

PAA Extended Abstract - please do not cite

Estimating adult death rates from sibling histories: a network approach

Dennis M. Feehan^{*} and Gabriel M. Borges[†]

April 7, 2018

Abstract

Most of the world's poorest countries do not have complete death registration systems, meaning that critical quantities like life expectancy cannot be directly measured. The sibling survival method is a leading approach to estimating adult mortality in the absence of death registration. The idea is to ask a survey respondent to enumerate her siblings and to report about their survival status. In many countries and time periods, sibling survival data are the only nationally-representative source of information about adult mortality. Although a huge amount of sibling survival data has been collected, important methodological questions about the method remain unresolved. To help make progress on this issue, we propose re-framing the sibling survival method as a network sampling problem. This approach enables us to formally derive a statistical estimator from sibling survival data. Our derivation clarifies the precise conditions that sibling history estimates rely upon; it leads to closed-form variance estimators; and it reveals important quantities that could potentially be measured to relax assumptions in the future.

^{*}UC Berkeley, feehan@berkeley.edu

[†]UC Berkeley, gmendesb@berkeley.edu

1 Introduction

Death rates at adult ages are a core component of population health and a central topic of study for demography. Unfortunately, most of the world’s poorest countries are victims of the *scandal of invisibility*: they do not have complete death registration systems, meaning that most people die without ever having their existence officially recorded (C. AbouZahr et al. 2015; P. W. Setel et al. 2007). This lack of complete death registration means that critical quantities like life expectancy cannot be directly measured. Improving death registration systems is the long-term solution to the scandal of invisibility, but progress has been very slow (Mikkelsen et al. 2015). Until complete death registration systems are available everywhere, sample-based approaches to adult mortality estimation will continue to play a pivotal role in understanding population health and wellbeing.

The leading approach to estimating adult mortality in the absence of death registration is the sibling survival method (Rutenberg and Sullivan 1991). The idea is to ask survey respondents to report the number of siblings they have¹, and to then ask for each sibling’s gender, date of birth and date of death (where appropriate). This data collection strategy produces *sibling histories* which contain information about the survival status of all of the members of the respondent’s sibship.

Since high-quality household surveys are routinely conducted in most countries—including countries that lack death registration systems—the sibling survival method offers the opportunity to try to estimate adult death rates in many places that have no other nationally-representative adult mortality data. Over the past two decades, a huge amount of sibling history data has been collected; for example, as a part of the DHS program alone, sibling histories have been collected in more than 150 surveys from dozens of countries around the world (Corsi et al. 2012; Fabic et al. 2012).

Despite the importance of adult mortality and the lack of alternative sources of national-level information about adult deaths, relatively few studies have made use of the available sibling history data. This under-use of the available sibling history data may be the result of considerable methodological uncertainty about how sibling histories should be analyzed. Researchers have long been aware that the method suffers from many possible sources of bias (Gakidou and King 2006; Graham et al. 1989; B. Masquelier 2013; Reniers et al. 2011; Trussell and Rodriguez 1990). Previous studies have concluded that sibling history estimates can be problematic if (i) there is a relationship between sibship size and mortality (*e.g.* larger sibships face higher death rates); (ii) there are sibships with no surviving members who could potentially be sampled and interviewed in the survey; and (iii) respondents’ reports about their siblings are inaccurate (*e.g.* respondents may omit siblings or misreport a sibling’s survival status). There has also been confusion about whether the survival status of the respondent herself should be included in the calculations, since respondents are always alive (B. Masquelier 2013; Reniers et al. 2011).

Researchers have worked on addressing these concerns about the sibling survival method in three main ways: they have collected empirical information about possible sources of bias in sibling reports (*e.g.*, Helleringer et al. 2014); they have used microsimulation to illustrate how large certain sources of bias can be under different scenarios (B. Masquelier 2013); and they have used regression models to pool information from different countries and time periods (Gakidou and King 2006; Obermeyer et al. 2010; Timaeus 1991). Together, these studies have produced many important insights about the sibling survival method. However, these insights have not yet brought about a consensus on how sibling histories should be analyzed.

¹Respondents are typically asked to consider ‘siblings’ to be all children born to their mother.

What may be lacking is a single analytical framework that can put the various insights together. Currently, there is partial evidence about many individual sources of possible bias, but there is no way to integrate all of this evidence together. Thus, even if we knew the exact size and direction of all the different sources of possible error, we still would not understand how the errors would combine to affect estimated death rates. More generally, little has been proven about the precise conditions under which sibling survival estimates can be expected to have attractive statistical properties such as consistency, unbiasedness, or efficiency.

In this study, our goal is to help resolve some of the methodological uncertainty about sibling survival. Our analysis is based on the insight that the sibling relation induces a particular type of social network among the members of a population. In this network, two people are connected to one another if they are siblings; thus, estimating death rates from sibling histories can be understood as a problem in network sampling. By conceptualizing sibships as networks, we are able to bring together insights from two historically distinct literatures: one literature on mortality estimation and a second literature on sampling in social networks. Starting from the principles of network reporting, we describe how to mathematically derive a sibling survival estimator. Deriving an estimator from first principles in this way enables us to (i) clarify the precise assumptions that the estimator requires in order to be consistent, unbiased, and efficient; (ii) describe how violations of any and all assumptions can combine to affect estimated death rates; (iii) identify quantities that could potentially be measured in the future to relax assumptions; and (iv) derive a closed-form expression for sampling variance and confidence intervals.

In the remainder of this extended abstract, we explain the intuition behind our approach and we sketch several steps in the derivation of an estimator for adult mortality based on sibling history data. The Appendix contains several derivations, and we present highlights in the main text. We explain the framework for sensitivity analysis and illustrate how the different assumptions that underlie the estimator interact with one another to affect death rate estimates. The full paper will fill in some of the technical details; it will introduce an estimator for sampling variance; and it will provide additional empirical examples.

2 Deriving a sibling history estimator

Figure 1 illustrates how we understand sibling histories as a network reporting problem. The left-hand panel shows a small population whose members are joined together in a network; the network is defined by the sibling relation (*i.e.*, two nodes are connected if they have the same mother). Since the sibling relation is transitive, the network is entirely composed of fully connected components, or cliques; each of these cliques is one sibship. Clear nodes are alive and grey nodes are dead at the time of the survey. The middle panel shows one specific sibship, and the right-hand panel illustrates the bipartite reporting network that is generated when all of the surviving siblings are asked to report about their sibship. In the bipartite reporting network, Each directed edge represents a report so that, for example, the edge $13 \rightarrow 12$ indicates that node 13 reports about node 12². Feehan and Salganik (2016a) describes bipartite reporting networks in greater detail.

The quantity of interest is M_α , the death rate for a specific group α (for example, α might be all

²The dead person (in grey) cannot be interviewed, and so is not shown on the left-hand side of the bipartite reporting network.

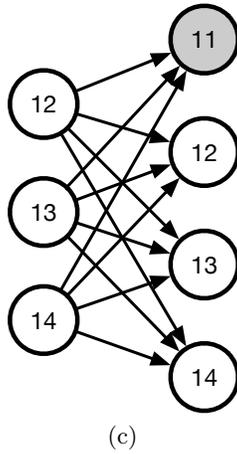
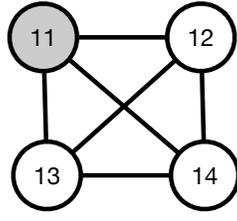
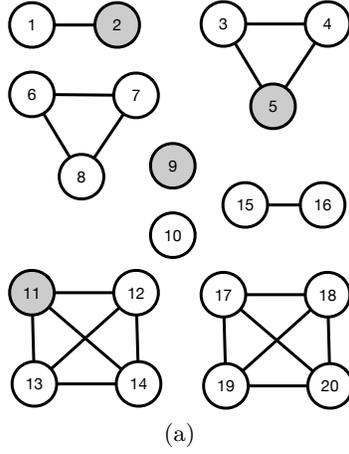


Figure 1: Framing sibling survival as a network reporting estimator. (a) A population connected through a sibship network. Because the sibling relation is transitive, the network is composed entirely of cliques. (b) The network for a single sibship. (c) Bipartite reporting graph for a single sibship.

women aged 30-34 in 2017). M_α is defined as

$$M_\alpha = \frac{D_\alpha}{N_\alpha}, \quad (1)$$

where D_α is the number of deaths in group α and N_α is the person-years of exposure among members of group α . We can develop an estimator for M_α by separately estimating the numerator and the denominator of Equation 1.

Challenges

There are two key challenges that arise when using sibling histories to estimate death rates: reporting accuracy and differential visibility. We will analyze each key challenge in turn.

The first challenge is that respondents' reports about their siblings may not be perfectly accurate. Using the network reporting framework enables us to introduce two ways to help address this problem: first, we can use the framework to derive *internal consistency checks*. These internal consistency checks use sibling histories to produce two independent estimates of the same quantity. If sibling reports are accurate, these internal consistency checks should agree; if, on the other hand, sibling reports contain many errors, then these internal consistency checks can help us detect that fact. Second, we develop a mathematical framework for sensitivity analysis. This framework is useful because (i) it enables us to understand how much estimated death rates would change when reporting is not accurate; and (ii) it identifies several empirical quantities that could potentially be measured in the future, allowing sibling history estimates to be adjusted and improved.

The second key challenge in estimating M_α from sibling history data is that each sibling can potentially be reported as many times as she has living sibship members who are eligible to respond to the survey (Gakidou and King 2006; Sirken 1970). Inferences from sibling reports must somehow account for this fact.

Our approach is to distinguish between two types of people: first, some people have no siblings who are eligible to respond to the survey. These people will never appear in the sibling history data – they are *invisible* to the sibling histories. Second, *visible* people do have siblings eligible to respond to the survey. We treat these two groups separately.

Formally, we define person j 's *visibility*, $v_{j,F}$, to be the number of living siblings who would report person j in a census of all eligible survey respondents. Using the number of deaths in group α , D_α , as an example we can write:

$$D_\alpha = \underbrace{\sum_{\substack{j \in D_\alpha \\ v_{j,F} > 0}} 1}_{\text{visible deaths}} + \underbrace{\sum_{\substack{j \in D_\alpha \\ v_{j,F} = 0}} 1}_{\text{invisible deaths}} = D_\alpha^V + D_\alpha^I, \quad (2)$$

where D_α^V is the number of *visible deaths* that could be learned about using sibling reports and D_α^I is the number of *invisible deaths* that cannot be learned about using sibling reports. We can define analogous quantities for the denominator $N_\alpha = N_\alpha^V + N_\alpha^I$, where the N_α^V is the *visible exposure* and N_α^I is the *invisible exposure*. Finally, we define $M_\alpha^I = \frac{D_\alpha^I}{N_\alpha^I}$ to be the *invisible death rate*, $M_\alpha^V = \frac{D_\alpha^V}{N_\alpha^V}$ to be the *visible death rate*, and $M_\alpha = \frac{D_\alpha^I + D_\alpha^V}{N_\alpha^I + N_\alpha^V}$ to be the *total death rate*.

In order to address the fact that some people are invisible to the sibling histories, we derive mathematical expressions that characterize the error that results from using only information about visible people as an estimate for the death rate of all people. Our results suggest that, in many cases, the effect of omitting invisible deaths may not be very large, particularly when the focus is on estimating death rates for recent time periods (such as 12 months before the survey).

Next, we must address the fact that even people who are visible to the sibling histories can differ in the extent to which they are visible. We address this challenge by introducing statistical estimators that adjust for how visible reported siblings are. We describe these estimators in more detail in the next section.

To summarize: our approach is to focus on carefully estimating death rates among visible people, and to then develop a framework for sensitivity analysis that enables us to understand how differences in the visible and invisible death rates would affect our death rate estimate.

Adjusting for visibility in sibling reports

The visible death rate in a demographic group α can be estimated with an expression of the form

$$\widehat{M}_\alpha^V = \frac{\widehat{D}_\alpha}{\widehat{N}_\alpha}. \quad (3)$$

The main decision to be made is how to choose sibling history-based estimators for D_α and N_α that address the problem of overcounting. We consider two different approaches: *individual visibility* estimation and *aggregate visibility* estimation. In both cases, we discuss how to estimate Equation 1 in a census of all eligible survey respondents when there are no false positive reports. It is not necessary to incorporate sampling in order to distinguish between the individual and aggregate visibility approaches, so we defer this additional complication until later in the paper.

The *individual visibility* approach adjusts for the visibility of reported quantities at the level of each sibship (Feehan 2015; Gakidou and King 2006; Lavalley 2007; Sirken 1970). To illustrate this approach, we focus first on the reports about a specific deceased sibling $j \in D_\alpha$. We have $y_{F,j} = v_{j,F}$ by definition; thus, for every sibling j whose death is visible (so that $v_{j,F} \neq 0$), we have

$$\frac{y_{F,j}}{v_{j,F}} = 1.$$

Intuitively, each sibling is reported about $v_{j,F}$ times in a census of the frame population. The relationship between visible deaths and reports can thus be written

$$D_\alpha^V = \sum_{j \in D_\alpha^V} 1 = \sum_{j \in D_\alpha^V} \frac{y_{F,j}}{v_{j,F}} = \sum_{i \in F} \frac{y_{i,D_\alpha}}{v_{i,F}}.$$

The last step relies upon the fact that $v_{i,F} = v_{j,F}$ for all i and j who are in the same sibship; this is a result of the fact that the sibling relation is transitive. The deaths reported by survey respondent i all have the same visibility as i herself.

Since this approach is based on adjusting for the visibility of each individual reported sibling, we call it *individual visibility* estimation (Gakidou and King 2006; Lavalley 2007; Sirken 1970). Appendix

A formally derives a consistent and unbiased estimator for D_α^V , N_α^V , and the visible death rate M_α^V based on the individual visibility approach. The actual estimator for visible death rates based on the individual visibility estimator is developed in Appendix A.

The *aggregate visibility* approach adjusts for the visibility of reported quantities at the level of each sibship (Bernard et al. 1989; Feehan and Salganik 2016a; Shea Oscar Rutstein and Rojas 2006). An aggregate visibility estimate for D_α is based on the total reported number of siblings who have died:

$$D_\alpha = \frac{\# \text{ reported sibs who died}}{\text{avg. } \# \text{ of times each dead sib is reported}} = \frac{y_{F,D_\alpha}}{\bar{v}_{D_\alpha,F}}, \quad (4)$$

where y_{F,D_α} is the estimated total number of reports about sibling deaths and $\bar{v}_{D_\alpha,F}$ is the average number of times each dead sibling is reported. Since this approach is based on adding together reports about all of the siblings, the estimator based on Equation 4 is called an *aggregate visibility* estimator (Bernard et al. 1989; Feehan 2015; Feehan et al. 2017). The actual estimator for visible death rates based on the aggregate visibility estimator is developed in Appendix A (Result A.5). The derivation reveals that this approach can be expected to produce unbiased estimates as long as reports about siblings are accurate, and as long as there is no relationship between sibship size and mortality.

3 Internal consistency checks

In this section, we introduce several internal consistency checks that can be used to assess how accurate sibling reports are. In order to devise these internal consistency checks, we used the network reporting framework to identify several quantities that can be estimated in two different ways using independent subsets of the sibling history data. If reporting is highly accurate, then we expect these independent estimates to agree; when these independent estimates are very different, that suggests that there may be considerable amounts of reporting error.

We base our internal consistency checks on reports about siblings' ages. For a particular age, say 30, the symmetry of the sibling relation guarantees that in a census of the frame population,

$$\frac{\# \text{ connections people in } F \text{ aged 30}}{\text{report to sibs in } F \text{ not aged 30}} = \frac{\# \text{ connections people in } F \text{ not aged 30}}{\text{report to sibs in } F \text{ aged 30}}. \quad (5)$$

The quantity on the left-hand side of Equation 5 can be estimated from the survey respondents aged 30, and the quantity on the right-hand side of Equation 5 can be estimated from all of the survey respondents who are not aged 30. These two estimates can then be compared; if reporting is accurate, then we expect the estimates to agree.

Formally, for age α , we write $y_{F_\alpha, F_{-\alpha}}$ for the total reported connections from respondents aged α to siblings who are in F but not aged α ; similarly, we write $y_{F_{-\alpha}, F_\alpha}$ for the total reported connections from respondents not aged α to siblings who are in F and who are age α . In theory, these are same quantity. From sibling history data, we can independently estimate $y_{F_\alpha, F_{-\alpha}}$ and $y_{F_{-\alpha}, F_\alpha}$ and then compare how similar these two independent estimates of the same quantity are by calculating Δ_α :

$$\Delta_\alpha = \hat{y}_{F_\alpha, F_{-\alpha}} - \hat{y}_{F_{-\alpha}, F_\alpha}.$$

When the two estimates agree, Δ_α is close to zero. If there is considerable reporting error, then

Δ_α can be very different from zero.

Figure 2 illustrates this idea by showing internal consistency checks for each age from the 2000 Malawi DHS sibling histories (Malawi National Statistical Office and ORC Macro 2001). The figure shows an internal consistency check based on each age from 15 to 49. Each point shows the difference between two independent estimates for the same quantity. If these independent estimates agreed perfectly, they would all lie on the horizontal $y = 0$ line. The confidence intervals capture estimated sampling variation³. Most of the confidence intervals include 0, suggesting that reports are internally consistent; however, the figure also suggests that there is some age heaping or misreporting, particularly around ages 20 and 35.

Note that internal consistency checks need not be based only on age; any quantity that is measured for survey respondents and for siblings can be the basis of an internal consistency check. Furthermore, these internal consistency checks could form the basis of a model whose goal is to remove some of the reporting errors by ensuring that the internal consistency checks are satisfied. It would be helpful for future research to explore this possibility.

4 Framework for sensitivity analysis

In order to estimate death rates from sibling histories, several formal conditions must be satisfied (Appendix A). In practice, researchers may not be able to ensure that these conditions are all exactly met. We therefore introduce a mathematical framework for understanding how sensitive sibling history estimates are to the conditions that they rely upon⁴. Our analysis also reveals several key quantities that could potentially be measured in the future, enabling estimated death rates to be adjusted and improved. In this extended abstract, we focus on the sensitivity of the aggregate visibility estimator.

The derivation in Appendix B reveals that the aggregate visibility estimand can be written as:

$$\underbrace{\frac{y_{F,D_\alpha}}{y_{F,N_\alpha}}}_{\text{aggregate visibility estimand}} = \underbrace{M_\alpha}_{\text{total death rate}} \times \underbrace{\left[\frac{d_{D_\alpha,F}^V}{d_{N_\alpha,F}^V} \right]}_{\text{relationship between visibility and mortality}} \times \underbrace{\left[\frac{\eta_{F,N_\alpha} \times \tau_{F,D_\alpha}}{\eta_{F,D_\alpha} \times \tau_{F,N_\alpha}} \right]}_{\text{accuracy of reporting}} \times \underbrace{\left[\frac{p_{D_\alpha}^I + K(1 - p_{D_\alpha}^I)}{K} \right]}_{\text{difference between visible and invisible death rates}}. \quad (6)$$

Equation 6 shows that the aggregate visibility estimand can be decomposed into the total death rate times several *adjustment factors*. These adjustment factors are related to the conditions that the aggregate visibility estimator relies upon. When all of these adjustment factors are equal to 1, the aggregate visibility estimand is equal to the total death rate⁵.

Equation 6 is written to combine substantively related adjustment factors into meaningful groups. We will now discuss each of these groups in turn.

³Sampling uncertainty is estimated using the rescaled bootstrap method (Feehan and Salganik 2016b; Rao and Wu 1988). This approach accounts for the complex sampling design.

⁴In this extended abstract, we focus on the sensitivity of the aggregate visibility estimator; Appendix B contains more detail, including results for the individual visibility estimator.

⁵More generally, if these adjustment factors multiply out to be 1, the aggregate visibility estimand will be the total death rate. This means that the conditions that the estimator relies upon are sufficient, but not necessary.

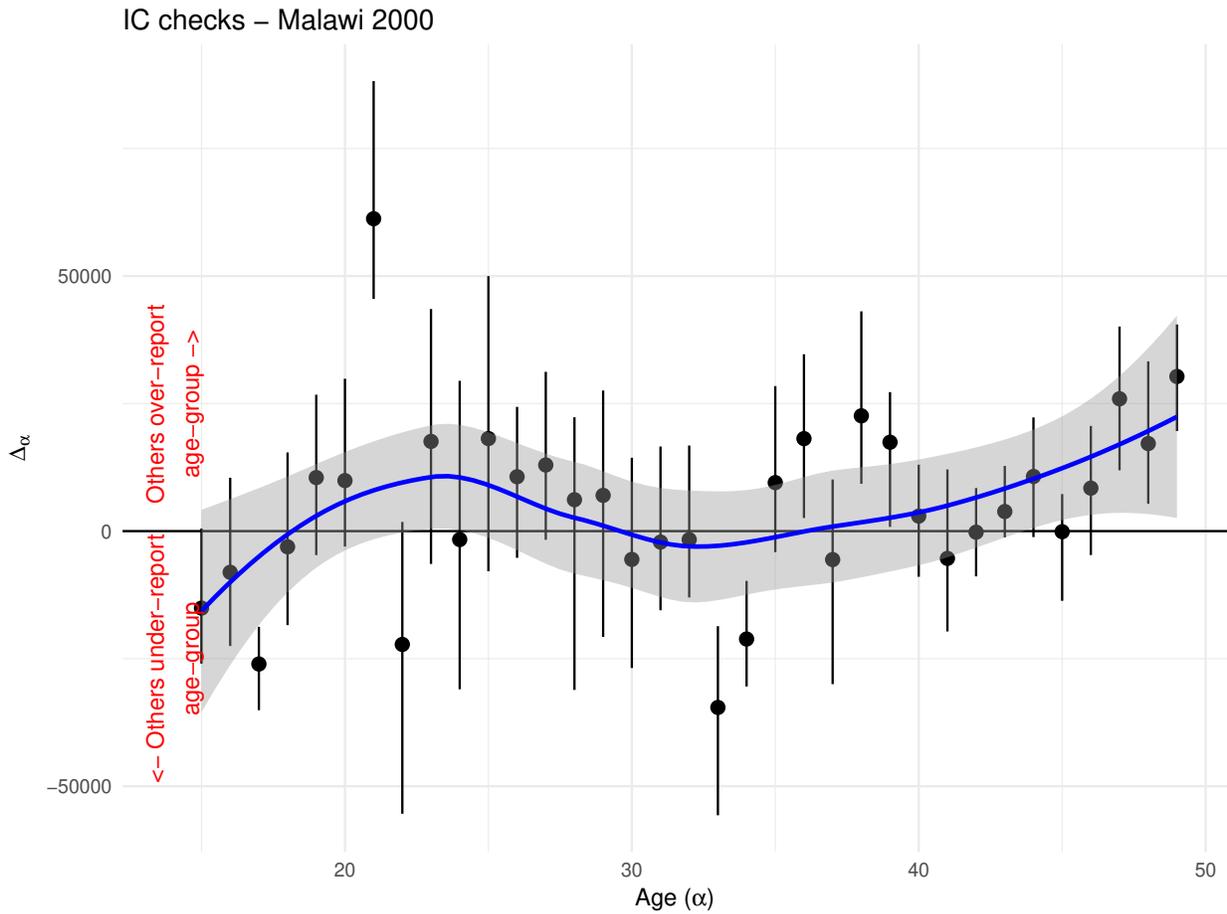


Figure 2: Internal consistency checks using sibling reports from the 2000 Malawi DHS sibling histories. Each point shows the difference between two independent estimates for the same quantity. If these independent estimates agreed perfectly, they would all lie on the horizontal $y = 0$ line. Most of the checks' confidence intervals include 0, suggesting that reports are internally consistent. However, the figure also suggests that there is some age heaping or misreporting, particularly around ages 20 and 35. Confidence intervals show sampling uncertainty, accounting for the complex sample design; they were produced using the rescaled bootstrap method (Feehan and Salganik 2016b; Rao and Wu 1988).

The first group captures a long-standing methodological concern from the sibling survival literature: the relationship between visibility and mortality. When there is no relationship between these two quantities, the average number of sibling connections between the frame population and any sibling will be the same as the average number of sibling connections between the frame population and a sibling who died. When siblings who have died tend to be in more visible sibships than siblings overall, this factor will tend to be larger than 1 (meaning that the death rate will be over-estimated).

The next group captures the extent to which reports about siblings are accurate. Two particularly salient findings emerge from the derivation of this group of adjustment factors (Appendix B): first, the estimator requires that reporting be accurate *in aggregate*, but not necessarily at the individual level; as long as reporting errors cancel one another out, estimates will not be affected. Second, this group of adjustment factors shows that estimates will also not be affected if reporting errors about deaths and reporting errors about siblings balance out. In other words, if respondents tend to, say, omit older siblings at a constant rate, independent of the survival status of older siblings ($\tau_{F,D_\alpha} = \tau_{F,N_\alpha} < 1$), then Equation 6 shows that this will not affect death rate estimates because the reporting about deaths and about all siblings will cancel out.

Finally, the last group captures the conditions needed to be able to use only information about visible deaths to estimate the total death rate. This group depends on two quantities: $p_{D_\alpha}^I$, the fraction of deaths that is invisible; and $K = \frac{M^I}{M^V}$, an index for how different the invisible and visible death rates are. Two figures help us to better understand the aggregate visibility estimator’s sensitivity to using only information about visible deaths. First, in order to assess roughly what share of deaths we expect to be visible by age, Figure 3 shows the estimated fraction of respondents to the 2000 Malawi DHS who would not be visible in a sibling sample; thus, this figure gives us a sense for the type of values that we might expect to see for $p_{D_\alpha}^I$. Next, Figure 4 uses the sensitivity framework to illustrate the relative error that results from using the visible death rate to estimate the total death rate over a range of values of $p_{D_\alpha}^V$ and K . Even relatively large values for the two parameters appear to result in modest relative errors.

5 Conclusion

In this extended abstract, we showed how the sibling survival method can be understood as a network approach to estimating adult death rates. We explained how to derive network-based estimators for adult death rates, how to devise internal consistency checks, and how to understand how sensitive death rate estimates can be to the different conditions that the estimators rely upon. Future versions of this paper will derive a variance estimator; fill in technical details in the Appendix; and supplement the analysis here with additional empirical examples.

References

- AbouZahr, C., de Savigny, D., Mikkelsen, L., Setel, P. W., Lozano, R., & Lopez, A. D. (2015). Towards universal civil registration and vital statistics systems: The time is now. *The Lancet*.
- Bernard, H. R., Johnsen, E. C., Killworth, P. D., & Robinson, S. (1989). Estimating the size of an average personal network and of an event subpopulation. In M. Kochen (Ed.), *The Small World* (pp. 159–175). Norwood, NJ: Ablex Publishing.

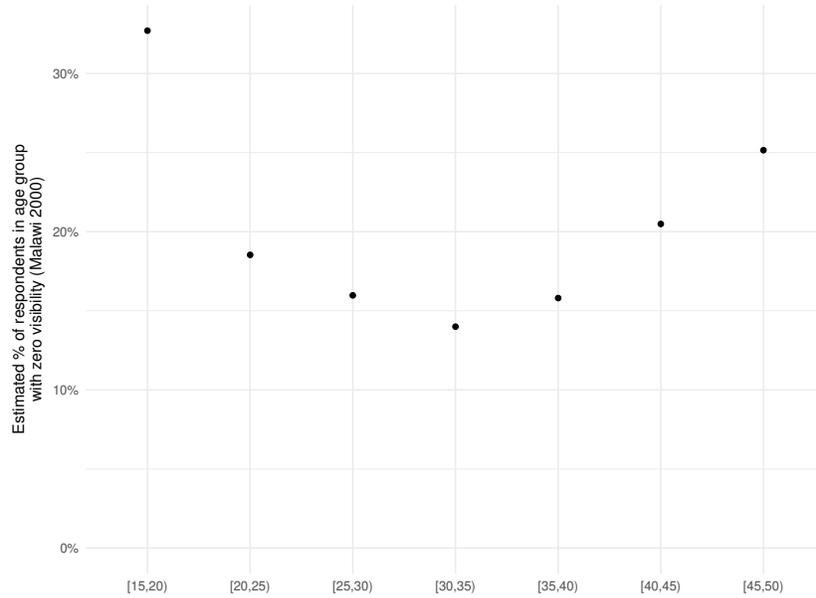


Figure 3: Estimated fraction of respondents to the Malawi 2000 DHS who would not be visible in sibling history data.

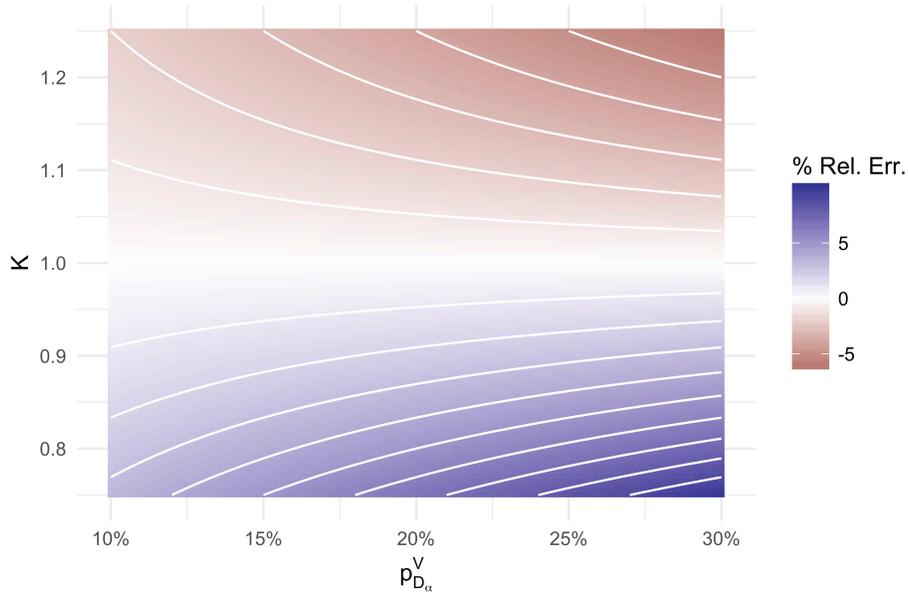


Figure 4: Illustration of the relative error in using the visible death rate M^V as an estimate for the total death rate M . The proportion of deaths that is invisible, $p_{D_\alpha}^V$, varies along the x axis; the relationship between the visible and invisible death rates, captured by the parameter K (Equation 10), varies along the y axis. The colors show the percentage relative error; so if 20% of deaths are invisible and the invisible death rate is 10% higher than the visible death rate, the relative error is about 2 percent. Relative error increases as K gets farther away from 1 and as $p_{D_\alpha}^V$ increases.

- Corsi, D. J., Neuman, M., Finlay, J. E., & Subramanian, S. V. (2012). Demographic and health surveys: A profile. *International Journal of Epidemiology*, *41*(6), 1602–1613.
- Fabic, M. S., Choi, Y., & Bird, S. (2012). A systematic review of Demographic and Health Surveys: Data availability and utilization for research. *Bulletin of the World Health Organization*, *90*(8), 604–612.
- Feehan, D. M. (2015). *Network reporting methods* (PhD thesis). Princeton University.
- Feehan, D. M., & Salganik, M. J. (2016a). Generalizing the Network Scale-Up Method: A New Estimator for the Size of Hidden Populations. *Sociological Methodology*.
- Feehan, D. M., & Salganik, M. J. (2016b). *Surveybootstrap: Tools for the Bootstrap with Survey Data*.
- Feehan, D. M., Mahy, M., & Salganik, M. J. (2017). The network survival method for estimating adult mortality: Evidence from a survey experiment in Rwanda. *Demography*, *54*(4), 1503–1528.
- Gakidou, E., & King, G. (2006). Death by survey: Estimating adult mortality without selection bias from sibling survival data. *Demography*, *43*(3), 569–585.
- Graham, W., Brass, W., & Snow, R. W. (1989). Estimating maternal mortality: The sisterhood method. *Studies in Family Planning*, 125–135.
- Helleringer, S., Pison, G., Kanté, A. M., Duthé, G., & Andro, A. (2014). Reporting errors in siblings’ survival histories and their impact on adult mortality estimates: Results from a record linkage study in Senegal. *Demography*, *51*(2), 387–411.
- Lavallee, P. (2007). *Indirect sampling*. New York: Springer-Verlag.
- Malawi National Statistical Office, & ORC Macro. (2001). *Malawi Demographic and Health Survey 2000*. Zomba, Malawi: National Statistical Office.
- Masquelier, B. (2013). Adult mortality from sibling survival data: A reappraisal of selection biases. *Demography*, *50*(1), 207–228.
- Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., de Savigny, D., Lozano, R., & Lopez, A. D. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *The Lancet*.
- Obermeyer, Z., Rajaratnam, J. K., Park, C. H., Gakidou, E., Hogan, M. C., Lopez, A. D., & Murray, C. J. L. (2010). Measuring adult mortality using sibling survival: A new analytical method and new results for 44 countries, 1974–2006. *PLoS medicine*, *7*(4), e1000260.
- Rao, J. N. K., & Pereira, N. P. (1968). On double ratio estimators. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *30*(1), 83–90.
- Rao, J. N. K., & Wu, C. F. J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, *83*(401), 231–241.
- Reniers, G., Masquelier, B., & Gerland, P. (2011). Adult mortality in Africa. *International Handbook of Adult Mortality*, 151–170.
- Rutenberg, N., & Sullivan, J. M. (1991). Direct and indirect estimates of maternal mortality from the sisterhood method. In *Proceedings of the Demographic and Health Surveys World Conference* (Vol. 3, pp. 1669–1696).

- Rutstein, S. O., & Guillermo Rojas, M. C. S. (2006). *Guide to DHS statistics*. ORC Macro, Calverton, MD.
- Rutstein, S. O., & Rojas, G. (2006). Guide to DHS statistics. *Calverton, Maryland: ORC Macro*.
- Sarndal, C. E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling*. Springer Verlag.
- Setel, P. W., Macfarlane, S. B., Szreter, S., Mikkelsen, L., Jha, P., Stout, S., & AbouZahr, C. (2007). A scandal of invisibility: Making everyone count by counting everyone. *The Lancet*, 370(9598), 1569–1577.
- Sirken, M. G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65(329), 257–266.
- Thompson, S. K. (2002). *Sampling* (2nd ed.). Wiley.
- Timaeus, I. M. (1991). Measurement of adult mortality in less developed countries: A comparative review. *Population Index*, 552–568.
- Trussell, J., & Rodriguez, G. (1990). A Note on the Sisterhood Estimator of Maternal Mortality. *Studies in Family Planning*, 21(6), 344–346.
- Wolter, K. (2007). *Introduction to Variance Estimation* (2nd ed.). New York: Springer.

A Appendix

In this appendix, we introduce the formal setup for design-based sampling and we derive the various components of the different sibling survival estimators we discuss.

There are two types of aggregate visibility estimators, according to whether the analyst decides to use the sibling histories to estimate (1) the numerator of the death rate, or (2) the numerator and denominator of the death rate.

Although, in principle, both of these approaches are possible, the second approach – that is, using sibling histories to estimate both the numerator and the denominator of death rates – seems much more promising. The challenge in the first approach is that the sibling histories enable researchers to estimate visible deaths. In order to even estimate the visible death rate, an estimate for the number of visible deaths would have to be combined with an estimate for the visible exposure (not the total exposure.) We therefore focus only on estimators that use sibling history data for both the deaths and the exposure here.

A.1 Sampling setup

We use the design-based sampling framework described in Sarndal et al. (2003), repeating a few key definitions here for convenience. We assume we have a probability sample s from a frame population F ; common frame populations include all adults, all adults aged 15-59, and in many DHS surveys, all women aged 15-59. Each $i \in F$ has a nonzero *probability of inclusion* π_i and the sampling weights are given by $w_i = \frac{1}{\pi_i}$.

Suppose some quantity y_i is defined for every $i \in F$. Then the *Horvitz Thompson estimator* for the population total $Y = \sum_{i \in F} y_i$ from a probability sample s is given by

$$\hat{Y} = \sum_{i \in s} w_i y_i.$$

Sarndal et al. (2003) shows that Horvitz-Thompson estimators are consistent and unbiased⁶, a fact that will be useful below.

Result A.1. *Suppose a Horvitz-Thompson estimator*

$$\hat{Y}^{HT} = \sum_{i \in s} w_i f(\mathbf{x}_i)$$

is design-unbiased for a total Y . Then \hat{Y}^{HT} is also (Fisher) consistent for Y .

Proof. This result follows from taking the sampling design to assign $p(i) = 1$ for all $i \in F$. We then have $s = F$ and $w_i = 1$ for all i . Since the estimator is unbiased, Fisher consistency follows. □

⁶In this paper, we use the framework of design-based sampling, so the properties of estimators – such as unbiasedness and consistency – are with respect to the probability sampling mechanism. There are many types of consistency; we refer in this work to design-consistency, also called Fisher consistency.

Next, we state a Result that is helpful when devising estimators that are ratios of other estimators.

Result A.2. *Suppose that $\hat{y}_1, \dots, \hat{y}_n$ are estimators that are consistent and unbiased for Y_1, \dots, Y_n respectively. Then the compound ratio estimator*

$$\hat{R} = \frac{\hat{y}_1 \cdots \hat{y}_k}{\hat{y}_{k+1} \cdots \hat{y}_n}.$$

is consistent and essentially unbiased for $R = (Y_1 \cdots Y_k)/(Y_{k+1} \cdots Y_n)$.

Proof. See Rao and Pereira (1968), Wolter (2007) (pg. 233), and Feehan and Salganik (2016a) for more details.

□

The references that derive the compound ratio estimator in Result A.2 show that, technically, the estimator is biased. However, a long literature has concluded that the estimator is very nearly unbiased in many applied situations; this is why we refer to it as *essentially unbiased*. The literature on ratio estimators has revealed that bias can be a problem if the denominators are very small; we do not expect this to be the case for our applications here.

Individual visibility estimators

Definition A.1. The **individual visibility** estimator for a total $\mathbf{Y} = \sum_{i \in U} y_i$ is defined to be

$$\hat{Y}^{\text{ind}} = \sum_{i \in s} w_i \sum_{j \sim i} \frac{y_j}{v_{j,F}}, \quad (7)$$

where $i \sim j$ indexes the neighbors j of respondent i , w_i is the design weight for respondent i , and $\hat{v}_{j,F}$ is the visibility of respondent j to the frame population F . The visibility $v_{j,F}$ has also been called the multiplicity of person j .

Result A.3. *Suppose that there are no false positive reports. The individual visibility estimator is consistent and unbiased for the total \mathbf{Y} .*

Proof. First, we show that the estimator is unbiased. If π_i is i 's probability of inclusion under the

sampling design, then the design weights are $w_i = \frac{1}{\pi_i}$. Thus, we have

$$\begin{aligned}
\widehat{\mathbf{Y}}^{\text{ind}} &= \sum_{i \in s} \frac{1}{\pi_i} \sum_{j \sim i} \frac{y_j}{v_{j,F}} \\
&= \sum_{i \in F} I_i \frac{1}{\pi_i} \sum_{j \sim i} \frac{y_j}{v_{j,F}} \\
&= \sum_{i \in F} \sum_{j \sim i} \frac{y_j}{v_{j,F}} \\
&= \sum_{l \in F} v_{l,F} \frac{y_l}{v_{l,F}} \\
&= \sum_{l \in F} y_l = \mathbf{Y}.
\end{aligned}$$

The last step follows because, as long as there are no false positive reports, in a census of F , each unit j appears once for each time it is visible to F ; that is, $v_{j,F} = y_{F,j}^+$ (see Feehan and Salganik (2016a) for details).

Note that the derivation above reveals that the individual visibility estimator can be written as a Horvitz-Thompson estimator (Sirken 1970; Thompson 2002): to see how, define $z_i = \sum_{j \sim i} \frac{y_j}{v_{j,F}}$ for all $i \in F$. The individual visibility estimator in Equation 7 then becomes $\widehat{Y} = \sum_{i \in s} w_i z_i$. Since $\widehat{\mathbf{Y}}^{\text{ind}}$ can be written as a Horvitz-Thompson estimator, Result A.1 shows that unbiasedness implies consistency. □

Good references for individual visibility estimators include Sirken (1970) and Lavalley (2007).

Aggregate visibility estimators

Definition A.2. The **aggregate visibility** estimator for a total $\mathbf{Y} = \sum_{i \in U} y_i$ is defined to be

$$\widehat{Y} = \frac{\sum_{i \in s} w_i \sum_{j \sim i} y_j}{\widehat{v}_{Y,F}},$$

where $\bar{v}_{Y,F} = N_F^{-1} \sum_{l \in U} v_{l,F}$ is the average of the individual visibilities of each person who could be reported about in a census of F .

The network scale-up estimator and related approaches are examples of estimators based on the idea of aggregate visibility (Bernard et al. 1989; Feehan and Salganik 2016a).

Result A.4. *Suppose that there are no false positive reports. Then the aggregate visibility estimator is consistent and essentially unbiased for the total \mathbf{Y} .*

Proof. See Feehan (2015) or Feehan and Salganik (2016a). □

Result A.5. Let \widehat{y}_{F,D_α} be a consistent and unbiased estimator for y_{F,D_α} , the total reported connections from F to siblings who died in group α . Let \widehat{y}_{F,N_α} be a consistent and unbiased estimator for y_{F,N_α} , the total number reported connections from F to siblings in group α (whether or not they have died). Suppose that reports about siblings in D_α or N_α are accurate in aggregate, so that $y_{F,D_\alpha} = d_{F,D_\alpha}$ and $y_{F,N_\alpha} = d_{F,N_\alpha}$. Finally, suppose that, on average, people who die have the same number of siblings as people who survive ($\bar{d}_{D_\alpha,F} = \bar{d}_{N_\alpha,F}$). Then

$$\widehat{M}^V = \frac{\widehat{y}_{F,D_\alpha}}{\widehat{y}_{F,N_\alpha}}$$

is consistent and essentially unbiased for $M_\alpha^V = \frac{D_\alpha}{N_\alpha}$, where the exposure N_α is approximated by the number of people, living or dead, in group α .

Proof. By Result A.2, \widehat{M}^V is consistent and essentially unbiased for the estimand $\frac{y_{F,D_\alpha}}{y_{F,N_\alpha}}$. Using the fact that reports are accurate in aggregate, we have

$$\frac{y_{F,D_\alpha}}{y_{F,N_\alpha}} = \frac{d_{F,D_\alpha}}{d_{F,N_\alpha}}.$$

Next, using the fact that the sibling relation is symmetric (so that $d_{A,B} = d_{B,A}$), and re-writing the total degree in terms of averages, we obtain

$$\frac{d_{F,D_\alpha}}{d_{F,N_\alpha}} = \frac{\bar{d}_{D_\alpha,F} N_{D_\alpha}}{\bar{d}_{N_\alpha,F} N_\alpha}.$$

Finally, using the fact that on average people who die have the same number of siblings as people who survive, we get

$$\frac{\bar{d}_{D_\alpha,F} N_{D_\alpha}}{\bar{d}_{N_\alpha,F} N_\alpha} = \frac{N_{D_\alpha}}{N_\alpha} = M_\alpha^V.$$

□

To recap, Result A.5 shows that using an aggregate visibility estimator for the visible death rates relies on two particularly important types of condition: first, it requires that reports about siblings be accurate *in aggregate*. Thus, Result A.5 does not require that each individual report about a sibling be exactly correct; instead, the Result requires that any error in one respondent's reports must be cancelled out when all respondents' reports are combined. In fact, as the sensitivity framework in Appendix B reveals, the condition that reports be accurate in aggregate is stronger than is actually necessary: it turns out reporting errors for the numerator and denominator of \widehat{M}^V in Result A.5 can cancel one another out.

Result A.5 also relies upon the condition that, on average, people who die have the same number of siblings as people who survive ($\bar{d}_{D_\alpha,F} = \bar{d}_{N_\alpha,F}$). This condition captures the concern from the literature about difference in death rates by sibship size. Again, the sensitivity analysis in Appendix B shows that there are scenarios in which this condition can be relaxed.

B Sensitivity framework

In Appendix A, we showed that under certain conditions, sibling survival estimators are consistent and essentially unbiased. In practice, researchers may suspect that some of the conditions the estimators rely upon may not hold. Therefore, in this section of the Appendix, we develop a sensitivity framework that researchers can use to understand how sibling survival estimates are sensitive to the different conditions.

First, we introduce some background that will be helpful in understanding the sensitivity of the estimators to invisible deaths. We then analyze the sensitivity of the estimator to invisible deaths. Next, we analyze the sensitivity of the estimator to reporting errors and other structural factors. Finally, we combine these two analyses into a universal framework for understanding the sensitivity of sibling survival estimates.

B.1 Background: aggregating death rates across groups

In order to develop a sensitivity framework, we first need a couple of technical results. These results help us understand how death rates can be expected to aggregate when different groups are combined. This will turn out to be important to our understanding of how sibling survival estimates can be affected by invisible deaths.

Definition B.1. Let $\mathbf{x}, \mathbf{w} \in \mathbb{R}^k$ and let $x_i > 0$ and $w_i > 0$ for all i . Then the **Weighted Harmonic Mean** of the \mathbf{x} values, with weights given by the \mathbf{w} values, is

$$H[\mathbf{x}; \mathbf{w}] = \frac{\sum_{i=1}^k w_i}{\sum_{i=1}^k \frac{w_i}{x_i}}.$$

The weights can be rescaled without affecting the weighted harmonic mean, as we see in the next derivation.

Result B.1. Let $\mathbf{x}, \mathbf{w} \in \mathbb{R}^k$ and let $x_i > 0$ and $w_i > 0$ for all i . Let \mathbf{w}' be defined so that $w'_i = K w_i$ for all i and for $K > 0$. Then $H[\mathbf{x}; \mathbf{w}] = H[\mathbf{x}; \mathbf{w}']$.

Proof. This property follows directly from the definition of the weighted harmonic mean (Definition B.1):

$$H[\mathbf{x}; \mathbf{w}'] = \frac{\sum_{i=1}^k w'_i}{\sum_{i=1}^k \frac{w'_i}{x_i}} = \frac{\sum_{i=1}^k K w_i}{\sum_{i=1}^k \frac{K w_i}{x_i}} = \frac{\sum_{i=1}^k w_i}{\sum_{i=1}^k \frac{w_i}{x_i}} = H[\mathbf{x}; \mathbf{w}].$$

□

We can connect these insights about harmonic means to deepen our understanding of how death rates aggregate across groups, as Result B.2 shows:

Result B.2. Suppose two demographic groups have death rates $M_1 = \frac{D_1}{N_1}$ and $M_2 = \frac{D_2}{N_2}$, where D_1 is the number of deaths in the first group, N_1 is the person-years of exposure in the first group, D_2

is the number of deaths in the second group, and N_2 is the person-years of exposure in the second group. Now suppose we combine the two groups and treat them as one aggregate group. Then the death rate for the aggregate group is the weighted harmonic mean of the death rates in the subgroups, with weights given by the number of deaths in each subgroup:

$$M_{agg} = H[(M_1, M_2); (D_1, D_2)] = \frac{D_1 + D_2}{N_1 + N_2}.$$

Proof. In the combined group, the total exposure is $N_1 + N_2$ and the total number of deaths is $D_1 + D_2$. Thus, the combined death rate is indeed $M_{agg} = \frac{D_1 + D_2}{N_1 + N_2}$. It remains to show that this is the weighted harmonic mean. By the definition of the weighted harmonic mean (Definition B.1), we have

$$\begin{aligned} H[(M_1, M_2); (D_1, D_2)] &= \frac{D_1 + D_2}{\frac{D_1}{M_1} + \frac{D_2}{M_2}} \\ &= \frac{D_1 + D_2}{\frac{D_1}{N_1} + \frac{D_2}{N_2}} = M_{agg}, \end{aligned}$$

where the last step follows because $\frac{D_1}{M_1} = D_1 \times \frac{N_1}{D_1} = N_1$.

□

Note that, as a Corollary, Result B.2 can be extended to more than two groups; in general, the death rate for an aggregation of several groups will be given by the weighted harmonic mean of the component death rates, with the weights given by the number of deaths in each component group. Below, we will see that Result B.1 provides a useful way to understand how death rates combine when groups are aggregated.

B.2 Sensitivity to invisible deaths and exposure

In order to analyze the sensitivity of sibling survival estimates to invisible deaths, we need to develop notation that can be used to distinguish between visible and invisible deaths. For a demographic group α (for example, women aged 15-25), let

$$\begin{aligned} p_{D_\alpha}^V &= \frac{D_\alpha^V}{D_\alpha^V + D_\alpha^I}, \quad \text{be the fraction of deaths that is visible;} \\ p_{N_\alpha}^V &= \frac{N_\alpha^V}{N_\alpha^V + N_\alpha^I}, \quad \text{be the fraction of exposure that is visible.} \end{aligned} \tag{8}$$

We define analogous quantities for the fraction of deaths and exposure that is invisible, $p_{D_\alpha}^I$ and $p_{N_\alpha}^I$.

Note that then

$$\begin{aligned} \frac{p_{D_\alpha}^V}{p_{N_\alpha}^V} &= \frac{D_\alpha^V}{D_\alpha^V + D_\alpha^I} \times \frac{N_\alpha^V + N_\alpha^I}{N_\alpha^V} \\ &= \frac{M_\alpha^V}{M_\alpha}. \end{aligned} \tag{9}$$

Thus, the ratio of the fraction of deaths that is visible to the fraction of exposure that is visible is equal to the ratio of the visible death rate to the total death rate.

Result B.2 shows that the total death rate M_α can be understood as a weighted harmonic mean of the invisible death rate M_α^I and the visible death rate M_α^V , where the weights are given by the number of visible and invisible deaths. We now use this insight in Result B.3, which helps us understand the formal relationship between the invisible death rate, the visible death rate, and the total death rate.

Result B.3. *Suppose that, for a demographic group α , the invisible death rate (M_α^I) and the visible death rate (M_α^V) differ by a factor of K , so that*

$$M_\alpha^I = KM_\alpha^V \quad (10)$$

for $K > 0$. Then

$$M_\alpha = M_\alpha^V \left[\frac{K}{p_{D_\alpha}^I + K(1 - p_{D_\alpha}^I)} \right], \quad (11)$$

where $p_{D_\alpha}^I$ is the proportion of deaths that is invisible.

Proof. Using the fact that M_α is the weighted harmonic mean of M_α^I and M_α^V , we find

$$\begin{aligned} M_\alpha &= \left[\frac{p_{D_\alpha}^I}{M_\alpha^I} + \frac{p_{D_\alpha}^V}{M_\alpha^V} \right]^{-1} \\ &= \left[\frac{p_{D_\alpha}^I}{KM_\alpha^V} + \frac{p_{D_\alpha}^V}{M_\alpha^V} \right]^{-1} \\ &= \left[\frac{p_{D_\alpha}^I + Kp_{D_\alpha}^V}{KM_\alpha^V} \right]^{-1} \\ &= M_\alpha^V \left[\frac{K}{p_{D_\alpha}^I + Kp_{D_\alpha}^V} \right] \\ &= M_\alpha^V \left[\frac{K}{p_{D_\alpha}^I + K(1 - p_{D_\alpha}^I)} \right]. \end{aligned}$$

□

Result B.3 reveals that there is a relationship between between K , the difference between the visible and invisible death rates, and $p_{D_\alpha}^I$, which is related to the number of invisible deaths relative to the number of visible deaths. Equation 11 shows that

- when $K = 1$, $M_\alpha^V = M_\alpha$
- when $p_{D_\alpha}^I = 0$, $p_{D_\alpha}^V = 1$ and so $M_\alpha^V = M_\alpha$

In order to further develop intuition about how large we might expect biases due to invisible deaths to be, we can investigate different scenarios.

For example, suppose that 10% of deaths are invisible, and the death rate is 20% higher among the invisible population than among the visible population. Then $K = 1.2$, $p_{D_\alpha}^I = 0.1$, and the factor in square brackets in Equation 11 is about 1.02; in other words, in this scenario, death rate estimates based on the visible population alone will be too low by about 2 percent.

Finally, it can be helpful to use Result B.3 to obtain an expression for the relative error that would follow from using the visible death rate M_α^V as an estimate of the total death rate M_α :

$$\frac{M_\alpha^V - M_\alpha}{M_\alpha} = 1 - \frac{K}{p_{D_\alpha}^V + K(1 - p_{D_\alpha}^V)}. \quad (12)$$

Figure 4 illustrates this relative error for a range of values of K and $p_{D_\alpha}^I$.

B.3 Sensitivity to reporting and structural factors

Throughout our discussion of sensitivity, we consider a census of the frame population. This allows us to avoid, for the time being, the complication and additional notation that come along with samples. Later, we will integrate these results with the results about sampling (Appendix A).

The following Fact will be useful in some of our analysis below; see, for example, Feehan and Salganik (2016a) for a derivation.

Fact B.1. *Suppose we have a finite population U of size n and that $a_i, b_i \in \mathbb{R}$ are defined for all $i \in N$. Then*

$$\sum_{i \in U} a_i b_i = N \left[\bar{a} \bar{b} + \text{cov}_U(a_i, b_i) \right],$$

where $\bar{a} = N^{-1} \sum_{i \in U} a_i$, $\bar{b} = N^{-1} \sum_{i \in U} b_i$, and $\text{cov}_U(a_i, b_i)$ is the finite population covariance of the a_i and b_i values.

Sensitivity of aggregate visibility estimators

Our sensitivity framework relies upon several quantities that are related to other network reporting analyses, including Feehan and Salganik (2016a) and Feehan et al. (2017). We call these quantities *adjustment factors*; both of these adjustment factors will be equal to 1 under ideal conditions.

The first adjustment factor is the *true positive rate* for reports about visible deaths, τ_{F, D_α} ; it is defined as

$$\tau_{F, D_\alpha} = \frac{\text{average \# times a visible death in } \alpha \text{ would be reported by someone in } F}{\text{average number of connections a visible death in } \alpha \text{ has to } F} = \frac{\bar{v}_{D_\alpha^V, F}}{\bar{d}_{D_\alpha^V, F}}. \quad (13)$$

Note that we can also write $\tau_{F, D_\alpha} = \frac{\bar{y}_{F, D_\alpha}^+}{\bar{d}_{F, D_\alpha^V}^+}$, since $v_{D_\alpha^V, F} = y_{F, D_\alpha}^+$.

The second adjustment factor is the *precision* for reports about visible deaths, η_{F, D_α} ; it is defined as

$$\eta_{F, D_\alpha} = \frac{\# \text{ of reported connections from } F \text{ to } D_\alpha \text{ that actually lead to } D_\alpha}{\# \text{ of reported connections from } F \text{ to } D_\alpha} = \frac{y_{F, D_\alpha}^+}{y_{F, D_\alpha}} \quad (14)$$

We define analogous adjustment factors τ_{F, N_α} and η_{F, N_α} for F 's reports about siblings' exposure.

By incorporating the adjustment factors introduced in the previous section, we can decompose the aggregate visibility sibling survival estimator as

$$\begin{aligned}
M_\alpha &= \frac{D_\alpha}{N_\alpha} \\
&= \frac{D_\alpha^V}{N_\alpha^V} \times \frac{p_{N_\alpha}^V}{p_{D_\alpha}^V} \\
&= \frac{D_\alpha^V}{N_\alpha^V} \times \frac{M_\alpha}{M_\alpha^V} \\
&= \frac{y_{F,D_\alpha}}{y_{F,N_\alpha}} \times \frac{\bar{d}_{N_\alpha,F}^V}{\bar{d}_{D_\alpha,F}^V} \times \frac{\eta_{F,D_\alpha}}{\eta_{F,N_\alpha}} \times \frac{\tau_{F,N_\alpha}}{\tau_{F,D_\alpha}} \times \frac{M_\alpha}{M_\alpha^V}.
\end{aligned} \tag{15}$$

Finally, to combine the expression for sensitivity to invisible siblings, let the visible and invisible death rates differ by a factor K so that $M_\alpha^I = KM_\alpha^V$. Equation 11 then tells us that

$$\frac{M_\alpha^V}{M_\alpha} = \frac{p_{D_\alpha}^V + K(1 - p_{D_\alpha}^V)}{K}. \tag{16}$$

Combining Equation 11 and Equation 15, we obtain

$$M_\alpha = \frac{y_{F,D_\alpha}}{y_{F,N_\alpha}} \times \frac{\bar{d}_{N_\alpha,F}^V}{\bar{d}_{D_\alpha,F}^V} \times \frac{\eta_{F,D_\alpha}}{\eta_{F,N_\alpha}} \times \frac{\tau_{F,N_\alpha}}{\tau_{F,D_\alpha}} \times \left[\frac{K}{p_{D_\alpha}^I + K(1 - p_{D_\alpha}^I)} \right]. \tag{17}$$

This decomposition relates the quantities we can observe or estimate from a survey—i.e., y_{F,D_α} and y_{F,N_α} —to the quantities that we actually wish to estimate—i.e., D_α and N_α . The decomposition in Equation 15 produces several factors that are the ratio of (i) an adjustment factor for deaths; and (ii) the same adjustment factor for exposure (for example, $\frac{\eta_{F,D_\alpha}}{\eta_{F,N_\alpha}}$). To the extent that reporting about deaths and reporting about exposure is similar, this is advantageous: these adjustment factors can cancel or counteract one another.

Sensitivity of individual visibility estimators

In order to understand the sensitivity of individual visibility estimators, we must define individual-level adjustment factors:

$$\tau_{i,D_\alpha} = \frac{\# \text{ times a visible death in } \alpha \text{ is reported by } i}{\# \text{ of connections } i \text{ actually has to visible deaths in } \alpha} = \frac{v_{D_\alpha,i}^V}{d_{D_\alpha,i}^V}$$

and

$$\eta_{i,D_\alpha} = \frac{\# \text{ of reported connections from } i \text{ to } D_\alpha \text{ that actually lead to } D_\alpha}{\# \text{ of reported connections from } i \text{ to } D_\alpha} = \frac{y_{i,D_\alpha}^+}{y_{i,D_\alpha}}.$$

τ_{i,D_α} and η_{i,D_α} are the individual-level analogues of the quantities we introduced in the previous section.

We also define a combined adjustment factor for each individual i :

$$\gamma_{i,D_\alpha} = \frac{\eta_{i,D_\alpha}}{\tau_{i,D_\alpha}}.$$

Using γ_{i,D_α} instead of τ_{i,D_α} and η_{i,D_α} will help simplify the expressions we derive below.

We also define analogous adjustment factors τ_{i,N_α} , η_{i,N_α} , and γ_{i,D_α} for i 's reports about siblings' exposure.

Let the average of these individual-level adjustment factors across all of the members of the frame population be

$$\bar{\tau}_{F,D_\alpha} = \frac{1}{N_F} \sum_{i \in F} \tau_{i,D_\alpha},$$

$$\bar{\eta}_{F,D_\alpha} = \frac{1}{N_F} \sum_{i \in F} \eta_{i,D_\alpha},$$

and

$$\bar{\gamma}_{F,D_\alpha} = \frac{1}{N_F} \sum_{i \in F} \gamma_{i,D_\alpha} = \frac{1}{N_F} \sum_{i \in F} \frac{\eta_{i,D_\alpha}}{\tau_{i,D_\alpha}}.$$

Finally, let

$$z_{i,D_\alpha} = \frac{d_{i,D_\alpha}}{d_{i,F}}.$$

First, we will use these adjustment factors to decompose the number of deaths, D_α . We have

$$\begin{aligned} D_\alpha &= \frac{D_\alpha^V}{p_{D_\alpha}^V} \\ &= \frac{1}{p_{D_\alpha}^V} \sum_{i \in F} \frac{y_{i,D_\alpha}^+}{v_{i,F}} \\ &= \frac{1}{p_{D_\alpha}^V} \sum_{i \in F} z_{i,D_\alpha} \gamma_{i,D_\alpha} \\ &= \frac{1}{p_{D_\alpha}^V} N_F [\bar{z}_{F,D_\alpha} \bar{\gamma}_{F,D_\alpha} + \text{cov}_F(z_{i,D_\alpha}, \gamma_{i,D_\alpha})]. \end{aligned} \tag{18}$$

The step from the second to third line follows because $v_{i,F} = d_{i,F} \tau_{i,F}$ and $y_{i,D_\alpha}^+ = y_{i,D_\alpha} \eta_{i,D_\alpha}$, and the step from the third to the fourth line follows from Fact B.1.

An analogous derivation reveals that

$$\begin{aligned}
N_\alpha &= \frac{N_\alpha^V}{p_{N_\alpha}^V} \\
&= \frac{1}{p_{N_\alpha}^V} N_F [\bar{z}_{F,N_\alpha} \bar{\gamma}_{F,N_\alpha} + \text{cov}_F(z_{i,N_\alpha}, \gamma_{i,N_\alpha})].
\end{aligned} \tag{19}$$

Combining Equation 18 and Equation 19, we have

$$\begin{aligned}
M_\alpha &= \frac{D_\alpha}{N_\alpha} = \frac{N_F [\bar{z}_{F,D_\alpha} \bar{\gamma}_{F,D_\alpha} + \text{cov}_F(z_{i,D_\alpha}, \gamma_{i,D_\alpha})] \frac{1}{p_{D_\alpha}^V}}{N_F [\bar{z}_{F,N_\alpha} \bar{\gamma}_{F,N_\alpha} + \text{cov}_F(z_{i,N_\alpha}, \gamma_{i,N_\alpha})] \frac{1}{p_{N_\alpha}^V}} \\
&= \frac{[\bar{z}_{F,D_\alpha} \bar{\gamma}_{F,D_\alpha} + \text{cov}_F(z_{i,D_\alpha}, \gamma_{i,D_\alpha})]}{[\bar{z}_{F,N_\alpha} \bar{\gamma}_{F,N_\alpha} + \text{cov}_F(z_{i,N_\alpha}, \gamma_{i,N_\alpha})]} \times \frac{p_{N_\alpha}^V}{p_{D_\alpha}^V} \\
&= \frac{[\bar{z}_{F,D_\alpha} \bar{\gamma}_{F,D_\alpha} + \text{cov}_F(z_{i,D_\alpha}, \gamma_{i,D_\alpha})]}{[\bar{z}_{F,N_\alpha} \bar{\gamma}_{F,N_\alpha} + \text{cov}_F(z_{i,N_\alpha}, \gamma_{i,N_\alpha})]} \times \frac{M_\alpha}{M_\alpha^V}.
\end{aligned} \tag{20}$$

In the last step, we use Equation 9, which showed that $p_{N_\alpha}^V/p_{D_\alpha}^V = M_\alpha/M_\alpha^V$.

Following our analysis of the aggregate visibility estimator above, we can substitute Equation 16 into Equation 20 to obtain

$$M_\alpha = \frac{[\bar{z}_{F,D_\alpha} \bar{\gamma}_{F,D_\alpha} + \text{cov}_F(z_{i,D_\alpha}, \gamma_{i,D_\alpha})]}{[\bar{z}_{F,N_\alpha} \bar{\gamma}_{F,N_\alpha} + \text{cov}_F(z_{i,N_\alpha}, \gamma_{i,N_\alpha})]} \times \left[\frac{K}{p_{D_\alpha}^I + K(1 - p_{D_\alpha}^I)} \right].$$

C Relationship to existing estimators

There are basically three main direct sibling survival estimators. The first one is often called the *naive* estimator and is simply given by the ratio of the reported deaths D_α^V by the reported births N_α^V , including the respondent:

$$M_\alpha = M_\alpha^V = \frac{D_\alpha^V}{N_\alpha^V} \tag{21}$$

This approach has the obvious limitation that respondents are, by definition, alive, and including them in the denominator produces a downward bias in the estimate.

Furthermore, this estimator can be biased through two other different mechanisms: the relationship between fertility and mortality; heterogeneity in mortality itself. In the first case, selection bias arises from the fact large sibships are over represented in the sample, since each sibling is reported about multiple times. In other words, since the number of death and surviving siblings is equal to the number of surviving sibling to report them, we are more likely to learn about the mortality experience of sibships that have more survivors when the survey is conducted. It is clear that this estimator gives more weight to mortality experience of larger sibships, but whether large sibships have higher or lower mortality is still under debate in the literature. Gakidou and King (2006) argue that large sibships tend to have higher mortality, whereas B. Masquelier (2013) states that

this positive correlation can arise from cohort trends in mortality and fertility, and, in fact, this correlation may be negative, although this could also be related to reporting problems.

The second mechanism that could bias the estimator is heterogeneity in mortality itself. High mortality population groups, not only in regard to sibship size, would be underrepresented in the sample, which would inevitably lead to an underestimation of mortality calculated by the sibling survival method [B. Masquelier (2013); Zaba and David 1996].

To address some of these limitations, the vast majority of studies that have used the direct sibling method (B. Masquelier 2013), including the DHS reports (S. O. Rutstein and Guillermo Rojas 2006), calculates mortality rates based on the deaths D_α^V and exposure reported by the survey respondents, but excluding the respondents from the denominator, represented here by B_α^V :

$$M_\alpha = M_\alpha^V = \frac{D_\alpha^V}{B_\alpha^V} \quad (22)$$

If information used to calculate mortality rates through this approach come from surveys such as the DHS, both the numerator and the denominator are also weighted by the respondent's sample weight.

This approach is supported by the findings of Trussell and Rodriguez (1990) study, which shows that: i) mortality estimates calculated using information on the total number of siblings and dead siblings (excluding the respondent) are unbiased for each sibship size, assuming that they have the same probability of dying p ; ii) if mortality is unrelated to sibship size, then p is the same for every woman regardless of sibship size and the aggregate mortality estimate is also unbiased.

This estimator will still be biased if mortality differs by sibship size or, more generally, if there is any kind of mortality heterogeneity that challenges the assumption that the probability of dying p is the same for all women.

Gakidou and King (2006) developed a weighting scheme that corrects for these biases in two steps. First, they propose to reweigh the estimator to give less weight to sibships where many siblings survived, by computing family-level weights of the form $W_i = \frac{B_i}{S_i}$, where B_i is the number of siblings at the start of the observation period and S_i is the number of surviving siblings at the time of the survey. Notice that in this approach respondents are counted as well. A weighted average of the proportion of dead siblings reported by each survivor, using W_i as weights, will give the true proportion for sibships with at least one survivor. When applied to individual-level data files, this weight takes the form $\frac{1}{S_i}$.

$$M_\alpha = \frac{\sum_{i=1}^n M_i W_i}{\sum_{i=1}^n W_i} \quad (23)$$

where M_i are the proportion of respondent i 's sibship that has died, $\frac{\sum_{j=1}^N M_j}{N}$.

The second adjustment Gakidou and King (2006) propose is intended to estimate the number of deaths in families with no survivors at the time of the survey, denoted by ξ , and should be added to both numerator and denominator of Equation 23:

$$M_\alpha = \frac{\sum_{i=1}^n M_i W_i + \xi}{\sum_{i=1}^n W_i + \xi} \quad (24)$$

The authors propose to estimate ξ by extrapolating to sibship size equals zero the results of a regression model that relates sibship size and the proportion of deaths (Gakidou and King 2006, B. Masquelier (2013)).

Although the weighting applied in the first step is an exact correction, there remain some questions about the accuracy of extrapolating to zero survivors sibships.

In addition no reinforce that the adjustment for the zero-survivor is not satisfactory, B. Masquelier (2013) also states that Gakidou and King weights have not been applied correctly in practice (Obermeyer et al. 2010), arguing that the standard approach should be adopted instead, since his microsimulations show that the selection bias problems appear to cancel out.