Genealogies as a Source for Demographic and Genetic Studies: some estimates of bias.

Jim Oeppen[*][†]

University of Southern Denmark

Work in Progress: please do not cite without the author's consent.

1 Introduction

In 1975, a meeting of the Societé de Démographie Historique explored the theme 'Que sont les généalogies pour les généalogistes? Que devraient–elles être pour les historiens démographes et les démographes?'[1] Despite the valuable efforts of the participants at the meeting to answer the two questions, and the subsequent work of these and other authors, genealogies remain a source of both promise and disappointment for demographic researchers. In these notes, I will take the potential uses of genealogies as given, and try to draw together some ideas on the difficulties that genealogies can present for demographic analysis. At the same time, I shall try to go beyond speculation on some topics by attempting to measure the bias for estimates of fertility and mortality in some genealogies, even when the data are perfectly recorded. I hope that the other participants in this workshop will fill the gaps in my knowledge, and that they can relate these results to the particular requirements of studies of genetics and longevity.

Demographers have always been worried about genealogies. They suspect that they are biased towards high levels of nuptiality, survival and fertility, and that the incidence of missing data is not random. These two issues apply particularly to inheritance chains. In Britain, genealogies have largely been ignored, with the notable exception of Hollingsworth's study of the British Peerage, but this was specifically, and skilfully, constructed for demographic analysis.[2] In France and the United States there seems to have been a greater willingness to engage with genealogists and their data.[3] The most notable integrations of historical demography, genealogy, and genetics have taken place in Utah and Quebec.[4]

An informal way to approach the problem of bias is to ask the question: 'How would I design the demography of a family to ensure the male succession in the direct line?' One answer might be for everybody to survive to at least 50, everybody to marry, and at a young age, and have large numbers of sons. The holder of the title should die as quickly as possible after 50 to maximise the probability

---

[1]See Annales de Démographie Historique (1976, pp. 73–170)

[2]Hollingsworth, T.H. (1977) 'Mortality in the British Peerage Families since 1600.' *Population,* XXXII, 323–52.

[3]Dupaquier, J.. 'Généalogie et démographie historique.' *Annales de Démographie Historique,* 1993, 391–5. Fogel, R.W. (1993) 'New sources and new techniques for the study of secular trends in nutritional status, health, mortality, and the process of aging.' *Historical Methods,* 26, 1, pp. 5–43.

[4]Bean, L.L. 'The Utah population database: demographic and genetic convergence and divergence.' Ch. 15 in Adams, J. *et al.* (Eds. 1990) *Convergent Issues in Genetics and Demography.* Oxford. Bouchard, G. 'Population and genetics: a new frontier for the social sciences.' *Annales de Démographie Historique,* 1993, pp. 397–412.

that one of the sons is alive to continue the chain. Going beyond this kind of speculation is rare in the literature and will need some simplification.

2 A Classification of Genealogies

When we examine particular genealogies it is clear that there are various construction methods. For the purposes of this paper, I propose to distinguish two dimensions on which they can be classified. The observations that follow can then refer to these 'pure' forms without getting bogged down in particular variants. I shall use the term 'ego' to refer to any selected or 'marker' individual in a genealogy, and 'population' as a shorthand term for either the whole population or an unbiased, homogeneous sub–set.

The first dimension is **Ascendant** versus **Descendant,** where an ascendant genealogy is concerned with the ancestors of ego (often called a pedigree), while in the latter, ego is the founder or patriarch.[5] Both are defined here to include lateral kin. For example, in my definition of an ascendant genealogy, father's brother is included, but not father's brother's son. Similarly, in my notion of a descendant genealogy, son's wife is included, but not son's wife's father.

The second dimension distinguishes between a **Lineage** and a **Succession.** I will define the former as a genealogy in which all individuals are regarded as equal, except for the ego who starts the tree in a descendant genealogy, or ends an ascendant one. The lineage emphasises continuity within the close kin, without emphasis on a particular individual. In a succession, the label 'ego', or perhaps 'headship', is passed between individuals in successive generations according to a set of inheritance rules conditional on some aspects of survival, sex, birth order, and kin relationship. All four cells in the cross–classification of these two dimensions are feasible.

Further distinctions can be made. Some genealogies have been constructed for multiple egos or successions, rather than a single lineage or title, but the rest of the paper should be equally applicable to these. Fogel, in his own classification of genealogies, identifies an important group that cuts across the one used here.[6] These he called **Family Histories,** as they reflect the interests of family historians.

Typically, an ego constructs an ascendant lineage for themselves as far back in time as they are able, and then constructs a descendant lineage from the earliest identifiable patriarchs. Fogel (p.30) seems to regard Family Histories as unbiased with respect to mortality if the compiler has sought to fill in the entire descendant lineage. I shall return to the validity of this view in a later section.

Another important variant of genealogy is **Family Reconstitution,** invented by Hannes Hyrenius and first published in 1942, before being re–invented and brought to wider attention by Louis Henry in the 1950s.[7] The primary aim is demography, rather than lineages. In the absence of censuses or population registers to define the at–risk population, event histories are constructed and used to define spells of exposure and estimate rates. A body of analytic methods has been carefully defined to avoid demographic biases, and take explicit note of truncation, censoring, and competing risks. In practice, event–histories are constructed around marriages, which form the primary unit of data organisation.[8] However, since individuals are births in one marriage generation, and spouses in the next, there is an explicit inter–generational linkage. Although the quantity of data analysed in this way has been large, the problem for genetic studies is that reconstitutions are usually constructed for small geographical units such as parishes or towns, where migration means that many of the lineages are incomplete.[9] There are also potential problems for longevity studies because, for those without a recorded death, there are few demographic events after age 50 to indicate that the person is still part of the risk–set

---

[5]See Jette, R. and Charbonneau, H. 'Généalogies déscendantes et analyse démographique.' *Annales de Démographie Historique,* 1984, pp. 45–54, and other papers in this volume.

[6]*Op. Cit.* Fn. 3

[7]See Terrise, M. 'Aux origines de la méthode de réconstitution des families: les Suédois d'Estonie', *Population,* 1975, numéro special, 143–155. Fleury, M. and Henry, L. (1985) *Nouveau Manuel de Dépouillement et d'Exploitation de l'Etat Civil Ancien.* Paris.

[8]In cases where there must have been a marriage, but it is unknown, a dummy marriage record is created to hold the data.

[9]For other causes of linkage failure, see the Record Linkage section below.

for mortality and has not yet migrated.[10]

The advantage of these data is that they are constructed under a set of rules, whose character is well understood, and can potentially cover the mass of the population of both sexes, all ages, and social status. They also cover wide spans of time, since the fact that lineages are broken is irrelevant to the demographic purpose. However, it is true that there is in some sense a 'reconstitutable minority' of families. Individuals who stay in one place, marry, and have lots of children are much more visible in the historical record. Finally, both genealogies and family reconstitution, with its emphasis on defining marital exposure, tend to ignore illegitimate or abandoned children, and the unmarried, yet these formed a significant part of historic Europe's demography. Similar arguments apply to the practice of ignoring childless women and all people under a certain age in Chinese genealogies.[11]

The most extensive usage of family reconstitution for genetic studies has been made in the Province of Quebec in Canada, where all the parishes have been reconstituted, thus avoiding the migration problem to a large degree.[12] The extent of the shared fertility experience between generations in English reconstitution data has been studied by Langford and Wilson.[13]

All genealogies, whatever their character, require the skills of record linkage.

2.1 Record Linkage.

Genealogy, and most research in historical demography, is based on some process of nominal record linkage. In practice, the classification for genealogies defined above is difficult to apply as it is often hard to establish what rules were used by the genealogist, let alone the consistency of their application. Unfortunately, there seems to be no tradition of explicitly stating these rules in the manner established by Louis Henry for family reconstitution and adopted by others.[14] The same observation applies to the 'pruning' of branches in a genealogy, which is often employed to prevent the task of construction becoming too large. To give an example, the ascendant succession for the British Peerage constructed by Bloore has a sex–ratio of 50 males per 100 females in one child families.[15] It seems likely that the genealogy has been implicitly pruned at the edges to exclude siblings of women who marry into the peerage, although their parents are recorded. In such cases it is impossible to make the important distinction between no data, missing data, and pruning. In any genealogy, the making of a link between individuals has some probability of failure. Perhaps the record is missing, or there is a slight variation in the spelling of a name, the person has left the scope of the available data sources, or an unambiguous link may not be possible because there are 'competing' individuals with the same names. The demographic biases that such problems may cause have been the subject of much research in family reconstitution studies, and most of this should be applicable to genealogies. More biases arise if we use records that are associated with sex, age, wealth, and power – such as land transactions, wills, and legal records.

One subtle aspect of linkage failure in an ascendant lineage concerns temporal coverage. By their very nature, they are more likely to fail the further one goes back in time. Since each link has a probability of failure, an analysis of say, sixteenth century demography using ascendant genealogies rooted in the present, will be biased towards those families that had the fewest links. Effectively, families with long generation lengths will be over–represented at early dates.

2.2 Missing and Incomplete Data.

Missing data is the Catch–22 of empirical research. If we knew exactly what was missing, it wouldn't be missing, so we can never know exactly what's missing. In addition to the conventional criticism of

---

[10]See Gill, R. (1997) 'Non–parametric estimation under censoring and passive registration.' *Statistica Neerlandica* 51, 1, pp. 35–54.

[11]Zhao, Z. 'Demographic conditions and multi–generational households in Chinese history. Results from genealogical research and microsimulation.' *Population Studies,* 1994, 48, 413–425.

[12]*Op. Cit.* Fn. 4

[13]Langford, C. and Wilson, C. (1985) 'Is there a connection between a woman's fecundity and that of her mother?' *J. Biosocial Sci.* 17, pp. 437–443.

[14]Wrigley, E.A. and Schofield, R.S. 'Nominal record linkage by computer and the logic of family reconstitution.' in Wrigley, E.A. (Ed. 1973) *Identifying people in the past.*

[15]Bloore, J. compiler of a CD of the British Peerage (S&N Genealogy Supplies, Salisbury, England.)

sources that underlies historical analysis, we are fortunate in historical demography to have a number of biological patterns that seem to be fairly universal, at least within populations of European origin. Age–specific patterns of fertility in non–contracepting populations practising breast feeding, and entry–sterility, are two examples. Sex–ratios by parity and completed family size can be useful indicators of the genealogical obsession with males, but of course they do not tell us the level of incompleteness , only if it is biased.[16] Excess male mortality of infants is a particularly useful diagnostic in successions. Twinning rates and the sex-ratios of twins can be used to check the plausibility of baptism registration. Historical demographers are less enthusiastic today than they were two decades ago about the age–patterns of mortality by sex, but they still provide a useful means of detecting missing infant deaths.[17]

It seems that genealogies could be constructed for genetic and demographic purposes with a more precise definition of missing events and links, to distinguish between true blanks, missing data, and pruning points. Rapid developments over the past decade in event–history analysis designed to deal with truncation, and passive, double, and interval–censoring, mean that both genealogy and traditional family reconstitution are producing data–sets that lag behind the capabilities of the statistical analyst.[18]

2.3 Bias in perfectly recorded genealogies.

There has been much speculation on demographic bias and its causes in genealogies, often conflating issues of representativeness and missing data with those of bias. I shall define bias, perhaps inadequately, as a systematic difference between what is actually measured and what one hopes to measure in a data–set. In the next two sections I am going to concentrate on bias, assume that the genealogies I am describing are perfectly recorded, and ignore the question of whether particular genealogies are representative of a wider population.

Examining the four cells in my cross–classification, I regard descendant lineages as potentially unbiased because the demographic processes are unconditional and therefore should reflect the 'population' process. This is not quite true of all the individuals, and some care must be taken in analysis, because spouses, in–laws, and adopted persons are not at risk from birth, but are left–truncated by the date at which they join the lineage.

Descendant successions are slightly more complicated. Where the succession of the label 'ego' proceeds from father to son, or grandson, I would argue that they are like descendant lineages and therefore unbiased. The problem comes when the succession fails in the direct line and the rules of inheritance are invoked to 'jump' to the next ego, without all the intervening kin being recorded. This person joins the lineage, conditioned on their surviving to the death of the previous ego and meeting the other inheritance rules, so they should be regarded as left–truncated by this date for the purposes of analysis. With this proviso, which has already been invoked for spouses, in–laws, and adoptees, I believe we can think of a descendant succession as one or more descendant lineages and potentially unbiased.

Ascendant lineages and successions must be biased towards high levels of nuptiality, survival, and fertility by definition: the issue here is the degree of bias, not the fact. At each backward link, a mother and father have to survive and reproduce. In a succession, an ego in one generation 'competes' in terms of birth order, sex, kin relationship, and survival to replace an ego in the previous one, adding further conditions. The question becomes: can any of the data be used without encountering bias?[19] For an ascendant lineage, we could ignore parents in the direct line and analyse the rest, although it may be worth checking this with the simulation methods described in the next section. In an ascendant male succession, it may be sufficient to ignore all males as well as the mother of ego, leaving wives (after allowing for left truncation) and sisters to be analysed.

---

[16]See Wrigley, E.A. *et al.* (1997) *English Population History from Family Reconstitution 1580–1837.* Cambridge. Fig. 7.6, p. 453, and Fig. 7.2, p. 395.

[17]See Wrigley, E.A. *et al.* (1997) Fig. 6.5, p. 232.

[18]Hollingsworth's prescient comments on this topic pre–date the statistical developments. 'Genealogy and historical demography.' *Annales de Démographie Historique,* 1976, 167–70.

[19]An interesting attempt to define what unbiased data can be used in a specially constructed ascendant genealogical sample has been described by Bideau, A. and Poulain, M. 'De la généalogie á la démographie historique: généalogie ascendante et analyse démographique.' *Annales de Démographie Historique,* 1984, 55–69.

2.4 Estimating the demographic bias in genealogies.

Estimating the bias in genealogies remains a major problem, and the idiosyncrasies of particular examples makes it unlikely that we can establish formal methods of bias assessment in all areas. However, there is at least one special case where we can measure the bias in fertility.

As an illustration, suppose that all women marry and that there are only two family sizes: two children and four children. In a descending lineage, women will have an average of three children each, reflecting the population mean. In an ascending lineage, each generation is linked to a pair of individuals selected from the previous one. A randomly selected individual is twice as likely to come from the larger family as the smaller one, creating a bias towards large sibling sets even though both family sizes are equally likely in the population. Obviously, the bigger the variance in the family size distribution, the bigger the gap between the fertility of the ascendant lineage and the total population. If it is true that many elite lineages are characterised by high levels of celibacy and sterility, then there are large numbers of zero–parity women, the variance is increased, and the problem is exacerbated.

Usually we don't have the parameters for the population at large, but if we do then we can assess the likely scale of the problem. In this case, we can calculate the extent to which its fertility will exceed that of the general population from the relation:

$$\bar{C} = \bar{X} + \frac{\sigma^2}{\bar{X}} \tag{1}$$

where for the general population: $\bar{C}$ is the average size of the sibling–set for a random individual, $\bar{X}$ is the average number of children per woman and σ^2 is the variance of the distribution of family sizes. This formula for the difference between the average number of children as seen by mothers, $\bar{X}$, and the average size, $\bar{C}$, of the sibling–set for a randomly chosen individual has been known for over a century.[20] It exposes the common fallacy of assuming that the family of orientation, $\bar{C}$, is equal to $\bar{X}$, the family of reproduction. Here, I am arguing that when fertility in the population, or descendant lineage is $\bar{X}$, that of an ascendant lineage will be $\bar{C}$.

To illustrate the scale of the bias, I have taken the Completed Family Size distribution for the period 1580 to 1837 in early modern England and assumed 10% of women never married.[21] Under these conditions and using the formula above, the average number of children per mother who survived to 50 in an ascendant lineage would be 7.1 children, compared with the population, or descendant lineage, value of 4.5 children, a bias of 58%.[22]

This huge difference helps to explain why family historians and genealogists are consistently surprised at the large numbers of aunts, uncles and cousins related to a given individual (an ascendant view), compared with the low Net Reproduction Rates estimated by demographic historians (a population or descendant view). As there is a bias towards large families, the consequence is that ascendant genealogies are automatically biased towards early marriage, and low mortality of the marriage partners, as these are associated with large families.

We can use microsimulation to investigate mortality bias in an ascendant lineage compared with a population, or descendant lineage.[23] Using the parameters for England from 1800–1837, I simulated the life–histories of one thousand female egos in an ascending lineage of one generation. Under the definition used here, this gives us ego and her sibling set experiencing unconditional mortality, plus her parents, and husband(s). The unconditional life–expectancy at birth, experienced by the general population, is 44.2 years, but the life–expectancy for this ascending lineage is 50.2 years – a considerable

---

[20] Bytheway, B. (1974) 'A statistical trap associated with family size.' *J. Biosoc. Sci.* 6, 67–72. Preston, S.H. (1976) 'Family sizes of children and family sizes of women.' *Demography,* 13,1,105–114.

[21] Completed Family Size is the number of children born to a woman whose first marriage survives to her fiftieth birthday. For the demographic estimates used see Wrigley, E.A. *et al.* (1997).

[22] The difference is, of course, even more marked if we allow mortality to have an effect by dropping the Completed Family assumption.

[23] These results were obtained by using the demographic parameters for the period as inputs to the kinship microsimulation program CAMSIM. It includes a facility for correlated random number streams, which allows specification of shared characteristics within and between generations. In these simulations, only marriage ages, and birth–intervals, have been correlated. See Smith,J.E. and Oeppen,J.E. 'Estimating numbers of kin in historical England using demographic microsimulation.' in Reher, D.S. and Schofield, R.S. (Eds. 1993) *Old and New Methods in Historical Demography.* Oxford.

upward bias.[24] However, Figure 1 shows that the age–specific differences in mortality disappear after the marrying and reproductive ages, as the mortality conditioning diminishes.
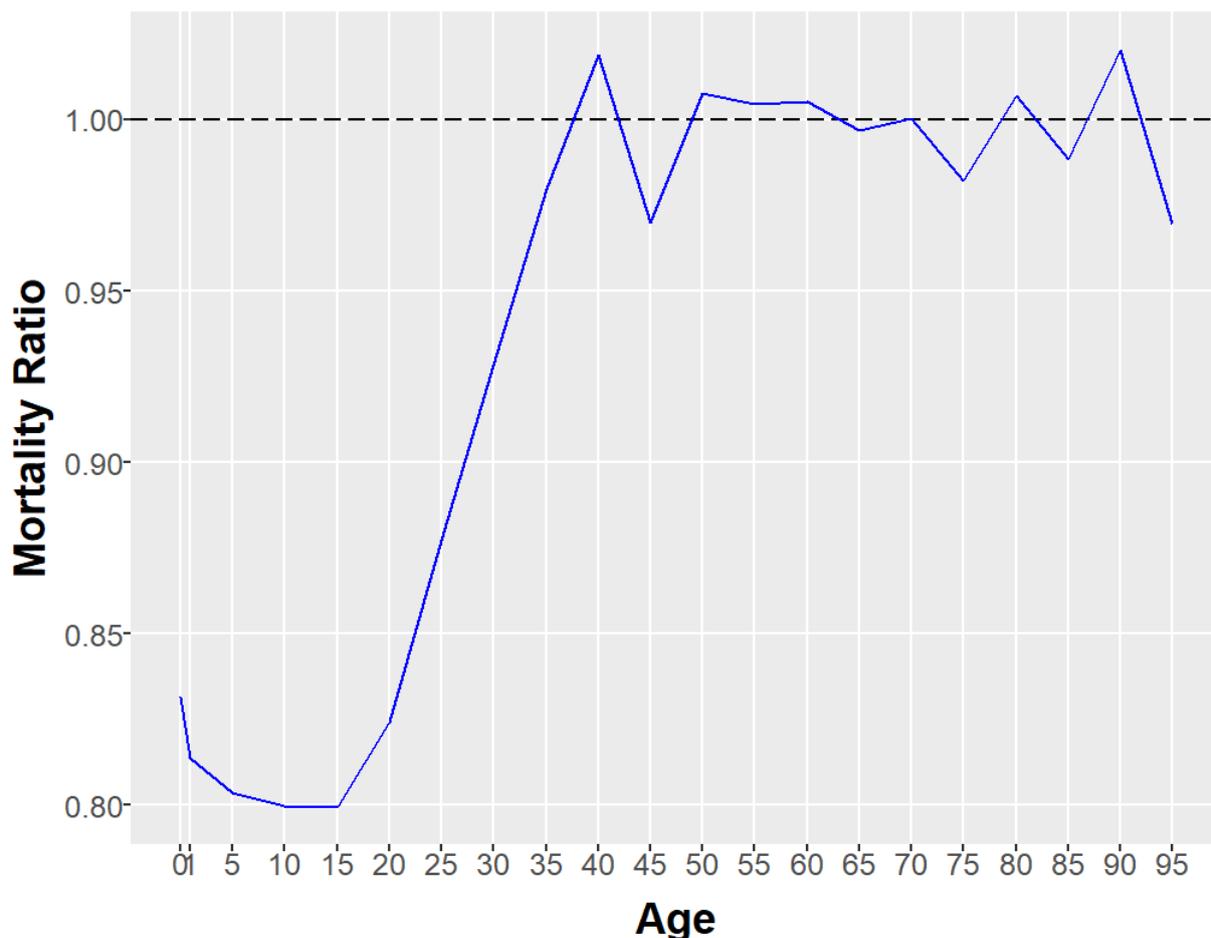

Figure 1: Mortality: Ascendant Lineage vs Population England 1800–1837.

While the gap below age fifty will depend on the interplay of particular patterns of family size, age at marriage, etc., this result suggests that it is reasonable to use all members of an ascendant lineage, as defined here, for mortality after age fifty.[25] If an actual ascending lineage includes some lateral descending elements, as they frequently do, we can expect the gap at lower ages to be diminished.

A similar experiment can be conducted for Family Histories. From an ego, the ascendant generations were simulated to the birth of all four grand–parents, who were then regarded as the egos of full descendant lineages extending downwards for three generations to the same level as the children of the original ego. Using the same parameters for 1800 to 1837 in England, and simulating a thousand such Family Histories, the life–expectancy at birth for all individuals was 47.9, compared with the population value of 44.2. Thus, even in its pure form as defined by Fogel, the Family History may overstate survival by c. 8%.[26]

Results for more complex or subtle effects could be obtained via simulation.[27] One would expect that there would be selection effects for nuptiality in an ascendant lineage, (i.e. lower celibacy and marriage age, less widowhood, and more remarriage) since these factors increase the chances of surviving offspring. The simulated lineage could be redefined as a succession by applying a specific set

---

[24]Of course, the husbands are really left–truncated by their marriage to ego. Eliminating them totally reduces the figure to 49.35, which is still a substantial over–estimate of the underlying survival chances, and reflects the conditioning on the parents' lifetimes. Add comment on Kaplanis.

[25]This assumes that all those who join the lineage by a means other than birth do so before age 50

[26]*Op. Cit.* Fn. 3

[27]See Zhao, *Op. Cit.* Fn. 11

of inheritance rules to make it more comparable with genealogies for the nobility. This would allow us to examine the complex effects of conditional survival on joint life–times that these rules create.

2.5 Inter– and Intra–Generational Correlations.

The possibility of estimating correlations within families and between generations is one of the attractions of genealogies for demographers. We know that certain characteristics, like infant–mortality, are heterogeneously distributed across families, and we would like to know if they can be shared between generations.[28]

Although there has been a recent vogue in demography for multi–level models,[29] which are specifically designed for hierarchical data of this type, there is little knowledge of the problems encountered with family data in other disciplines.

The most obvious point is that we cannot regard the children as independent observations in correlation or regression analysis, because they share unobserved characteristics, which leads to bias through auto–correlation in the errors. In biometrics, these are known as 'litter–effects' – the tendency for siblings to have more similar responses than randomly selected individuals.'[30] Secondly, if we want to look at parent–child relationships, and use the child as the unit of analysis, large families will contribute more observations to the data–set, which may be a disadvantage if stratification by family size is misleading.[31] A particular problem arises with lifetimes in male successions where, in the direct line, at least one son has to outlive his father.[32] Since lifetimes are effectively bounded variables, conventional regression methods are inappropriate for investigating the relationship. More importantly, the expected correlation between the lifetimes of the two egos under a succession rule is not zero.[33] Assuming the demographic conditions in England between 1800 and 1837, with independent mortality between individuals, the application of a first–born male succession rule leads to a correlation of 0.24 for 958 simulated father and successor lifetimes.[34] The correlation does not fall to zero if we require the father to live past 50, 60, or 70. With Replacement Rates close to unity in the past, any attempt to examine inherited male longevity in a succession may have very few usable pairs.

3 Methods of Genealogy

1. Unstandardised methods.

2. Genealogists: do they overlink?

3. Bias towards rare names - hence, low reproductive success.

4. Paper trail in the archives: biased towards long life, males, high social class, property.

5. Theoretically unaffected by migration. However, migration risk, and resulting loss–to–followup, increases with age. Therefore, the relative proportion of young deaths in birth cohorts should be over–estimated. In practice, the reverse seems to be observed, increasing the probability that young deaths are missing.

[28]Zaba, B. and David, P.H. 'Fertility and the distribution of child mortality risk among women: an illustrative analysis.' *Population Studies,* 1996, 50, 2, pp. 263–78. Das Gupta, M. 'Socio–economic status and clustering of child deaths in rural Punjab.' *Population Studies,* 1997, 51, 2, pp. 191–202.

[29]Goldstein, H. *Multilevel Statistical Models.* 1995, London.

[30]Segreti, A.C. and Munson, A.E. 'Estimation of the median lethal dose when responses within a litter are correlated.' *Biometrics,* 1981, 37, 153–6.

[31]Family size is a proxy for many other demographic characteristics.

[32]This can be thought of as a kind of truncation selection in breeding. Hartl, D.L. and Clark, A.G. *Principles of Population Genetics,* 1997, 3rd edn., p.407.

[33]Obviously, the longer the father lives, the longer the heir has to live to succeed him.

[34]As expected, the correlation between the mother's and the successor's lifetimes is approximately zero at 0.02.

4 Illustrations

4.1 Long–lived family studies

Examples: Leuven, Denmark, Perls.

4.2 Genealogies downloaded from the Web.

Examples: Kaplanis *et al.*

5 Bias Correction

6 Implications for Genetic Studies

7 Implications for Migration Studies

1. Assumption that bias and migration are independent.

2. Migration creates problems for genealogy: linkage 'cost' and distance.

3. Single women migrate to towns.

4. Childless women.

5. Problems shared with family reconstitution.

8 Conclusion.

In this paper, after ignoring issues of representativeness and missing data, I have argued that both ascendant lineages and successions, and ascendant/descendant family histories, are biased with respect to nuptiality, fertility and mortality, but that descendant lineages and successions can be unbiased with regard to the population they represent. These criticisms of ascendant genealogies apply even if they have no missing or incomplete data. The biases estimated here for fertility and mortality in ascendant lineages under pre–demographic transition conditions are substantial. The good news for analysts of longevity is that the biases in mortality from ascendant lineages seem to disappear after the fiftieth birthday.

   Does it matter for genetic research that genealogies can be biased? Willigan and Lynch say 'Looking at the reproductive and survival behavior of the most genetically 'fit' people will not bias the results of population–genetics research based on genealogies.'[35] Correlations within and between generations in genealogies seem to be fraught with difficulties so, at present, I would rather keep an open mind.

---

[35]Willigan, J.D. and Lynch, K.A. (1982) *Sources and Methods for Historical Demography.* London.