic medicine was practised and ill-health identified as needing immediate attention on a scale not paralleled anywhere else in India. The situation was similar in Sri Lanka, but, in addition, much energy and resources went into exorcism in the form of the 'devil dance'. At the 1901 censuses female age at marriage in these two societies was later than anywhere else in South Asia. Even then, educational levels in the two societies were rising, so that by 1921 Sri Lanka recorded the same proportion of females with schooling that Pakistan recorded half a century later.

These were the bases on which the later health revolutions were to be built. Yet at the beginning of the century Sri Lanka's life expectancy at birth was no more than 35 years, while Kerala's was probably as low as 26 years as late as 1916. There was continuing slow change but not at the rate that occurred once modern medical services became widely available. In Sri Lanka the improvement in mortality in the seven years following the Second World War was as great as in the previous half century. Certainly, the conquest of malaria played a role but was probably responsible for only part of the improvement, as is shown by the mortality decline in the south-west of the country which had been relatively free of the disease. The effective revolutions in modern medicine were the democratic ones which spread free or cheap services to the urban poor and the rural fastnesses, but it was largely curative services that were being so rapidly extended and they offered standard modern medical services.

If one goes back further into history, it is easy to argue that in the stationary or near-stationary populations which have characterized most of human history, neither social attitudes toward sickness and its treatment nor health practices played a very significant role in determining the mortality levels for whole societies. If we assume, as I think we must, that reproductive practices were largely independent of mortality levels, then clearly fertility determined mortality levels, ultimately through Malthusian checks. In a society characterized by the Princeton West family of model life tables a total fertility rate of six would have resulted in a life expectancy of 20 years and an infant mortality rate above 350, a total fertility rate of five would have allowed a life expectancy of 25 years and an infant mortality rate of 300, and a total fertility rate of four would have permitted a life expectancy of 32.5 years and an infant mortality rate below 250. It is, of course, probable that different societies would have distributed deaths by age and sex in different ways. It is also probable that the degree of maternal competence in providing child health care created differentials in the society in child survival. This is probably still the case but we lack objective measures of such competence and have to resort to measures of education, one input which can increase maternal effectiveness. It should also be conceded

that Notestein, in his original formulation of demographic transition theory, did regard the mortality level as the major demographic determinant with populations desperately contriving social institutions that would succeed in raising the birth rate to a level that would ensure that mortality did not decline (Notestein 1945: 39).

The communist route to low mortality

The only examples, in conditions where *per capita* incomes are small, of the rapid achievement of low mortality that can compete with the model offered by Sri Lanka and Kerala is that provided by various communist regimes: USSR before the Second World War, China after 1949, Cuba from 30 years ago, and Vietnam over the last 20 years. There is at present very little interest in how this was achieved.

There are probably two reasons. The first is that the disillusionment with regard to the extent of economic success has also led to a scepticism about the health achievements. The second has been the recent failure in Eastern Europe and the USSR to decrease mortality levels or even to maintain the levels already reached. Yet those systems earlier did remarkably well in attaining life expectancies of 65 or even 70 years.

Why they did so well is not absolutely clear. This is particularly brought out by China's experience during the 1950s. One ingredient was undoubtedly an emphasis on health access for all and on the provision of facilities. Yet those facilities had very little in the way of modern medicine to dispense. There are almost certainly social and behavioural aspects to the achievement. In spite of the command economy and political system, the revolution probably achieved not only greater egalitarianism and freedom for women and children but something of the spirit that goes along with these changes. It is also likely that the patriarchal family structure was dented in another way in that both neighbours and local party officials no longer felt any compunction in invading families' privacy and identifying sick children or women who needed treatment. It is possible that the changes secularized the society and gave support for scientific beliefs about the cause of illness and how it should be treated. There was an assumption that the revolution believed in modern science: a kind of compulsory Westernization.

This is a model that may not be employed again on any great scale. Yet there are obviously lessons to be learnt here for the health transition and some good analytical studies would be worthwhile.

Women's position, egalitarianism, family planning and the decline of family size

This is a heterogeneous area which needs much more research. It may all be the one big black box called 'individualism'. Changes in these regards may be encouraged by the development of a market economy, but clearly some societies already had characteristics long ago which would eventually lend themselves more than the characteristics of other societies to the achievement of low mortality. There are many interrelations and overlaps. Societies where women have greater independence are more likely to be egalitarian. Eventually, they are more likely to educate their girls. As fertility falls, children are more likely to have their health carefully looked after, but they are also more likely to achieve a certain independence which helps them both to look after themselves and to demand more care from others.

We have become so obsessed about the impact of parental care that we have tended to neglect the study of what persons can do for themselves. Yet the decline of excess female mortality in South Asia as girls age from 1–4 years to 5–9 years and then 10–14 years seems to be at least partly evidence of their growing control over their own fate (Caldwell and Caldwell 1990: 16). Work has begun on the control by adults of their own health destiny but it does not yet compare with what we know about children.

The characteristics of parents and the survival of their children

The major purpose of this paper is to reflect on important new evidence and its interpretation with regard to the determinants of child survival. That evidence comes from the recent publication of a book by Preston and Haines (1991a) *Fatal Years: Child Mortality in Late Nineteenth-Century America*.¹ Discussion of this book continues a debate recorded in Preston (1985), Caldwell (1986, 1990), and Ewbank and Preston (1990).

This book draws upon a sample of the returns for the 1900 US Population Census which asked questions about births to women and the survival of those children. These data, then, provide the opportunity for employing indirect methods of analysis for determining child survival and relating those survival levels to the characteristics of parents.

These calculations show that the level of child mortality in late nineteenth-century America was at a level found today in only a handful of poor countries, mostly in savannah West Africa, where life expectancy at birth is still under 50 years. Yet the authors argue that the real *per capita* income in America of that time was comparable to that now found in Hungary, Portugal, Romania, Argentina and Chile, all countries with life expectancies of 70 years or more (Preston and Haines 1991a:199).

Preston and Haines seek to answer two questions. The first is why such an income level in America of the 1890s could not be translated into better health. The second is why there were not the wide differentials in child survival by education, occupation or social class of parents that are found in the contemporary Third World. Their answer to these questions is by no means unambiguously set out. The authors are conspicuously reluctant to say, in so many words, that it has a great deal to do with biomedical advances.

They begin with the United States and conclude from multivariate analysis that the three most powerful determinants of child survival were, in order of importance, race, size of the place of residence, and region of residence. With regard to race, they say:

Being black in 1900 denoted a set of economic and social conditions that powerfully affected child mortality and that was not adequately captured by other variables that may have been associated with race (Preston and Haines 1991a:171).

It might be noted that less emphasis is placed here upon culture, possibly justifiably, than Preston and colleagues implied when examining differentials by ethnicity in other parts of the world in the United Nations (1985) monograph, *Socioeconomic Differentials in Child Mortality in Developing Countries*. With regard to place size, it is clear that cities were still the most unhealthy places, although that situation was already beginning to change in New York. Region of residence, controlled for state income, is described as an environmental effect. Of lesser importance, but not negligible, was whether the husband was employed or not, the level of income of the state of residence, and the presence or absence of boarders in the house. They note:

Economic discrepancies are clearly associated with a large variation in child mortality although no single economic variable shows striking effects by itself. In addition, the

¹ A shorter and perhaps more cautious treatment of this book was prepared after the conference and was published as Caldwell (1991). This immediate reaction to the book is presented here, partly because it is of greater length but partly also because in its incautious immediate reaction it raised issues in somewhat ambiguous areas that may be worth further exploration. Both treatments regard the book as a major contribution to our understanding of health transition in the West. To that addition to the debate and discussion by four other contributors to the *Health Transition Review* Forum, Preston and Haines (1991b) replied.

large mortality variation by race ... is most plausibly ascribed to the enormous economic disparities that existed between the races at the time (Preston and Haines 1991a:175).

The authors were clearly nonplussed that the expected care or behavioural variables did not show up, at least as would have been indicated by mother's literacy (in the absence of educational data), ethnicity, English-speaking ability, or husband's occupation. They decided 'Whatever behavioral variation was associated with these variables seems to be swamped in its effects by broad geographic and economic factors' (Preston and Haines 1991a:175). They went on:

This list paints individuals as relatively passive victims of time, place and labor markets. Behavioral factors were probably a key to child mortality declines during the twentieth century (Ewbank and Preston 1990), but their mark was not highly visible at the turn of the century in variables such as literacy and occupation where we might expect their effects to be most obvious. One reason may be that too little knowledge of specific ways to enhance child survival had developed to allow individuals to escape from the circumstances imposed by broad geographic and economic forces (Preston and Haines 1991a:175–76).

As we will come to see, there are problems about what the authors mean by the word 'knowledge'.

They first sought help from a comparative study of the published tables from the 1911 Census of England and Wales which, of course, referred to child mortality occurring eleven years later than the US deaths they had examined, probably a significant difference in occurrence in the early twentieth century. Clearly, by this time there may have been advances in the medical help available which the richer and more educated could largely monopolize. Nevertheless, the authors were undoubtedly right to be impressed that the occupational, social-class and economic differentials (for instance, number of rooms in the house or whether the wife had to go out to work) were so steep both in absolute terms and in comparison with America. They discarded the explanation that either society was substantially richer than the other, as this was apparently not the case, but they worried longer that occupational identity might be broader and less specific in the United States. In the end, they largely supported the view that the classes were more distinctive in England and that there were greater differentials in income and probably access to treatment between them (Preston and Haines 1991a:125ff). In a somewhat indirect parallel, they imply that the situation is closer to that of the division by race rather than occupation in the United States (Preston and Haines 1991a:198).

They then go on to compare the American situation with that of eleven developing countries (four in Africa, five in Asia, and two in Latin America) during the 1970s. Two striking contrasts are brought out, one with relation to the survivorship of children by their mothers' literacy and the other by urban-rural residence. In fact, the proxies for literacy in the contemporary Third World are years of schooling, but the fact that different durations in quite different types of schooling yield similar survival differentials indicates that there is little problem in estimating exactly when literacy is achieved. There may, as I shall argue below, be some problem with equating schooling with literacy.

The findings are clear. First, there is a much steeper differential in child survival by maternal education in the contemporary Third World than there was in America at the end of the nineteenth century. This is not a characteristic of only some parts of the Third World, because similar patterns were found in each continent, the differences within continents often being more striking than those between them. Secondly, the effect of urban residence has dramatically reversed. In America of almost a century ago child mortality was much higher in the cities and towns, while in the contemporary Third World the relativity is much the same but in the opposite direction.

How do Preston and Haines explain these findings?

With regard to America and Britain at the beginning of the century they argue that the key to understanding why the privileged sections of the population could not buy much greater health for their children than could the poor was the fact that 'they shared a...rather primitive...base of health knowledge' (Preston and Haines 1991a:198). Clearly, they mean something broader by 'base of health knowledge' than the curative and preventive medical services available to the people. They include the care which can be bestowed within the household and such protective measures as boiling milk or water or isolating children from infection, all actions likely to be taken most readily by persons aware of the germ theory of infection and so convinced of it that it has an impact on much of their behaviour.

The authors' reluctance to speak of the efficacy of medical services and the extent to which they were available across the community, together with the repeated employment of the word 'knowledge' without making it clear whether this is individual knowledge or the scientific knowledge available to the community or world, has the effect of tipping our interpretation towards accepting the primacy of household care as the explanation of child survival. They argue 'By modern standards, ignorance about both preventive and curative health care was widespread in the United States at the turn of the century' (Preston and Haines 1991a:200). Again,

Schooling effects should reflect only the accretion of material resources that result from the increased earnings opportunities; the additional portion of the effect that would reflect closer connection to good health knowledge and practice among the well-educated should be largely inoperative in a situation where education 'buys'... little knowledge (Preston and Haines 1991a:200).

The knowledge was apparently there among some but not the good sense about its use or the will to employ it in the broader community, and it appeared to be a message that would affect behaviour rather than technologies built upon scientific knowledge.

The most enlightened public health officials ... saw clearly the implications of the germ theory for preventive health care, but they despaired at the difficulties of getting the word across to physicians, let alone parents (Preston and Haines 1991a:200).

Elsewhere there is some reference to the fact that the basic problem may have been building up a body of proven knowledge and techniques and a system for applying them. There are references to the 'Level of technical knowledge about health' (Preston and Haines 1991a:199) and the fact that 'the United States in this period simply did not know how to effect this conversion' (Preston and Haines 1991a:199) of literacy and income into higher levels of life expectancy.

My interpretation of their evidence is the following. The small differentials in child survival in late nineteenth-century America by father's occupation or mother's literacy suggest that the level of home care and health maintenance, over and above what one's income would allow and one's social class dictate, was not in fact of major importance. This is somewhat surprising. It may mean that the middle class mother behaved as the middle class mother should, or as her husband or her church would expect, almost regardless of educational level. Being careful and clean was not as demanding on one's educational resources, as would prove to be the case in the challenge yet to come of deciding just when to use the more effective medical system and how to collaborate most effectively with it. There appears to be evidence here that later improvements in child mortality may have had much more to do with how caring parents collaborated with a health system that increasingly had something to offer than with the pure impact of that care. The economic evidence is impressive and does suggest that the slow improvement in mortality in the nineteenth century was the product of the greater wealth yielded by the

industrial revolution (and not just better nutrition) even if this was partly offset by the attack on the health by the industries themselves. The most fortunate were those whose heightened living standards were made possible by the industrial revolution but who themselves lived far from the cities and factories.

I believe that the analysis employs 'knowledge' in various senses, and, more seriously, that it underestimates just what a proven repository of scientific knowledge and technology is. Knowing that infection was caused by germs was not the same thing as knowing that there were germs in milk, that they existed there in sufficient densities to endanger a child's health, that they could be destroyed, at least temporarily, by boiling, and that the boiling did not have other harmful consequences for the value of the milk. True believers were carrying out experiments of this kind in New York City in the 1890s, but it was to be a generation before even all capable health scientists of good will came to a consensus on the matter.

The health science armoury that we now possess was created only slowly after the discovery of germs. Much of this was, of course, a series of biomedical breakthroughs as scientific knowledge was built up about the nature of these micro-organisms, how they infected human beings, how the human body reacted, and how this knowledge could be employed to devise means for combating the infections. The whole process was assisted by the development of research methodology and investment in the huge programs and institutions required. The building of a scientific health system depended also on other developments, equally made possible by the expanding wealth of the industrialized world. They included experimentation with community programs like boiling milk, the building of modern medical schools, and the training of doctors and other health personnel both rigorously and from properly tested experience. It is this last kind of authority which is important. It was something codified and believable and that was not merely provided by the occasional exhortations of public health personnel. Even their authority was not always right, but it was more often right than wrong. Ultimately, the availability of this health capacity, rather than knowledge, would depend also on the reduction of grinding poverty and the provision of health insurance and national health systems.

The small differential in child survival by parental education or occupation in late nineteenth century America was not evidence that the better-off or educated would not listen or learn or would not buy the best treatment for their children. It was just that there was little in the effective scientific health care that money could buy. A study of periodicals in a rather similar society, Australia, in the second half of the nineteenth century, revealed that the educated and prosperous section of society was, if anything, more earnest about their children's survival, more willing to pay for what was available to improve health or cure illness, and as oriented towards science and knowledge as are today's parents (P. Caldwell 1971). There was little in the way of anti-scientific movements or a disregard for science, and there was little belief in alternative therapies.

Preston and Haines may be saying all this, but it is not clear. They seem to be placing an emphasis on personal knowledge rather than the body of scientific knowledge and expertise and on individual thoughtless behaviour so as to support the argument that a considerable part of the decline of mortality in the twentieth century would have occurred even without the continued development of scientific medicine. However, a paper which is essentially a sequel to this book (Ewbank and Preston 1990) does argue that there was an all-important interaction between the enlarged health science capacity of society and the increasingly educated health consumers during the period 1900-30, and that the more educated were increasingly able to grasp these new opportunities to a greater extent than the less educated and poor. Even here, there is a curious tendency to refer to the non-behavioural side of that relationship as 'public health efforts' (Preston and Haines 1991a:143) rather than the totality of the scientific health system.

The contemporary Third World

Turning to the Third World, Preston and Haines put forward the explanation that the large differentials in child survival by maternal education are entirely a product of improved health behaviour in proportion to the extent of education:

In developing countries today, the mother's education or literacy appears to retain such a high degree of explanatory power because it is associated with such health behaviors as vaccination of children, maintenance of hygienic conditions in the home, and receipt of professional health care for maternity and for child illness (Preston and Haines 1991a:206).

I have two problems with this.

The first may be only a matter of emphasis. The list of effective interventions places curative services last and, in fact, at no stage employs that terminology. Our research in Sri Lanka (Caldwell et al. 1989) has convinced me that the low infant and child mortality in Sri Lanka is to a very large extent the result of the fact that the universally acceptable modern health services prevent most sick children from dying. The situation was not quite so clear in our research in South India (Caldwell, Reddy and Caldwell 1983; Caldwell, Caldwell et al. 1990) and Nigeria (Orubuloye and Caldwell 1975), but, nevertheless, the evidence certainly supported the importance of these interventions.

The second point depends very much on the evidence that Preston and his colleagues have brought forward. The differential in child survival by the level of maternal education (or the level of parental education) is much greater than in America (or the West more generally) not only in the late nineteenth century, but also in the period 1900-30, or at any other time. I have argued in a recent article (Caldwell 1990) that this is not a reflection of the general levels of either *per capita* incomes or health services in the Third World but is related to the nature of the educational experience. It was clear from the finding that maternal education improved child survivorship at even very short durations of education, that it was perhaps the experience or fact of schooling rather than its content or what was learnt that made the difference. Cleland (1990: 411-412) has explored this matter further. It might be noted here that it is this finding that makes it hazardous to equate literacy with schooling.

In the research in South India we found a multiplicity of reactions between the mothers of sick children and the health system which explained why educated mothers obtained more help from that system. They were more likely to take children for treatment at the modern health centre and they were likely to do so with less delay; they spent a longer period with the doctor because they were willing and able to describe symptoms more fully and because the doctor listened to them at greater length; they followed the prescribed treatment more exactly and were more likely to persist with it; and they were much more likely to report back, and to do so soon, if the child's condition was not improving. Their greater willingness to employ modern medicine, and to lay out a larger proportion of family income for it, arose because their schooling made them feel that they had a responsibility for taking an initiative, because they identified their school with other non-traditional institutions of the modern world and believed that they could and should use them, and because they were more likely to believe that the doctor and the health centre embodied a scientific truth about the world. The illiterates were more likely to belong to an older world, some believing more strongly in the Ayurvedic system, but most believing in spirit causation, inevitable punishments from previous errors or sins or just ill-fatedness. They believed less in scientific truth, more in alternative explanations, and more in the inevitability of things. Their failure to report symptoms adequately and also to report back the failure of treatment has deep roots in the Ayurvedic belief that the healer immediately recognizes the ailment and always prescribes the best cure. If that is insufficient, there is little more that can be done.

I have argued elsewhere (Caldwell 1980) that modern education is essentially a Westernization of the world. It embodies the Western experience and outlook. The Western outlook in the late nineteenth century was essentially scientific, partly because of prior economic development. Even the illiterate and poor believed in the efficacy of science and of scientific medicine, if indeed it could be found. The experience of education was essentially an intensification of their own system. The only cultural change was a tendency to teach the working class middle-class values.

Schooling systems develop slowly and with difficulty, and no one has bothered to invent them for a second time. The truth is, of course, even more complex than this. Schooling is largely a way of allowing people to communicate with and adapt themselves to the modern world, largely shaped by the West. It is a deculturating experience in the Third World in the sense that schooling is not in the West. It quite purposively draws people into the Western system and assumes from the day they set foot in school that they have agreed to that system. Even where most people believe in witchcraft, children, while in school, must not echo that belief (cf. Greenfield and Bruner 1966:84) and polygynous school teachers must not even defend polygyny in class (Caldwell 1980).

Rapid mortality decline in the Third World depends on access to both modern curative and preventive medicine and the fullest possible collaboration with these systems in both belief and action. The important role of modern health services is attested by the higher levels of child survival in urban areas. The collaboration is induced by the schooling system. Thus, developing countries import both their medical systems and the willingness to make maximum use of them. This is the explanation for the steep parental educational differentials in child survival. Educational revolution in the West of a hundred years ago would not have produced the same effects for two reasons: (1) the scientific medical system had little curative medicine to offer; (2) education would not have produced the same change in beliefs and behaviour that it does in non-Western societies.

A final note

The failure of parental educational or class differences to show up very clearly in American child survival a century ago appears to throw a broader doubt on the importance of household child care except in decisions about the use of medical services. This seems to fly in the face of commonsense and experience. The explanation may be that suggested by Preston and Haines that such influences were overwhelmed by economic ones, but, even if this is true, it suggests they are smaller than most of us suspected. It may be that we have here evidence that economic and social mobility in America at that time produced a fairly classless society and that the literacy-illiteracy divide is not subtle enough and is subject to misreporting at the census. The importance of class in Britain suggests its dominant importance there, and that data on mothers' education may have made little difference in that women were largely prisoners of their class and behaved towards their children as their social peers expected them to do.

Acknowledgements

My first acknowledgement is, of course, to Samuel Preston, and his various colleagues, Michael Haines, Douglas Ewbank and others for providing both the stimulus and much of the data for this discourse. My second is to Pat Caldwell and Wendy Cosford for assistance, and to Pat Goodall for word processing.

References

- Caldwell, John C. 1980. Mass education as a determinant of the timing of fertility decline. *Population and Development Review* 6,2:225–255.

- Caldwell, John C. 1986. Routes to low mortality in poor countries. *Population and Development Review* 12,2:171–220.
- Caldwell, John C. 1990. Cultural and social factors influencing mortality levels in developing countries. *Annals of the American Academy of Political and Social Science* 510:44–59.
- Caldwell, John C. 1991. Major new evidence on health transition and its interpretation, in Forum: *Fatal Years, Health Transition Review* 1,2:221–229.
- Caldwell, John C., Pat Caldwell, I. Gajanayake, I.O. Orubuloye, I. Pieris and P.H. Reddy. 1990. Cultural, social and behavioural determinants of health and their mechanisms: a report on related health programs. Pp. 534–541 in *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health*, ed. John C. Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid and Daphne Broers-Freeman. Canberra: Health Transition Centre, The Australian National University.
- Caldwell, John C., Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid and Daphne Broers-Freeman. eds. 1990. *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health*. Canberra: Health Transition Centre, The Australian National University.
- Caldwell, John C., I. Gajanayake, Pat Caldwell and I. Pieris. 1989. Sensitization to illness and the risk of death: an explanation for Sri Lanka's approach to good health for all. *Social Science and Medicine* 28,4:365–379.
- Caldwell, John C., P.H. Reddy and Pat Caldwell. 1983. The social component of mortality decline: an investigation in south India employing alternative methodologies. *Population Studies* 37,2:185–205.
- Caldwell, Pat. 1971. File on Australian newspapers and periodicals. Canberra: Department of Demography, The Australian National University.
- Caldwell, Pat and John C. Caldwell. 1990. Gender implications for survival in South Asia. Health Transition Working Paper 7. Canberra: Health Transition Centre, The Australian National University.
- Cleland, John. 1990. Maternal education and child survival: further evidence and explanations. Pp 400–419 in *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health*, ed. John C. Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid and Daphne Broers-Freeman. Canberra: Health Transition Centre, The Australian National University.
- Ewbank, Douglas C. and Samuel H. Preston. 1990. Personal health behaviour and the decline in infant and child mortality: the United States 1900–1930. Pp. 116–149 in *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health*, ed. John C. Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid and Daphne Broers-Freeman. Canberra: Health Transition Centre, The Australian National University.
- Greenfield, P.M. and J.S. Bruner. 1966. Culture and cognitive growth. *International Journal of Psychology* 1,2:89–107.
- Mateer, S. 1883. *Native Life in Travancore*. London: W.H. Allen.
- Notestein, F.W. 1945. Population: the long view. Pp. 36–57 in *Food for the World*, ed. T.W. Schultz. Chicago: University of Chicago Press.
- Orubuloye, I.O. and John C. Caldwell. 1975. The impact of public health services on mortality: a study of mortality differentials in a rural area in Nigeria. *Population Studies* 29,2:259–272.
- Preston, Samuel H. 1985. Resources, knowledge and child mortality: a comparison of the U.S. in the late nineteenth century and developing countries today. Pp. 373–386 in International Union for the Scientific Study of Population, *International Population Conference, Florence 1985*, Vol. 4. Liège: Ordina.
- Preston, Samuel H. and Michael R. Haines. 1991a. *Fatal Years: Child Mortality in Late Nineteenth-Century America*. Princeton: Princeton University Press.
- Preston, Samuel H. and Michael R. Haines. 1991b. Response to comments on *Fatal Years*, in Forum: *Fatal Years, Health Transition Review* 1,2:240–244.
- Schultz, T.W. ed. 1945. *Food for the World*. Chicago: University of Chicago Press.

- Sushama, P.N. 1990. Social context of health behaviour in Kerala. Pp. 777–787 in *What We Know about Health Transition: The Cultural, Social and Behavioural Determinants of Health*, ed. John C. Caldwell, Sally Findley, Pat Caldwell, Gigi Santow, Wendy Cosford, Jennifer Braid and Daphne Broers-Freeman. Canberra: Health Transition Centre, The Australian National University.
- United Nations. 1985. Socioeconomic differentials in child mortality in developing countries (by B. Mensch, H. Lentzner and S.H. Preston), ST/ESA/SERA/97. New York: Department of International Economic and Social Affairs, United Nations.