Using sampled social network data to estimate adult death rates

Dennis M. Feehan
UC Berkeley

Joint with:
Matthew J. Salganik (Princeton), Mary Mahy (UNAIDS),
Aline Umubyeyi (U. of Rwanda), Wolfgang Hladik (CDC)

Counting the dead is one of the world's best investments to reduce premature mortality

Prabhat Jha[1]

over the
cost giv
nology e
would t
ties, app
in releva
century
countries
the worl

Who Counts? 1

A scandal of invisibility: making everyone count by counting everyone

*Philip W Setel, Sarah B Macfarlane, Simon Szreter, Lene Mikkelsen, Prabhat Jha, Susan Stout, Carla AbouZahr, on behalf of the Monitoring of Vital Events (MoVE) writing group**

Most people in Africa and Asia are born and die without leaving a trace in any legal record or official statistic. Absence of reliable data for births, deaths, and causes of death are at the root of this scandal of invisibility, which renders most situation has arisen because, in some over the past 30 years. Net of debt relief, he weakness in recording vital statistics, on either mortality or poverty reduction. at enable progress towards Millennium ed, not only modelled. Vital statistics are ion has a dual function, both statistical not be overcome quickly, although new refocus attention on civil registration. in developing countries the expectation y sharing information and methods to appropriate use of complementary and

Lancet 2007; 370: 1569–77
Published Online October 29, 2007 DOI:10.1016/S0140-6736(07)61307-5

See Comment pages 1526 and 1527
See Perspectives page 1537

This is the first in a Series of four papers about the importance of collecting data for health development

*Members listed at the end of article

MEASURE Evaluation, Carolina Population Center and

Advance Access publication 28 July 2004
International Journal of Epidemiology 2004;33:710–717
doi:10.1093/ije/dyh099

REVIEW

Adult mortality: time for a reappraisal

Emmanuela Gakidou,[1] Margaret Hogan[1] and Alan D Lopez[2]

Accepted 19 December 2003

ny countries, lit
cessary to deve
available res
ed in surveys h
sibling survival
ehensively anal

State of the World's Data on Births, Deaths and Causes of Death, 2010

- 40 million births and 40 million deaths unrecorded each year
- No reliable data on causes of death for two-thirds of the world's population

HMN World Health Organization

per uses sibling
to generate est
lt mortality 45q
...en compared wi

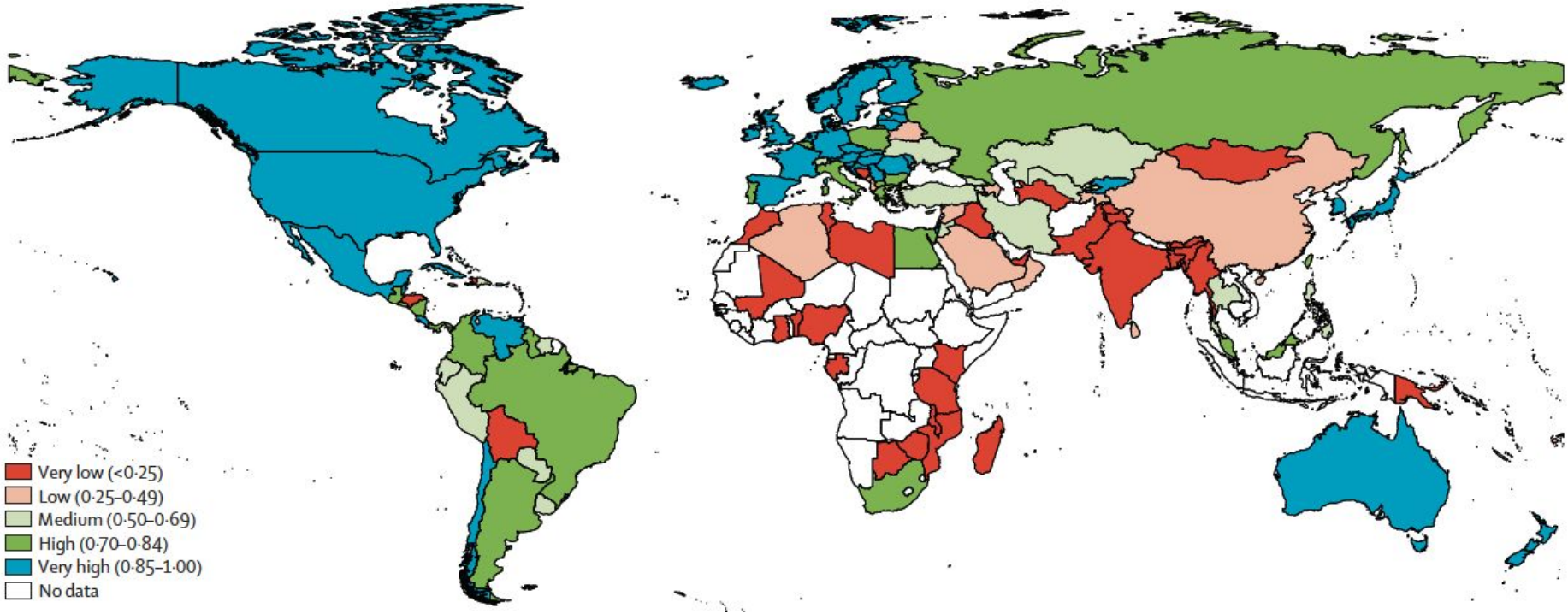
What do we really know about adult mortality worldwide?

Published Online April 30, 2010 DOI:10.1016/S0140-6736(10)60629-0

See Articles page 1704

As child mortality continues to decline globally, more children survive to adulthood, and it is imperative to prevent premature deaths in adults. But what do we really know about how many adults aged between 15 and 60 years—the most healthy and productive age group in our society—are dying today?

Despite the growing interest in the health of adults over the past two decades since the publication of the *World development report 1993: investing in health*,[1] a rigorous assessment of the levels and trends of adult mortality has been neglected, partly due to the huge measurement challenge (ie, adult deaths are rare events



Source: Mikkelsen et al (2015), *Lancet*

The challenge: measuring mortality on a survey

Adult deaths are challenging to measure with a survey

- We can’t sample and interview dead people
- Death is a rare event

The challenge: measuring mortality on a survey

Adult deaths are challenging to measure with a survey

- We can’t sample and interview dead people
- Death is a rare event

We’ll study two different approaches to overcoming these challenges

Sibling survival

Sibling survival method: ask respondents to list their siblings, when they were born, and whether or not they died

Sibling survival

Sibling survival method: ask respondents to list their siblings, when they were born, and whether or not they died

Good because

- We learn about people we don’t interview
- We learn about more than one person from each respondent

Sibling survival

But there are also challenges with sibling survival

- We don’t learn about enough siblings per interview to produce precise death rate estimates
- Not embedded in a statistical framework, leading to considerable disagreement about how data should be analyzed

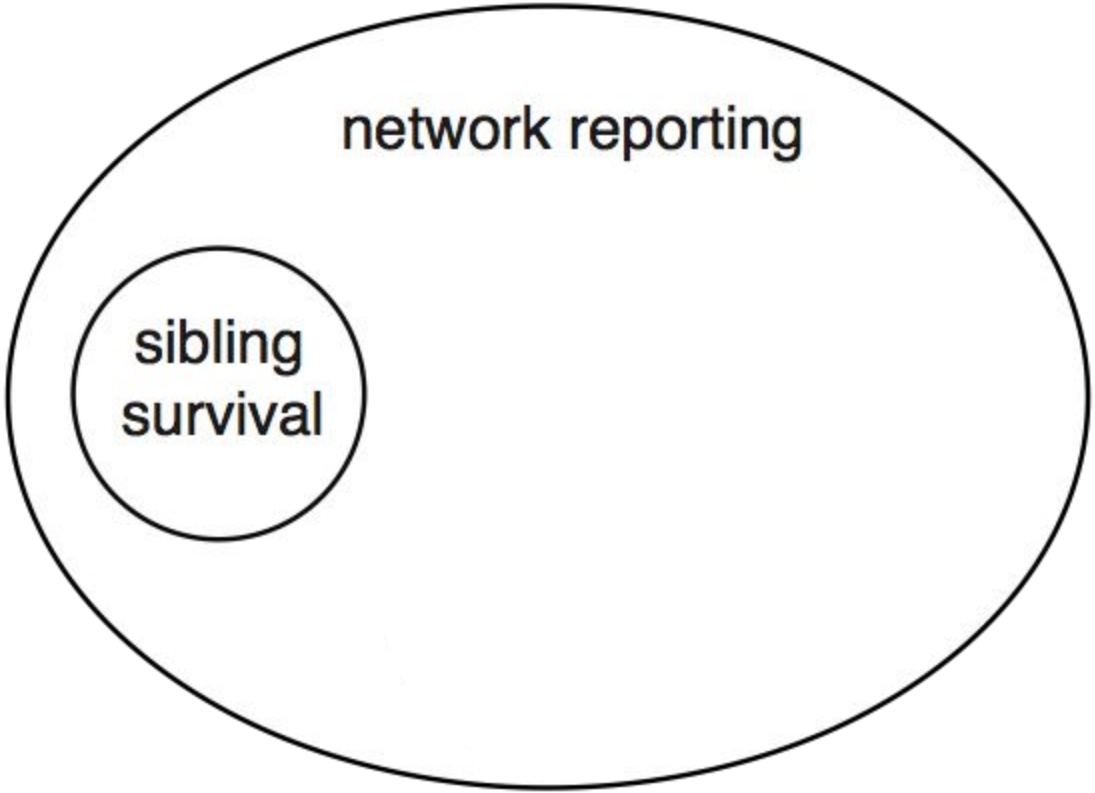
Sibling survival

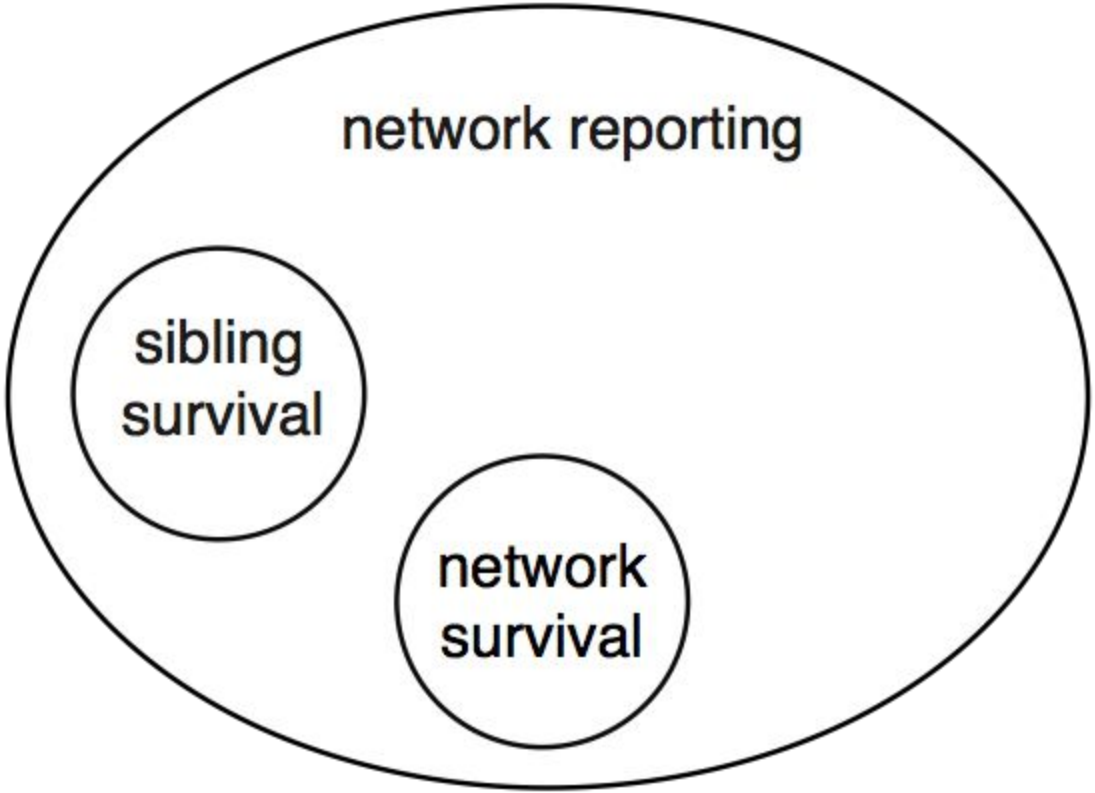
But there are also challenges with sibling survival

- We don’t learn about enough siblings per interview to produce precise death rate estimates
- Not embedded in a statistical framework, leading to considerable disagreement about how data should be analyzed

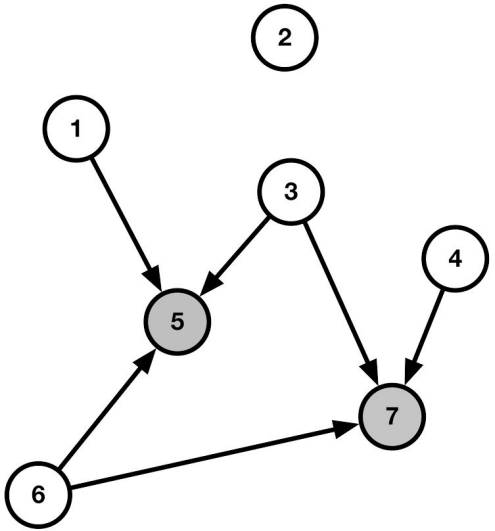
What about going beyond sibship and asking about other types of social relationships?

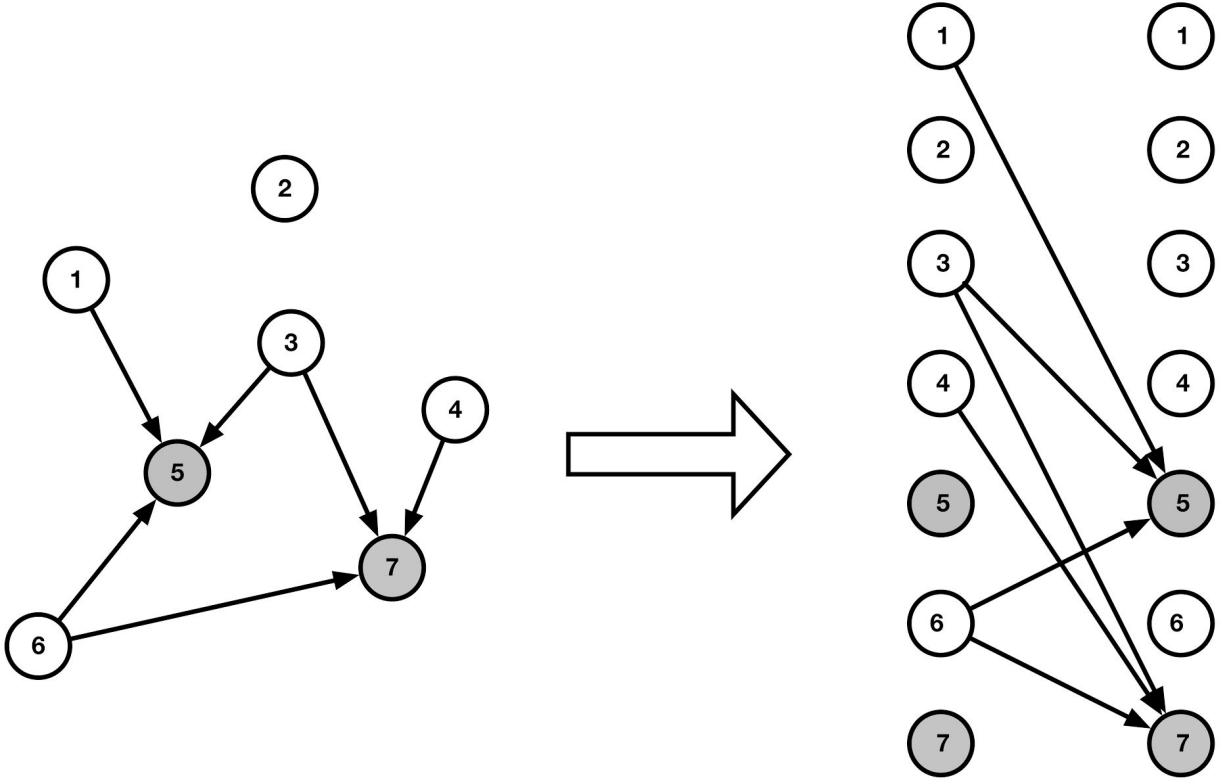

sibling
survival

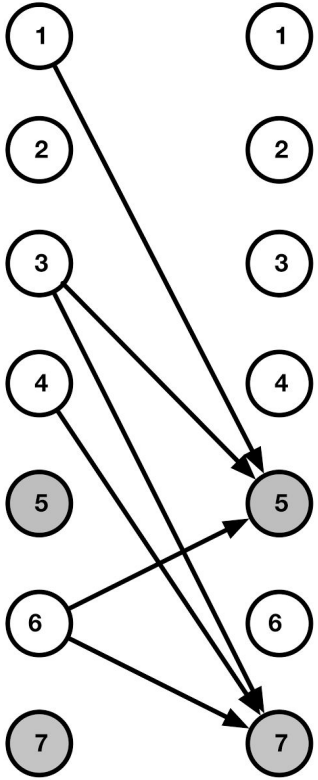

network reporting
sibling
survival

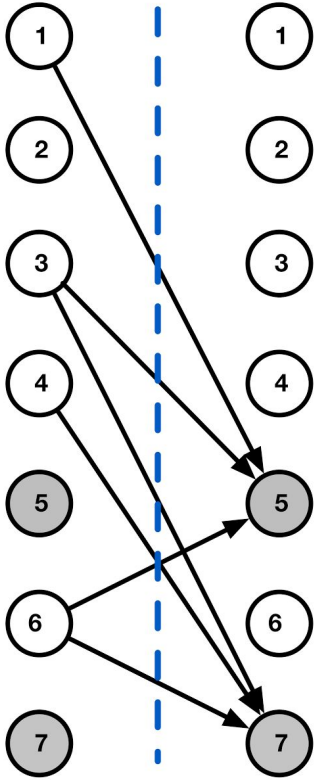

network reporting
sibling survival
network survival

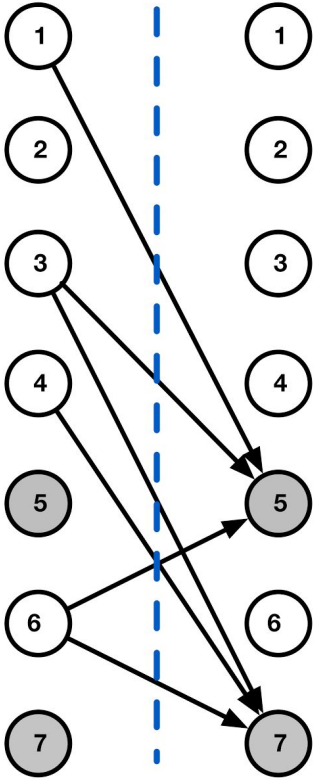
New approach:
network survival method


2
1
3
4
5
7
6


1
2
3
4
5
6
7
1
2
3
4
5
6
7
1
2
3
4
5
6
7


1
2
3
4
5
6
7
1
2
3
4
5
6
7

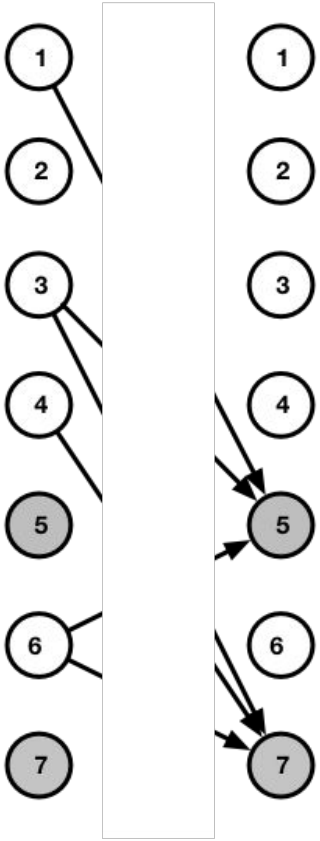

1
2
3
4
5
6
7
1
2
3
4
5
6
7



$$\text{total out-reports} = \text{total in-reports}$$

$$\text{total out-reports} = \text{number of deaths} \times \text{in-reports per death}$$

$$\text{number of deaths} = \frac{\text{total out-reports}}{\text{in-reports per death}}$$



$$\text{total out-reports} = \text{total in-reports}$$

$$\text{total out-reports} = \text{number of deaths} \times \text{in-reports per death}$$

$$\text{number of deaths} = \frac{\text{total out-reports}}{\text{in-reports per death}}$$

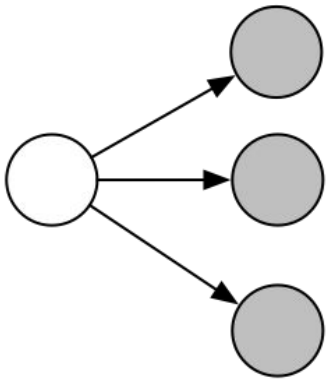
Out-reports: Deaths in the network

$$\text{number of deaths} = \frac{\text{total out-reports}}{\text{in-reports per death}}$$

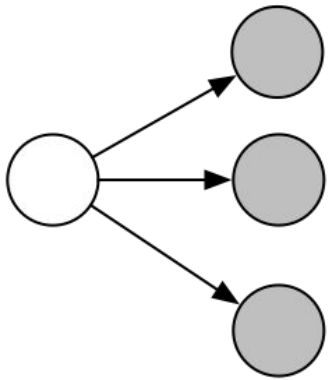
Out-reports: Deaths in the network

How many people do you know who died in the last year?

Out-reports: Deaths in the network



Out-reports: Deaths in the network



sex	age
m	39
f	55
m	67

Visibility: Number of in-reports per death

$$\text{number of deaths} = \frac{\text{total out-reports}}{\text{in-reports per death}}$$

Visibility: Number of in-reports per death

Lots of potential strategies for estimating visibility.

Visibility: Number of in-reports per death

Lots of potential strategies for estimating visibility.

Very simple way:

- Use the network sizes of our survey respondents to estimate the visibility of the people who died

Visibility: Number of in-reports per death

Lots of potential strategies for estimating visibility.

Very simple way:

- Use the network sizes of our survey respondents to estimate the visibility of the people who died

For example, if our survey results tell us that female respondents aged 50-59 have an average network size of 200

… then we assume that women aged 50-59 who died have an average visibility of 200.

Visibility: Number of in-reports per death

Lots of potential strategies for estimating visibility.

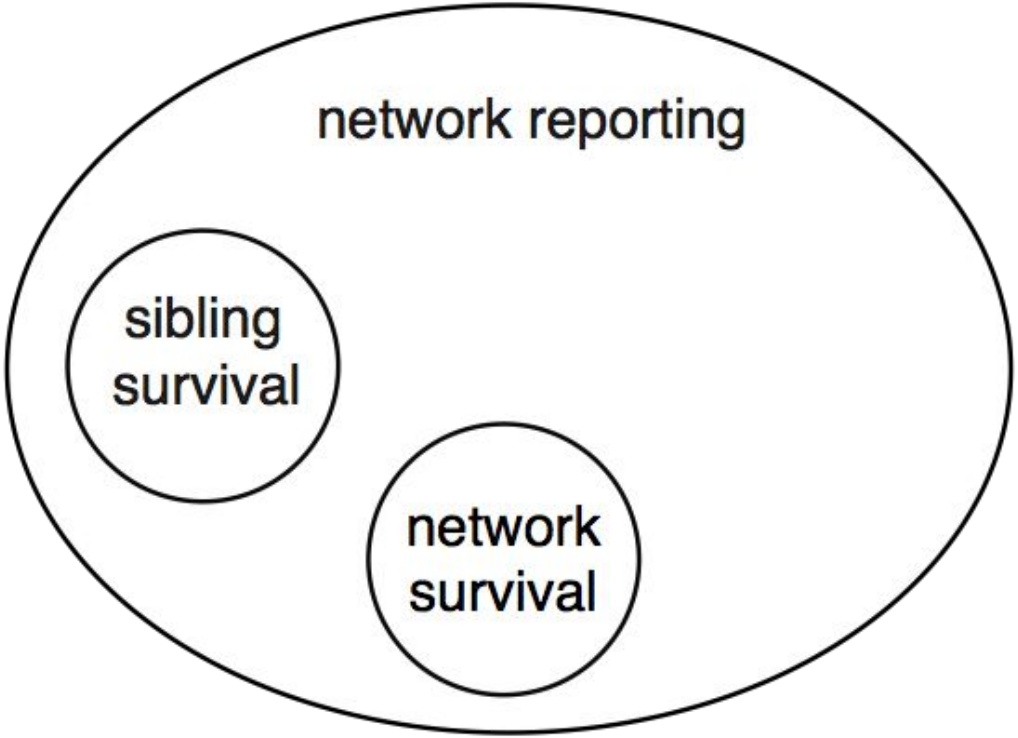
Very simple way:

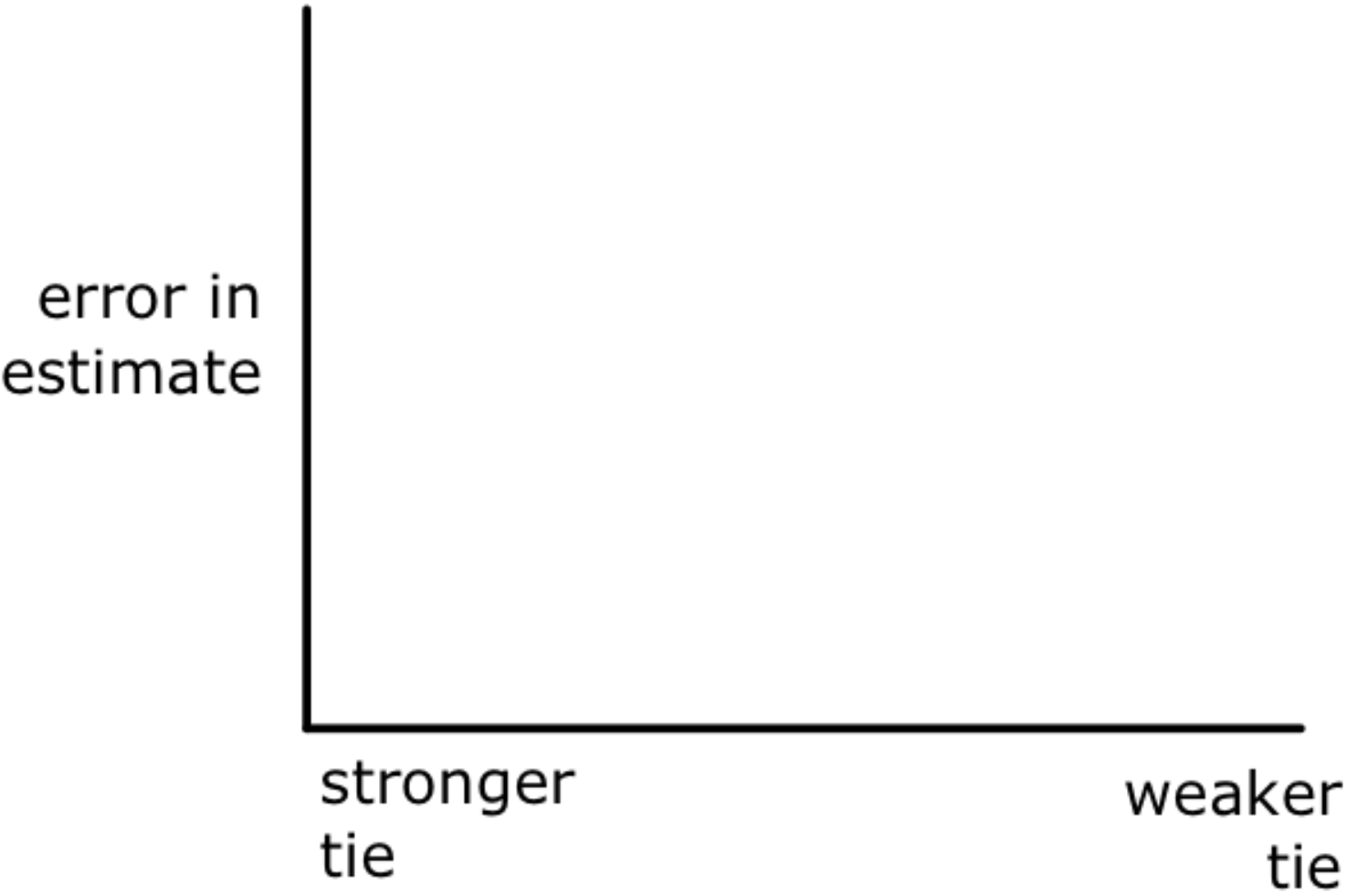
- Use the network sizes of our survey respondents to estimate the visibility of the people who died

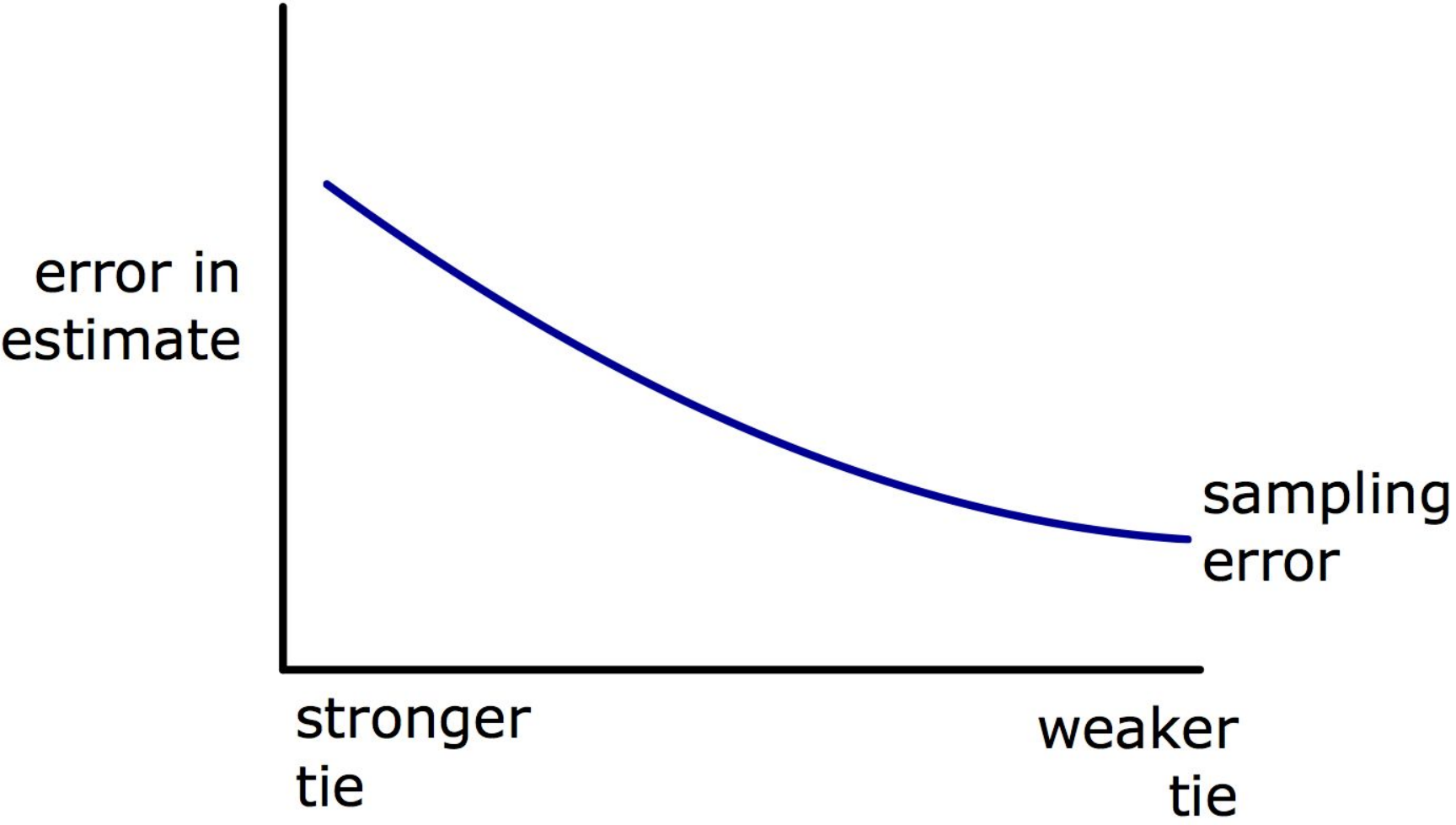
Will work well if

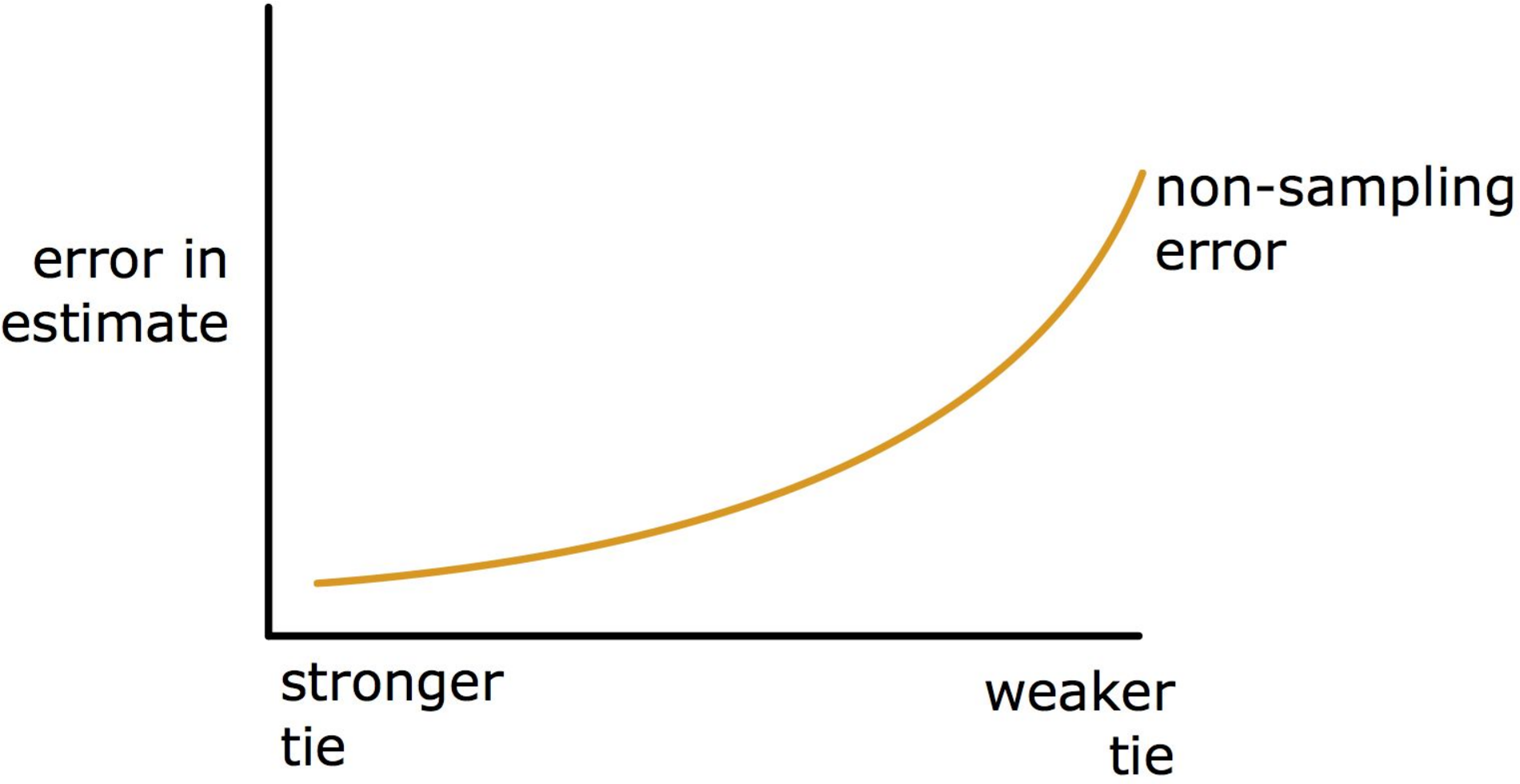
- Reports are accurate
- People are aware of which network members died
- People who died have networks that are similar to the people who respond to the survey

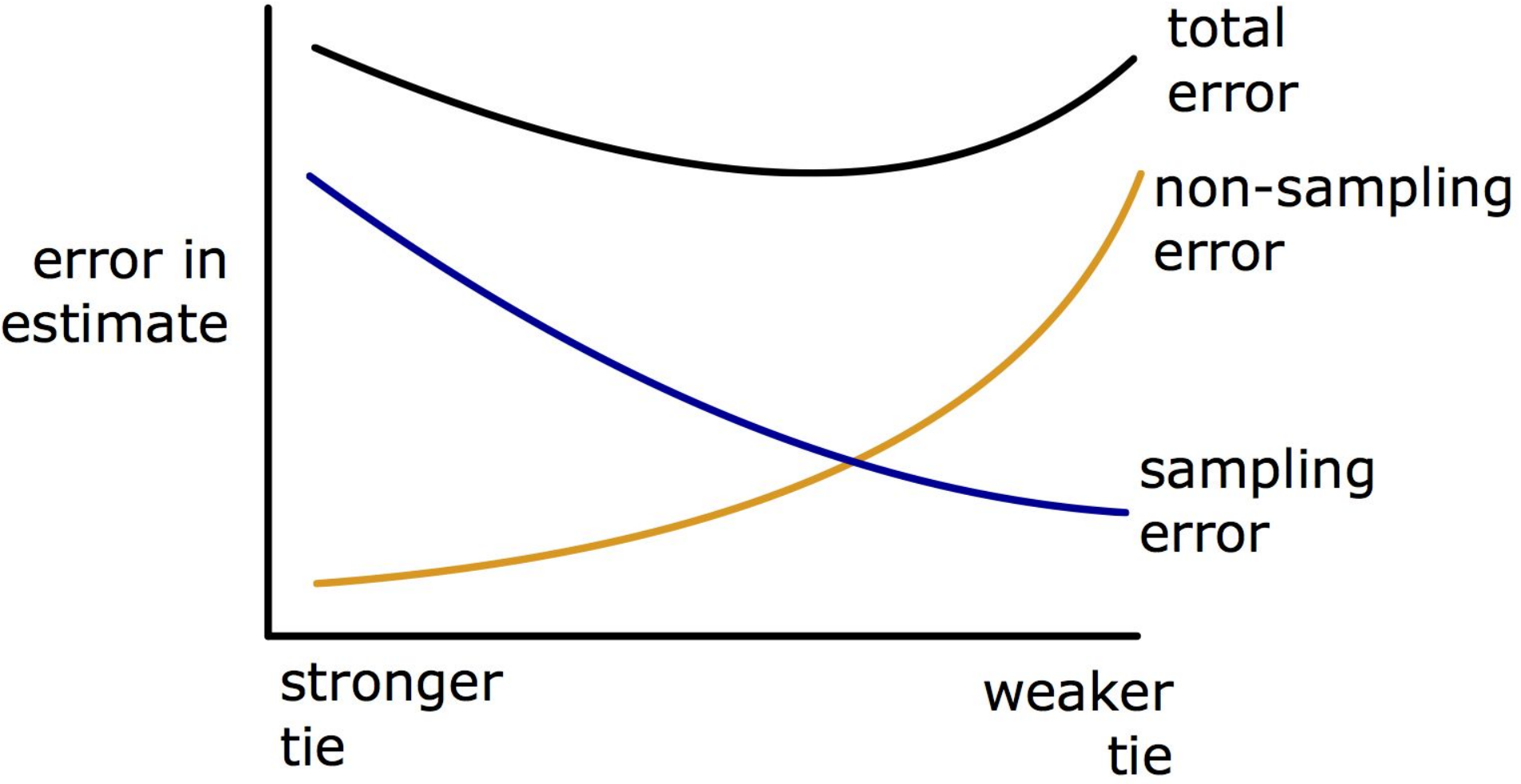
Framework for tie definitions

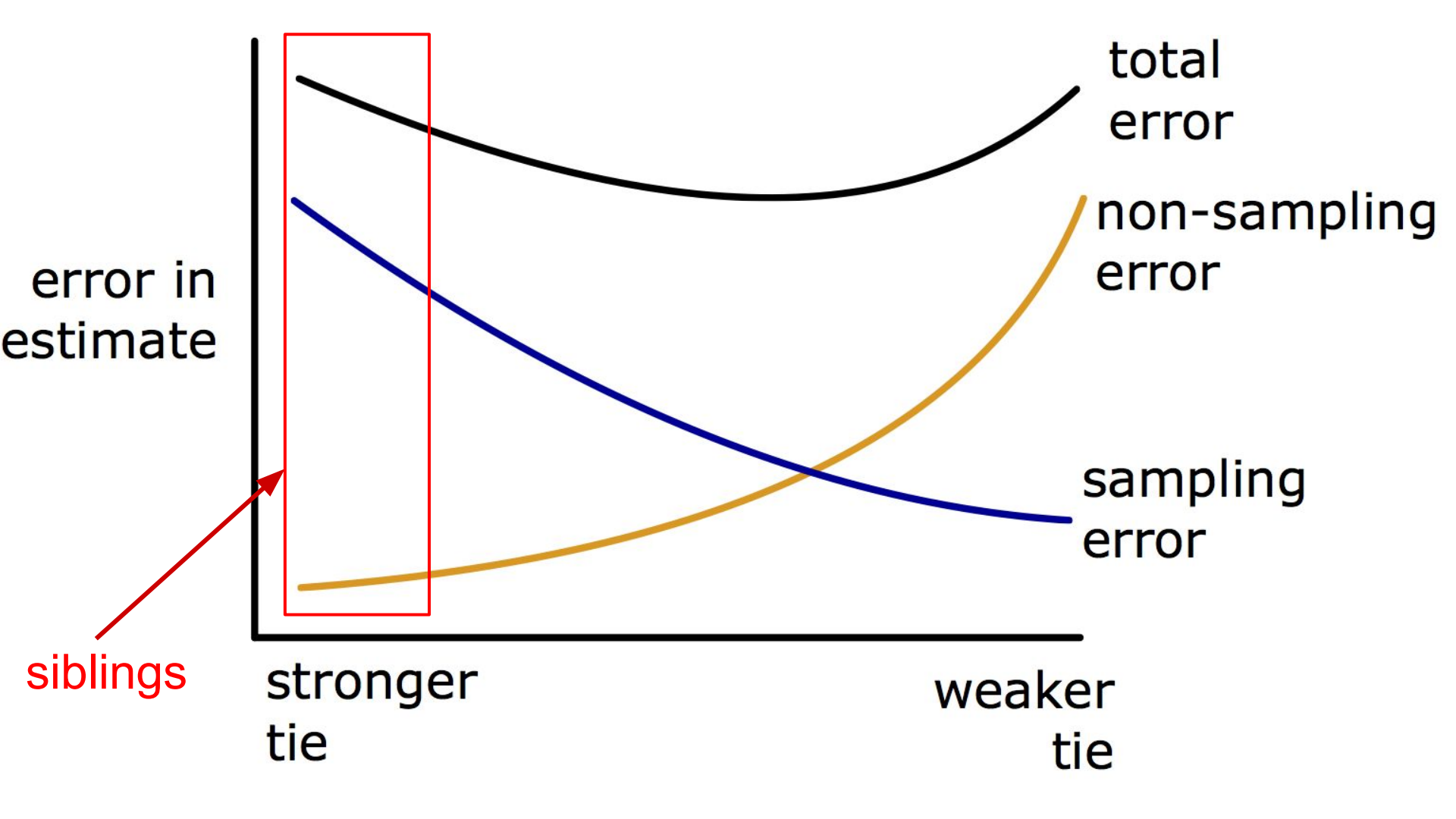


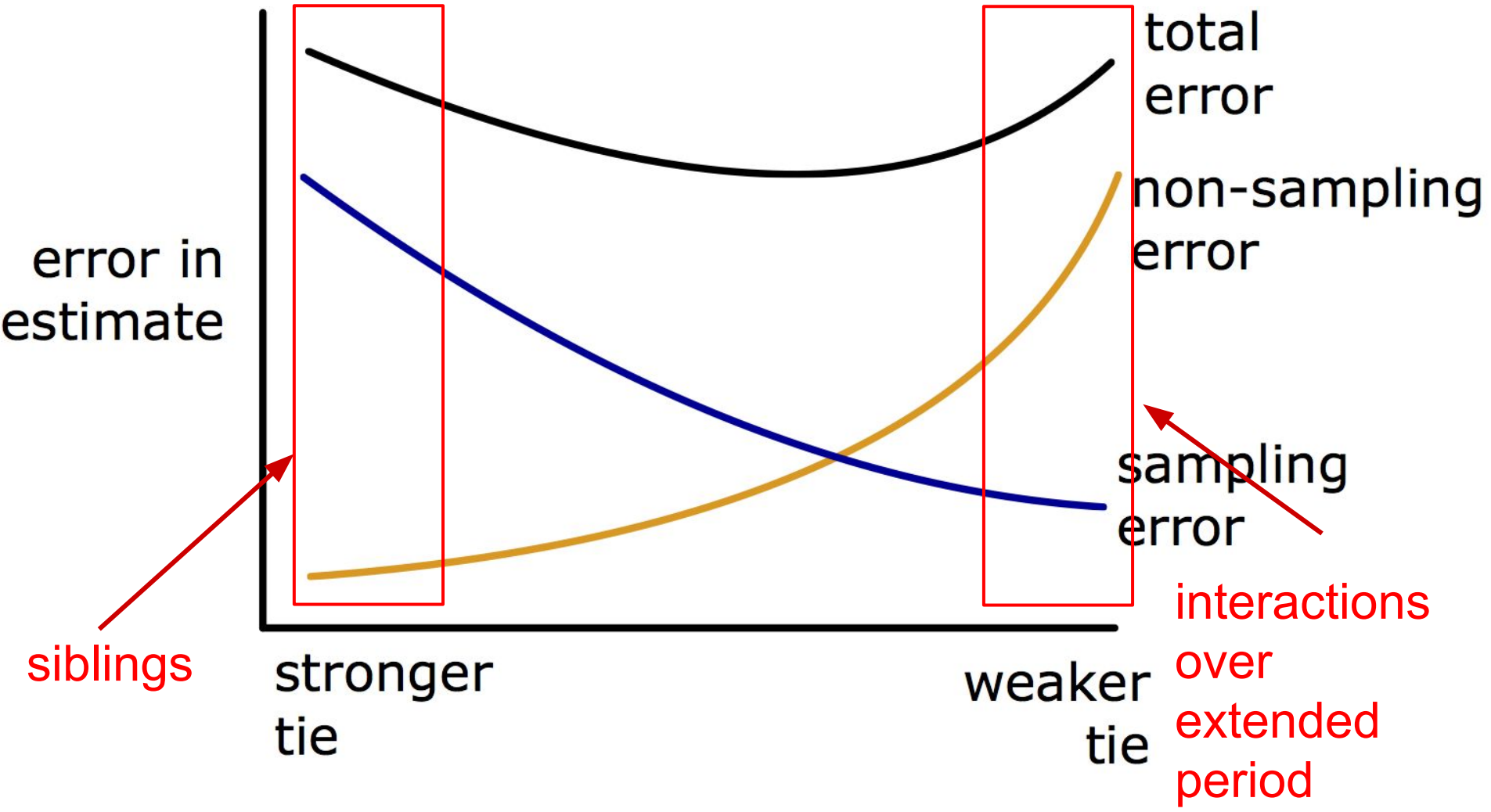

error in
estimate
stronger
tie
weaker
tie


error in estimate
sampling error
stronger tie
weaker tie


error in estimate
non-sampling error
stronger tie
weaker tie


total
error
non-sampling
error
error in
estimate
sampling
error
stronger
tie
weaker
tie


total
error
non-sampling
error
error in
estimate
sampling
error
siblings
stronger
tie
weaker
tie


total error
non-sampling error
error in estimate
sampling error
siblings
stronger tie
weaker tie
interactions over extended period

Data: household survey in Rwanda

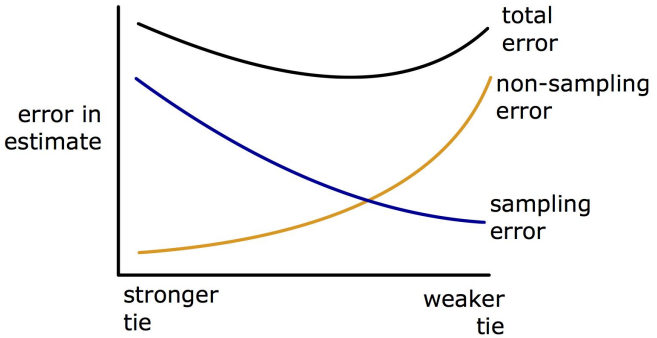
Map source: Wikipedia

Data: household survey in Rwanda

- Intended to mimic a Demographic and Health Survey
- Stratified, two-stage cluster sample of approximately 5,000 Rwandans aged 15 and over (oversampled Kigali)

Data: household survey in Rwanda

- Intended to mimic a Demographic and Health Survey
- Stratified, two-stage cluster sample of approximately 5,000 Rwandans aged 15 and over (oversampled Kigali)
- Experiment that tested questions about two types of networks - I won't have time to explain this in detail today



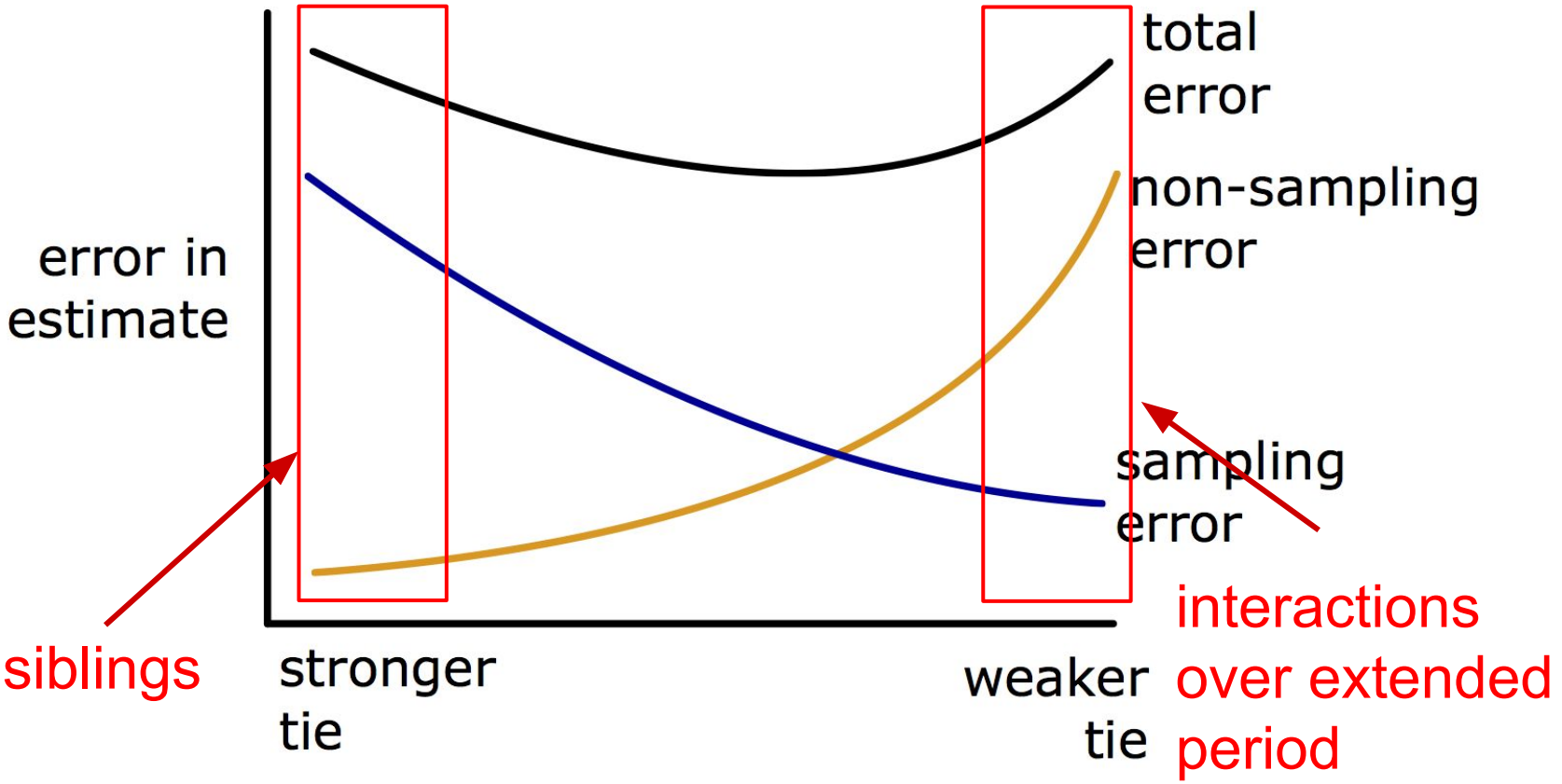
Data: Rwanda DHS

Sibling method results from Rwanda 2010-11 DHS

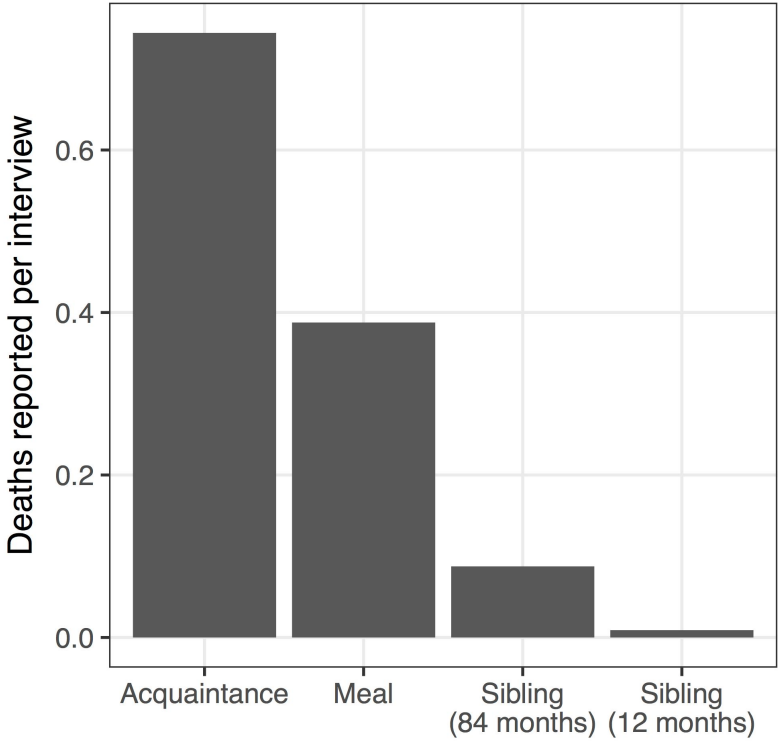
- Based on interviews with 13,761 women who were asked to report on their siblings
- The sibling estimates of death rates are based on the 7-year period before the interviews
  (the network results are for 1 year before the interview)

Deaths per interview

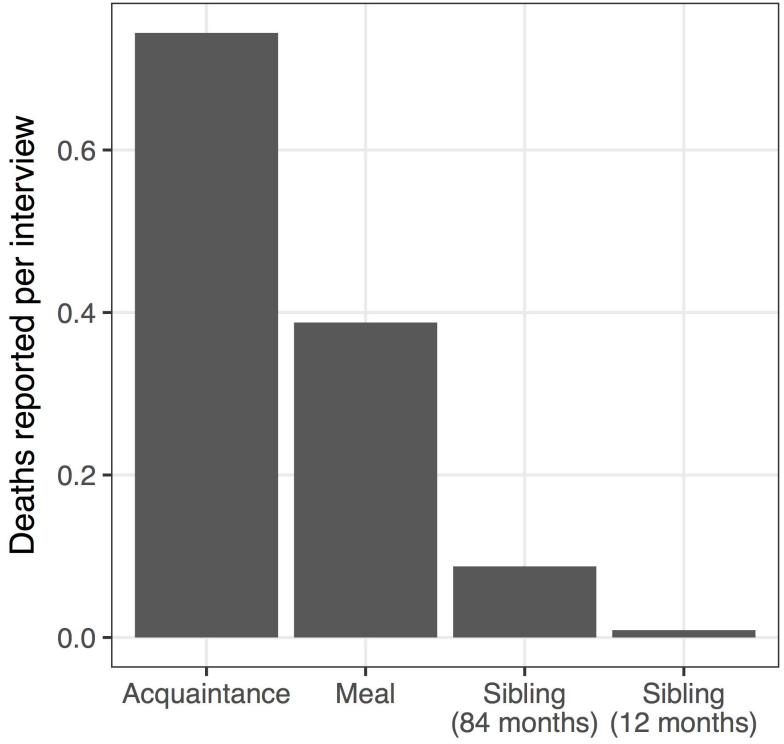
Deaths per interview



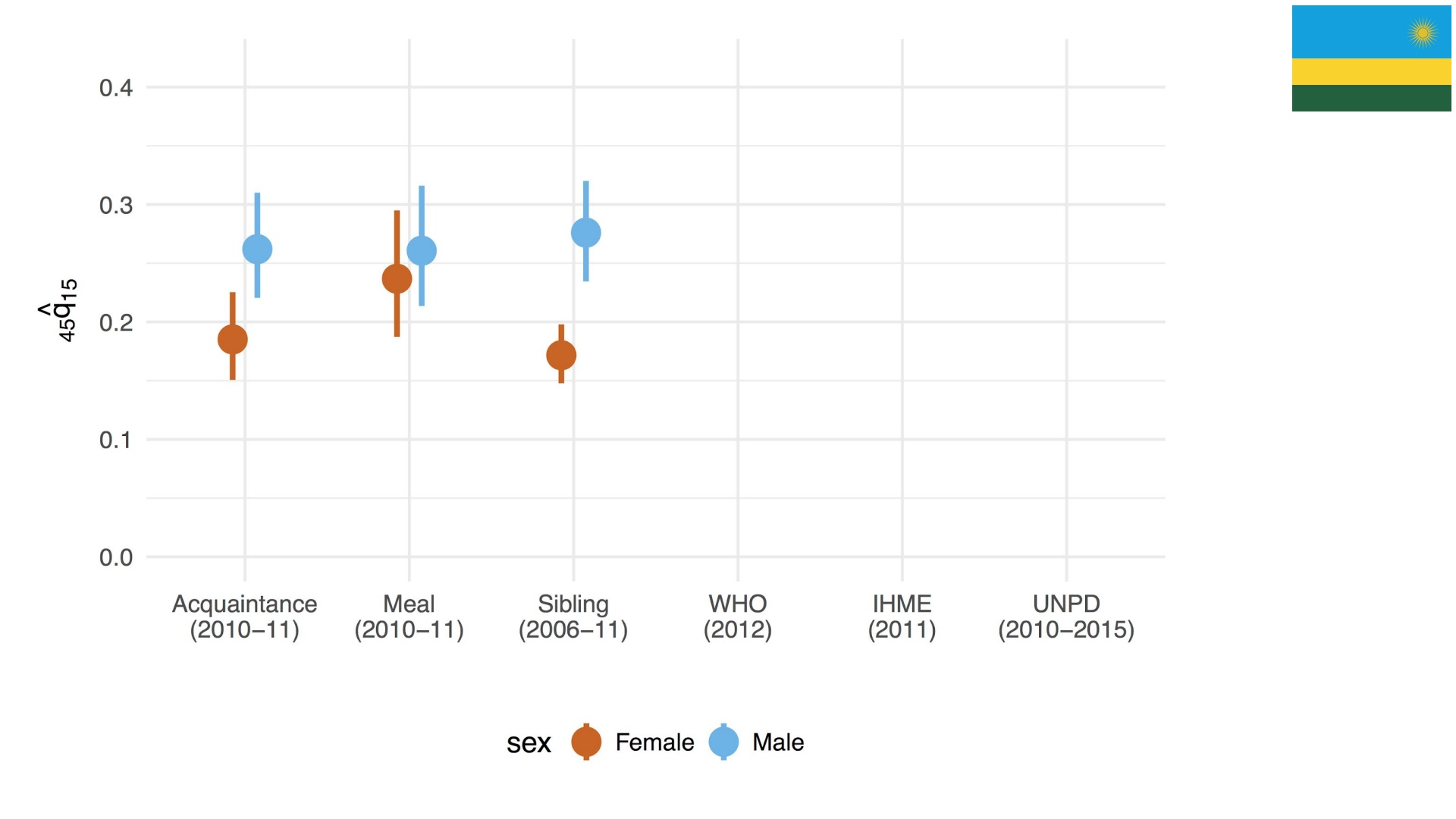
Deaths per interview

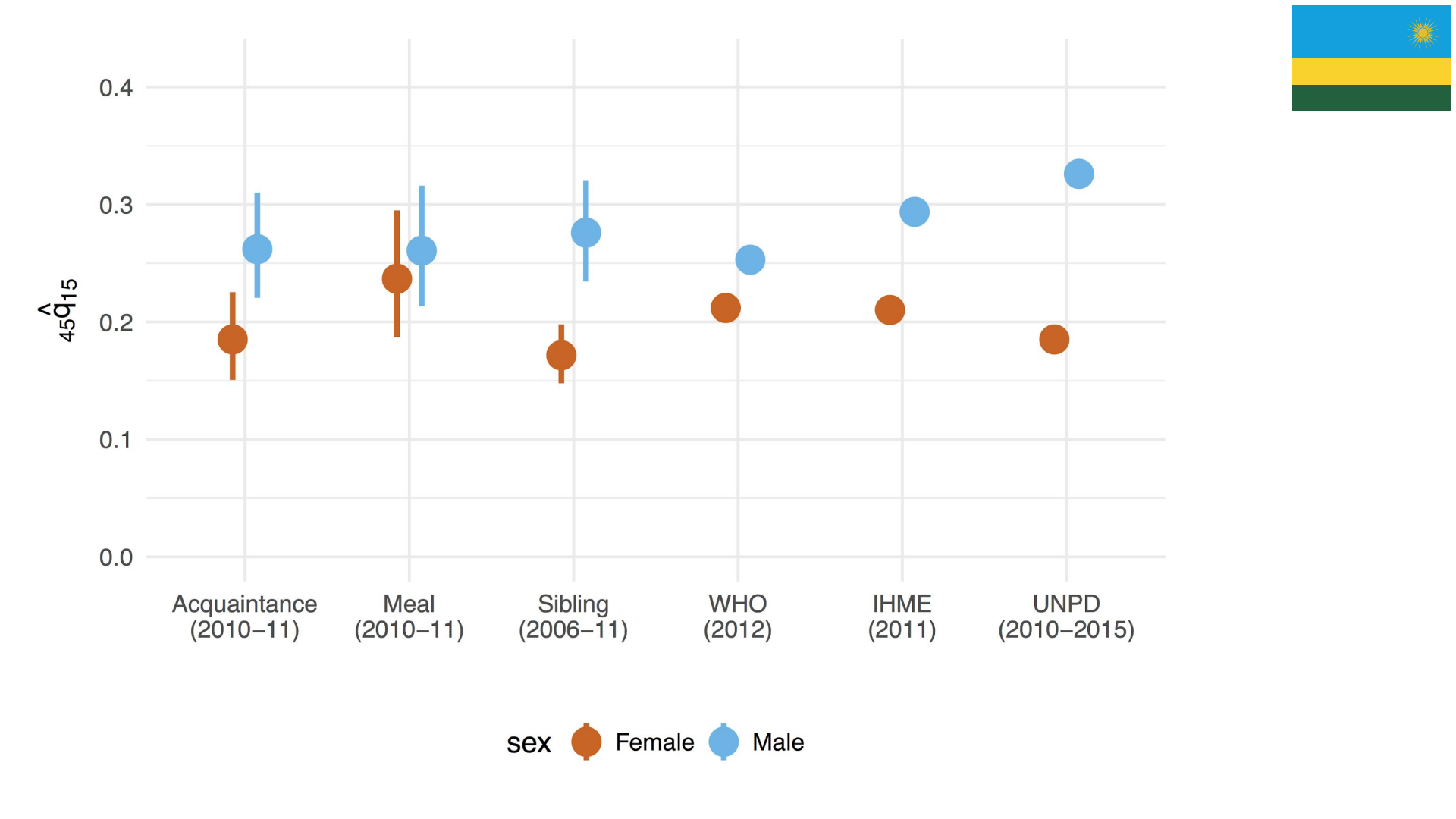


Deaths per interview



- Network reports produce between 4 and 7.5 times as many reported deaths as sibling (7 yrs)


0.4
0.3
0.2
0.1
0.0
$\hat{}_{45}q_{15}$
Acquaintance (2010–11)
Meal (2010–11)
Sibling (2006–11)
WHO (2012)
IHME (2011)
UNPD (2010–2015)
sex
Female
Male


0.4
0.3
0.2
0.1
0.0
$\hat{}_{45}q_{15}$
Acquaintance (2010–11)
Meal (2010–11)
Sibling (2006–11)
WHO (2012)
IHME (2011)
UNPD (2010–2015)
sex
Female
Male

Summary of Rwanda empirical results

- A network survival study is feasible on a Demographic and Health Survey
- We learned about more deaths from each interview using the network methods
- The estimated age-specific death rates are roughly similar for the sibling method and for the meal and acquaintance tie definitions (especially for males)

Network survival

- For some networks, nonsampling error could be higher than sibling survival
- In the Rwanda study, there is no gold standard - we can’t say for sure which approach is more accurate

Empirical question: which type of network produces more accurate estimates?

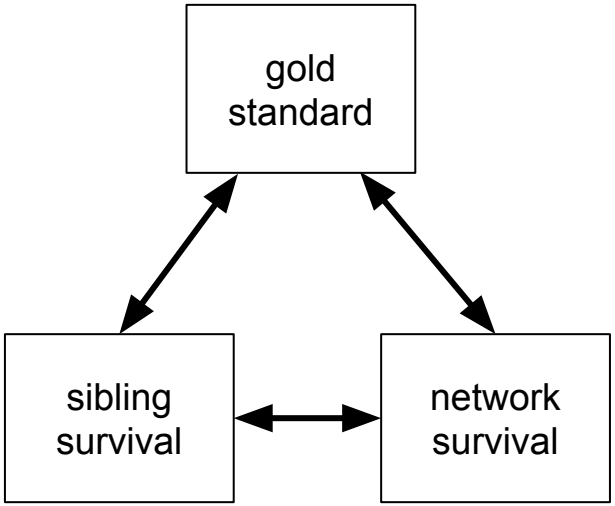
Study design

- 27 state capitals (with DF)
- Household survey: between 600 and 1500 interviews per city, about 25,000 in total
- Multi-stage probability sample
- The results here are preliminary
- Network qs based on people respondent knows and interacted with in the past year

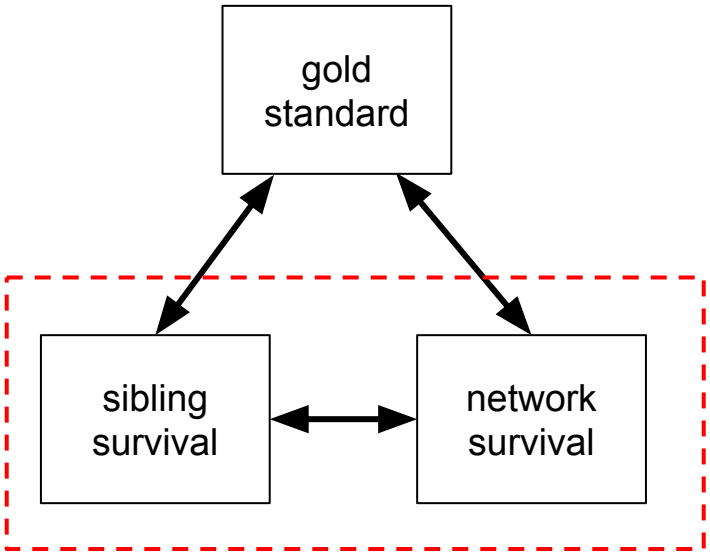
Study design

sibling survival ⟷ network survival

Study design

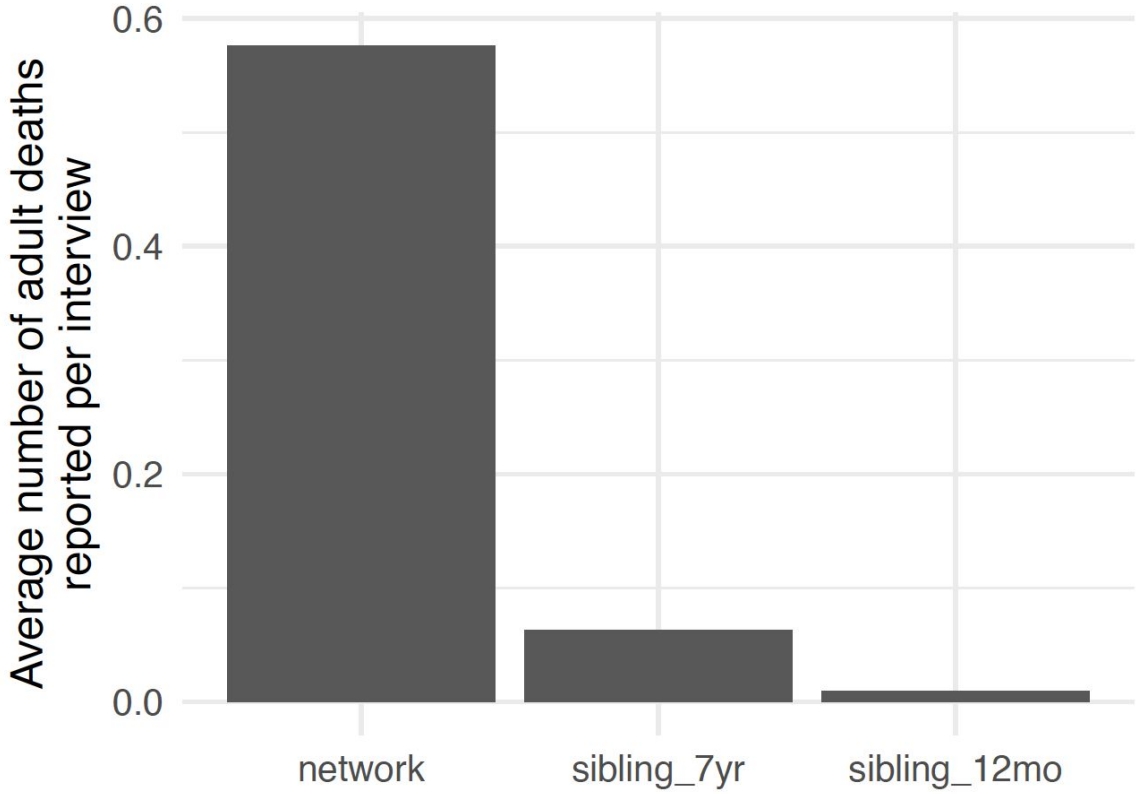


Study design

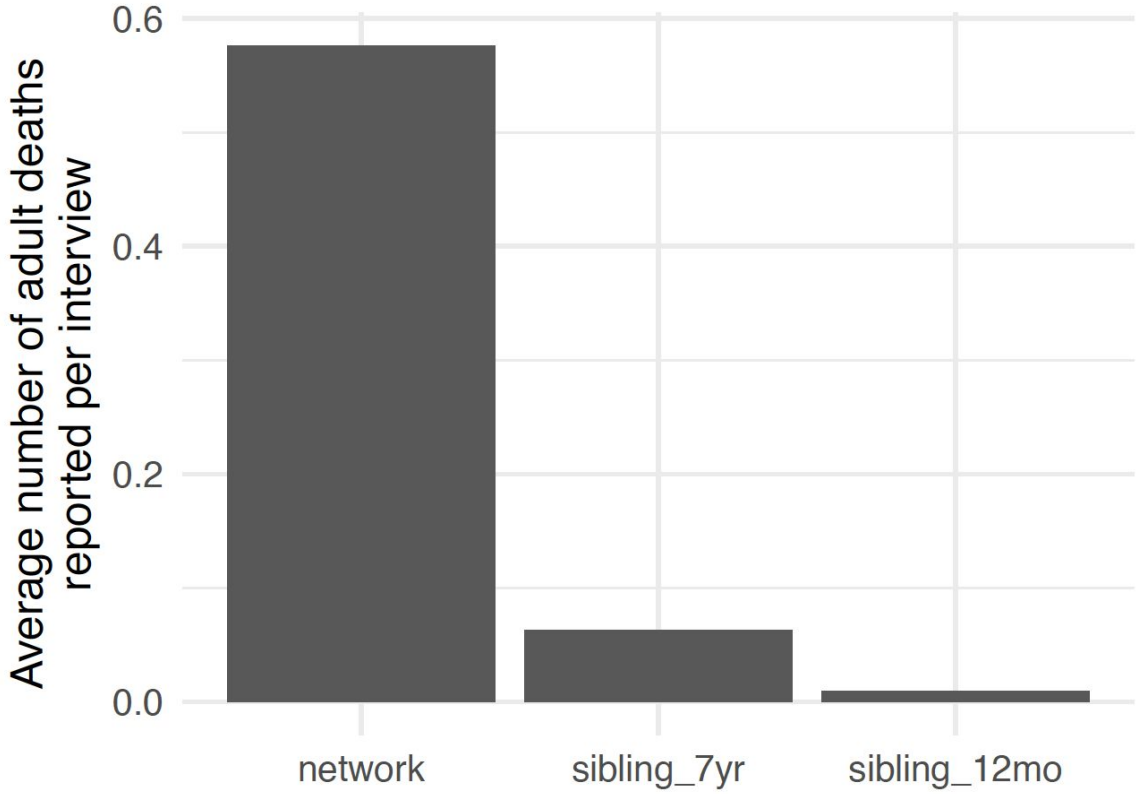


Results: number of reported deaths

Results: number of reported deaths



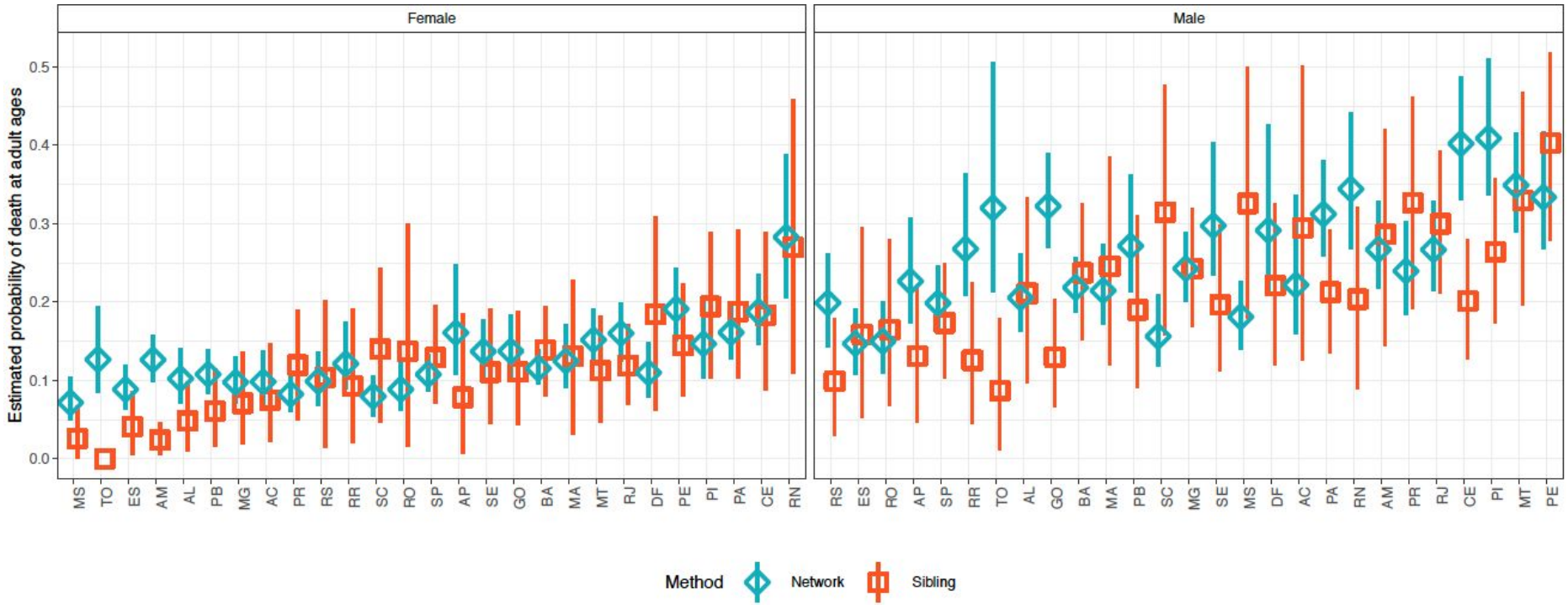
Results: number of reported deaths



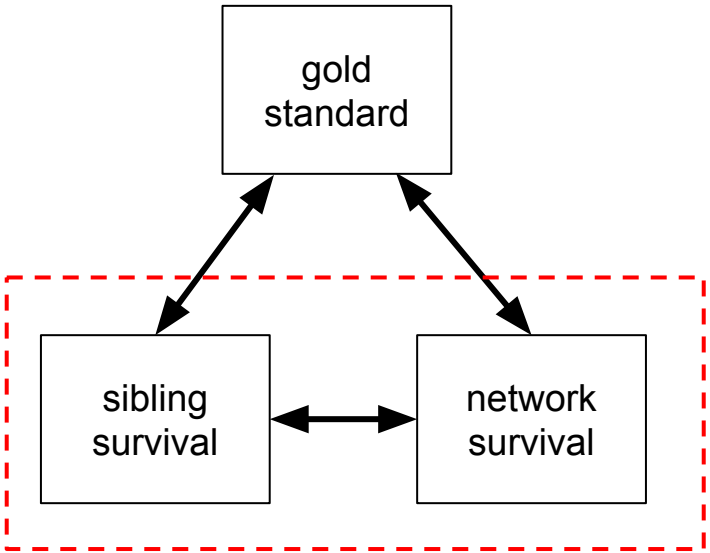
- Sibling (7 yrs) produces about 6.5 times as many reported deaths as sibling 1 year
- Network reports produce about 10 times as many reported deaths as sibling (7 yrs)

Results: sibling and network probabilities of death

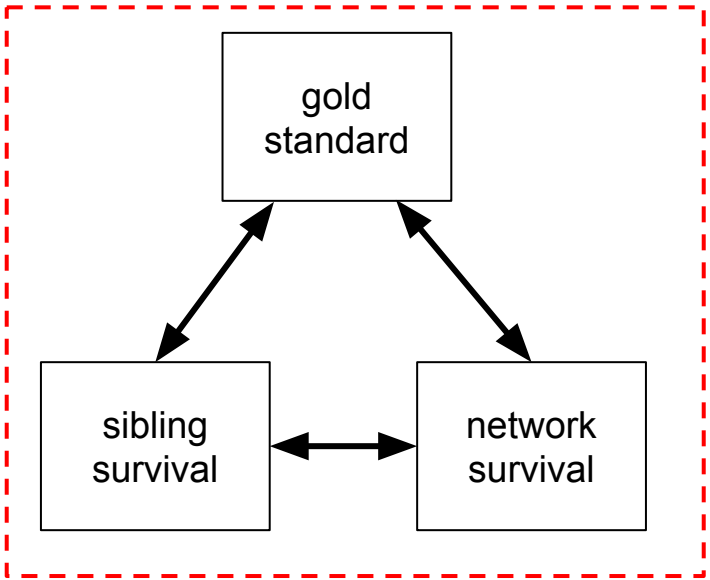
Results: sibling and network probabilities of death



Study design



Study design



Comparing to vital registration

- Lots of decisions go into death rate estimates
- Important not to overfit

Comparing to vital registration

- Lots of decisions go into death rate estimates
- Important not to overfit
- So we’re going to compare to the gold standard only at the very end of the analysis

Comparing to vital registration

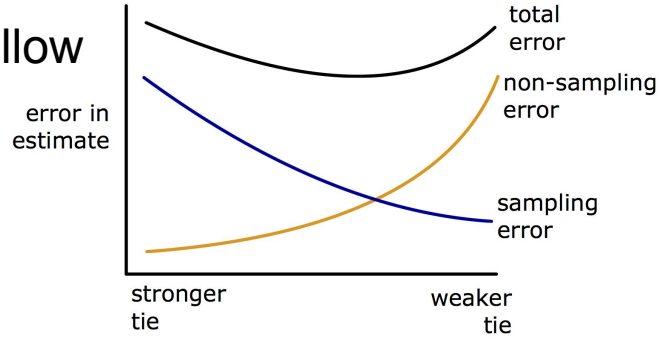
- Lots of decisions go into death rate estimates
- Important not to overfit
- So we're going to compare to the gold standard only at the very end of the analysis
- Important questions
  - What to compare?
    - Age-specific death rates
    - Probabilities of death at adult ages (45q15)
  - How to compare?
    - Relative error
    - Mean squared error across all estimates

Next steps

- **Critical step: comparing to gold standard**
  - **Decide on exactly how to measure discrepancy**
    - mean squared error in estimated death rates?
    - … in estimated probability of adult death?
- **After comparison**
  - Understand any systematic deviations each method has from gold standard
- **Additional modeling**
  - Using model life table information
  - Additional smoothness restrictions?

What I left out today

- **How to estimate network size**
- **Which network to ask about?**
  - It's possible to embed survey experiments that allow researchers to compare questions about two or more different networks
  - Over time, experiments like this can produce information about which sorts of network
- **What about reporting errors? Or differences in network structure?**
  - Experiment with different networks
  - Papers have a mathematical framework for sensitivity to reporting errors
  - In some cases, these reporting errors can potentially be measured and used to adjust estimates



Directions for future work

- **From Brazil survey**: also estimate out-migration and hidden population sizes
- **Network reporting surveys on the internet** -- can use an online sample to estimate characteristics of offline populations (just came out in *Demography*)
- **Sibling method analysis**: use network reporting framework to improve sibling survival estimates (working paper on website)
- **Improvements to data collection and estimates for size of weak-tie network** - upcoming study in Hanoi
- Many other possibilities

Thanks!

- Thanks to my collaborators on several related projects: Matthew J. Salganik (Princeton), Mary Mahy (UNAIDS), Aline Umubyeyi (U. of Rwanda), Wolfgang Hladik (CDC), Francisco Inacio Bastos (FIOCRUZ, Brazil), Neilane Bertoni (FIOCRUZ, Brazil)

- thanks to funders: UNAIDS, USAID, Government of Brazil, NIH

Thanks!

Feedback welcome: feehan@berkeley.edu

For papers and more info: http://www.dennisfeehan.org

Estimating personal network size

To estimate network size, we ask question about connections to groups of **known** size (Killworth et al, 1998).

Suppose that there are
30,000 bus drivers in Rio de Janeiro

and a respondent reportings having
connections to 2 bus drivers

Then we could estimate the respondent's network size with:

$$\widehat{d} = \overbrace{\frac{\text{number of connections to bus drivers}}{\text{total number of bus drivers in Rio}}}^{\text{proportion of bus drivers respondent is connected to}} \times \text{size of RJ's pop.}$$

$$= \frac{2}{30,000} \times 6,000,000$$

$$= 400 \text{ people}$$

In practice, we ask about many known populations to get a better estimate:

reported connections to each known population

$$\widehat{\bar{d}} = \frac{\sum_{i,j} \widehat{y}_{i,A_j}}{\sum_j N_{A_j}} \cdot N_F$$

size of the frame population

total size of each known population

Feehan and Salganik (2016) has the precise conditions that need to hold for this to produce unbiased estimates.