Biases in Survey Estimates of Neonatal Mortality: Results from a Validation Study in Urban Areas of Guinea-Bissau

Helleringer, Stéphane^{1*}
Liu, Li^{1*}
Chu, Yue²
Rodrigues, Amabelia³
Fisker, Ane Barent^{3,4}

Abstract:

Neonatal deaths (i.e., those occurring within 28 days of birth) account for close to half of all deaths among children under age 5 worldwide. In most low and middle-income countries, data on neonatal deaths come primarily from household surveys. We conducted a validation study of survey data on neonatal mortality in Guinea-Bissau (West Africa). We used records from an urban health and demographic surveillance system (HDSS) that monitors child survival prospectively, as our reference dataset. We selected a stratified sample of 599 women aged 15-49 years old among residents of the HDSS, and collected the birth histories of 422 participants. We cross-tabulated survey and HDSS data. We used a mathematical model to investigate biases in survey estimates of neonatal mortality. Reporting errors in survey data might lead to estimates of the neonatal mortality rate that are too high. This may limit our ability to track progress towards global health objectives.

¹ Johns Hopkins University, Bloomberg School of Public Health, Baltimore, Maryland (USA)

² Ohio State University, Department of Sociology, Columbus, Ohio (USA)

³ Bandim Health Project, Bissau (Guinea-Bissau)

⁴ University of Southern Denmark, Odense (Denmark)

* Joint first authors

Correspondence to: Stéphane Helleringer, Department of Population, Family and Reproductive Health, Bloomberg school of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA. Email: sheller7@jhu.edu

BACKGROUND

A neonatal death is a death that occurs during the first 28 days of life. In 2017, the UN Inter-Agency Group for Child Mortality Estimation (IGME) estimated that, worldwide, 47% of all children who died before age 5 died during the neonatal period (IGME 2018). Reducing the neonatal mortality rate (NMR) is now a key target of the 3rd Sustainable Development Goal (SDG): by 2030, countries should strive to reduce neonatal mortality to at least as low as 12 deaths per 1,000 live births (UN 2018). To facilitate progress towards this target, accurate data on the levels and trends of neonatal mortality are needed (Lawn et al. 2014; Moxon et al. 2015). This will allow better targeting interventions to reduce neonatal mortality, and monitoring the effects of changes in intervention coverage (Davey and Deribe 2017).

In most high-income countries, such data come from civil registration and vital statistics (CRVS) systems that operate continuously. In low- and middle-income countries (LMICs), these systems are often deficient (AbouZahr et al. 2015). A large proportion of births and newborn deaths occur at home, and are not registered by family members (Amo-Adjei and Annim 2015; Fisker, Rodrigues and Helleringer 2019). Even when these events occur at health facilities, their recording might be incomplete, delayed or inaccurate (Mikkelsen et al. 2015). As a result, official statistics based on CRVS data, and reported by Ministries of health and other agencies, under-estimate neonatal mortality by orders of magnitude (Lawn et al. 2008; Malqvist 2011; Setel et al. 2007)

Data on neonatal mortality are also collected periodically in most LMICs during household surveys, such as the Demographic and Health Surveys (DHS), and the Multiple Indicator Cluster Surveys (MICS). These surveys frequently conduct full birth

histories (FBH): women aged 15-49 years old are asked to report all their live births, to state the date of birth of each child, and whether he/she is still alive. For each deceased child, they are asked to indicate at what age the child died. From these data, demographers can directly estimate NMR for recent time periods (e.g., the past 5 years), by dividing the number of reported neonatal deaths by the number of births reported to have occurred over the same time frame (Moultrie et al. 2013). Such estimates form the cornerstone of global studies of neonatal mortality (IGME 2018; Liu et al. 2015). They also constitute the standard that new methods to interpret other data sources on child mortality (e.g., census data, health information systems) are evaluated against (Baschieri et al. 2019; Helleringer et al. 2018) or trained to emulate (Burstein et al. 2018).

Biases and errors in survey data on neonatal mortality: NMR estimates obtained from survey data may be affected by several issues. They may contain missing data, e.g., on the date of birth of a child, or on their age at death. These fields thus need to be imputed, often requiring strong assumptions about the causes of missing data. The frequency of missing data on dates of birth and age at death in FBH has however declined over time in household surveys (Pullum and Becker 2014).

Survey data may display “survivor bias” (Hill and Choi 2006), because FBH can only be collected from women who have survived until the time of the survey. If the children of the mothers who died before the survey were also more likely to die than other children (Armstrong Schellenberg et al. 2002; Finlay et al. 2015; Zaba et al. 2005), then NMR estimates will be too low. Simulation studies suggest that survivor bias might be corrected, even in contexts where infectious diseases transmitted from mother to child such as HIV are circulating (Hallett et al. 2010; Walker, Hill and Zhao 2012). Finally,

survey data on neonatal mortality may be affected by limited sample size because even in high mortality settings, neonatal deaths are a rare (statistical) event (Hill and Choi 2006).

The FBH data reported by survey respondents may also differ from their true birth history (Figure 1) because of four types of reporting errors. First, *date displacement* occurs when a respondent's answers to the FBH imply a date of birth of her child(ren) that differs from the true date(s). Date displacement affects NMR estimates when it shifts a birth/death into or outside of a reference period for which estimates are sought. Second, *age errors* occur when a respondent misstates the age at which her child died. This is often detected by measuring the extent of heaping in FBH data, i.e., a higher than expected number of events reported to have occurred at specific ages such as 7 days or 12 months (Guillot et al. 2012; Hill and Choi 2006). Age can be over-stated, for example when a child who died during the neonatal period is reported to have died at an age greater than 28 days. Age can also be under-stated, for example when a child who died in the post-neonatal period (i.e., between 28 and 364 days after birth) is reported to have died as a newborn. Age errors can affect NMR estimates when deaths are erroneously shifted in and out of the neonatal period.

[FIGURE 1 ABOUT HERE]

Third, *omissions* occur when respondents do not list all of their live births during the FBH. Omissions might affect NMR estimates if they occur among neonatal deaths at a different rate than among live births. Finally, *misclassifications* occur when respondents erroneously report stillbirths (i.e., a baby born with no signs of life at or after 28 weeks' gestation) as neonatal deaths, or leave a neonatal death out of the FBH, because they

considered that the child was not viable.

Sources of reporting errors in full birth histories: Several factors explain the emergence of reporting errors in FBH. Poor recall of events that have occurred several years ago, as well as limited numeracy among some population groups, might account for some errors in FBH data. Adjustment procedures have been proposed to correct for recall issues (Masquelier, Reniers and Pison 2014; Obermeyer et al. 2010), but they make strong assumptions. For example, they assume that respondents accurately report the most recent events, e.g., those that occurred in the past 0-3 years. In validation studies of survey data on adult mortality, this assumption did not hold (Helleringer et al. 2014a). They also consider that the proportion of events (e.g., deaths) that are recalled by a respondent declines linearly with the amount of time elapsed between the event and the survey. Studies in cognitive psychology however suggest that this assumption may not be met in practice: more distant events might be easier to recall than more recent ones depending on a complex array of emotional, neural and cognitive factors (Wixted 2004).

Respondents might also omit to list some births because they might evoke painful memories (e.g., the death of a child) or because they might be associated with stigmatized behaviors (e.g., out-of-wedlock childbearing). More generally, social desirability bias might explain inaccurate survey reports of adverse perinatal events such as neonatal deaths and stillbirths (Haws et al. 2010). Women might conceal these aspects of their birth histories during an interview because they fear judgmental attitudes, gossip or stigma, either from the interviewer, relatives or other community members.

Additional reporting errors may occur in FBH because the local terms used by women to

describe pregnancy outcomes in LMICs might not match the western biomedical concepts of stillbirths and neonatal deaths. Haws et al. (2010) documented that women in Tanzania used terms (e.g., “immature baby”) that could apply to both stillbirths and neonatal deaths. Because the fetal heart rate is not systematically monitored in under-resourced health facilities in LMICs, health workers might also mistakenly classify some babies who do not show signs of movement as stillbirths (Spector and Daga 2008).

These factors may be accentuated by the behaviors of data collectors. Interviewers may be tempted to skip recording some births reported by the respondent in order to speed up the interview, thus leading to omissions. They might also report some children as being born earlier than they are in order to skip long series of questions that are only applicable for specific age groups. This might occur in surveys like the DHS or MICS where time-consuming modules about post-natal care or child health are only applicable to children born within the past 5 years (Pullum and Staveteig 2017; Pullum 2006). This would then lead to an increased frequency of date displacement and/or age errors.

Prior assessments of the accuracy of full birth histories: Few studies have quantified the extent of reporting errors in FBH data. Liu et al. (2016) used data collected in Malawi to assess how common misclassifications might be in FBH. Compared to a more intensive verbal autopsy questionnaire conducted several months after the FBH survey, they found that 21% of the neonatal deaths reported during FBH might in fact have been stillbirths. Based on qualitative interviews, Haws et al. (2010) uncovered discrepancies between a woman’s reproductive history and her reports of reproductive events during a survey interview. They described several cases of women who reported a stillbirth while indicating that the baby had moved or breathed at the time of delivery, or who reported a neonatal death while describing that the baby’s body was cold or had deteriorating skin

(thus suggesting it might have been a stillbirth).

A more comprehensive validation study of FBH data was conducted in the 1990's in Bangladesh (Espeut and Becker 2015). In that study, FBH respondents reported 81.0% of the neonatal deaths they had experienced according to a reference prospective dataset extracted from the Matlab Health and Demographic Surveillance System (Alam et al. 2017). This study investigated the effects of reporting errors on estimates of the total fertility rate, but not on NMR estimates. Other assessments of the quality of FBH data have been reliability studies, in which estimates of mortality rates obtained from FBH were compared to estimates obtained from other sources or to predictions from statistical models (Hill and Choi 2006; Pullum and Becker 2014). Such studies are based on benchmarks with significant limitations (e.g., model life tables, census data). They have focused on infant mortality (Guillot et al. 2012), rather than on neonatal mortality.

The net impact of reporting errors on estimates of the NMR obtained from FBH data is difficult to predict. This is so because some errors might typically lead to downward biases (e.g., omissions), whereas other errors might lead to upward biases (e.g., misclassifications of stillbirths as neonatal deaths). In some settings, these errors might offset each other and yield a roughly accurate NMR estimate. In other settings though, they might cumulate and result in severely biased NMR estimates. We report the results of a detailed validation study of FBH conducted in an urban setting of Guinea-Bissau (West Africa). We used a simple mathematical model to show that the patterns of reporting errors observed in these urban neighborhoods might lead to estimates of the NMR that are too high.

DATA AND METHODS

Study setting: Guinea-Bissau is a low-income country with a population of ≈ 1.8 million and an estimated life expectancy of 56 years in 2010-15 (United Nations 2017). The most recent nationally representative survey that collected data on neonatal mortality was conducted in 2014, through the MICS program (Khan and Hancioglu 2019). It indicated that, nationwide, more than half of the births (55.2%) occurred at home, with large differences between urban and rural areas (28.0% vs. 71.0%, respectively). The proportion of deliveries that occurred by cesarean section was low (3.9%), but in urban areas, such surgeries were more common (8.0% of all births). Most cesarean sections (3 out of 4) were unplanned and had occurred due to an obstetrical emergency. The 2014 MICS documented a nationwide NMR of 36 deaths, and a post-neonatal mortality rate of 20 deaths, per 1,000 live births (INE 2014). The NMR varied between provinces of the country (from 30.2 in the north to 47.7 in the east), and by educational level of the mother (from 40.8 among those without schooling, to 27.7 among mothers with secondary schooling or more). The urban-rural gradient in NMR was more limited (34.4 in rural areas vs. 38.1 in urban areas).

Reference dataset: We worked in Guinea-Bissau due to the availability of a high-quality reference dataset on neonatal mortality in some areas of the country. This dataset is collected through the urban Health and Demographic Surveillance System (HDSS) of the Bandim Health Project. This is an open cohort that monitors the populations of six neighborhoods of Bissau –the capital city– since 1978 (Bjerregaard-Andersen et al. 2018; Rodrigues et al. 2008). The areas covered by the HDSS are situated approximately 2 km away from the city center and include more than 103,000 inhabitants. Every month, fieldworkers visit every household in HDSS areas to register

pregnancies and record their outcomes (e.g. stillbirth vs. live birth). After birth, children are followed every 3 months to record survival, nutritional status and health-seeking behaviors. The HDSS also registers all deliveries at the National Hospital Simão Mendes (Bjerregaard-Andersen et al. 2018) and at a health center serving parts of the HDSS population. The health records pertaining to each pregnancy/newborn compiled at these facilities are then linked with the HDSS records. New households (e.g. those established after construction of a new dwelling) are continuously added to the HDSS. In figure 2, we show trends in NMR since 2005 in the neighborhoods of Bissau covered by the HDSS. We also display trends in urban areas of the country over the same time period, as estimated from representative FBH surveys conducted in 2010 and 2014 (i.e., MICS).

[FIGURE 2 ABOUT HERE]

We used the HDSS as the reference dataset against which we evaluated FBH data. The HDSS in Bissau has one of the most rigorous protocols for monitoring perinatal events and infant survival, because it is based on monthly household visits to detect and monitor pregnancies and their outcomes (Akuze et al. 2017). Despite their high quality, however, HDSS data do not constitute a gold standard measure of neonatal mortality. Some pregnancies might be missed during monthly household visits and some stillbirths may be misclassified as neonatal deaths in HDSS data, and vice-versa.

Collection of full birth histories⁵: We selected a stratified random sample of women aged 15-49 years old from the lists of HDSS residents. We oversampled women in three

⁵ Individual identifiers are not provided with MICS datasets. It was thus not possible to validate the FBH data collected during the 2014 MICS survey in Guinea-Bissau through linkages with the HDSS. Instead, we collected new FBH data among the HDSS population.

strata: those who had experienced in the past 5 years either 1) a neonatal death among their live-born children, 2) a post-neonatal death among their live-born children, or 3) a stillbirth. This stratification was necessary to ensure that we had sufficient reported numbers of such events to evaluate the accuracy of FBH. The fourth sampling stratum was constituted of women of reproductive age who did not experience any of these events according to the HDSS over the same time frame. For example, all of their children were still alive, or had died after age 1 year. We designed a set of survey weights to account for differences in the probability of inclusion and participation rates in each of these strata.

Our questionnaire included a subset of the modules administered during the 2014 Guinea-Bissau MICS: a) questions about the respondent's socioeconomic background, b) summary questions about her fertility, and c) a standard FBH. The FBH module first asked respondents to list all the live births they have ever had in chronological order. Then, for each reported live birth, it asked the gender of the child, whether he/she was part of a multiple birth, whether he/she was still alive at the time of the survey, and the date of birth. For children who had died, the FBH asked respondents to state the age at death in days if they had died within their first month of life, in months if they had died between 1 and 23 months, and in years if they had died at older ages.

The questionnaire also included d) questions about pregnancies that did not result in a live birth. We asked respondents if they had ever experienced such a pregnancy termination. If so, we asked them how many such terminations they had experienced over their lifetime, and when was the most recent one. We did not ask respondents to state the type of pregnancy termination they had experienced (e.g., stillbirth, miscarriage or abortion). Finally, as in the 2014 MICS, we included e) questions about live births of

the past two years prior to the survey. These questions focused on antenatal and post-natal care, birth and (if applicable) death registration, and the care of recent illnesses such as cough, diarrhea and/or fevers.

We recruited interviewers who had worked on the 2014 MICS, and thus had extensive experience collecting FBH data. The interviewers were not given access to HDSS data: they did not know whether the women they had to interview had experienced any births, adverse perinatal events (e.g., stillbirths, neonatal deaths) or post-neonatal deaths over the past 5 years. We devised an electronic data collection tool that emulated the platform used in MICS and DHS surveys. We used Qualtrics (Snow and Mann 2013) on android tablets, because it allowed displaying all reported live births in a large table on a single screen, as currently done in DHS and MICS. As in those latter surveys, we incorporated data consistency checks, which alerted interviewers to potential errors or implausible FBH patterns. All study instruments were translated into Portuguese. As was done during the 2014 MICS, questions were asked in Creole, the vernacular language most commonly used in Bissau.⁶

Data analysis: We first described the constitution of the study sample, including reasons for non-participation. We investigated the selectivity of the study sample by comparing the HDSS records of the women who participated in the validation study, to the HDSS records of women who were selected but could not be interviewed. This comparison focused on maternal characteristics and recent birth histories, as well as on characteristics of the “index” pregnancy. That latter term refers to the pregnancy that prompted inclusion of the respondent in one of the sampling strata described above. For

⁶ Two interviews required a translator and were conducted in French (among Fulani migrants recently arrived from the neighboring Republic of Guinea).

example, consider a woman who had 3 pregnancies recorded by the HDSS over the past 5 years: two of these pregnancies resulted in live births who then survived until age 1, whereas the third pregnancy resulted in a live birth, but the child died after 7 days. This woman was then included in the “neonatal death” stratum, and the characteristics of the index pregnancy we considered in these analyses concern the pregnancy that led to a neonatal death.

Maternal characteristics included age at the time of the survey (in 5-year age groups), educational level, ethnic group, and a binary variable indicating whether the mother engaged in work outside of the home. We also considered several variables in HDSS datasets that denote the socio-economic status of residents. These included binary variables describing the roofing material, access to electricity and the type of toilet facility available at the house of a potential participant. Characteristics of the recent birth history of potential participants included the number of neonatal deaths, post-neonatal deaths, and stillbirths they had experienced over the past 5 years, as well as a binary variable indicating whether they had experienced at least one multiple birth (e.g., twins) over the same time frame.

Characteristics of the index pregnancy included the use of antenatal care and vaccination services during that pregnancy, the place of delivery, the number of live births a woman had already had prior to the index pregnancy and the mode of delivery of the index pregnancy (vaginal vs. cesarean section). Finally, when available, we also provided data on the birth weight of the index pregnancy. Some of these variables had a limited number of cases in some of their categories (e.g., $n < 10$). We thus used Fisher’s exact test (Agresti 1992) to assess differences in the distribution of these variables between women who were interviewed, and those who were not.

Evaluation of the accuracy of FBH data: We cross-tabulated FBH data with reference data from HDSS datasets. We carried our analyses separately for neonatal and post-neonatal deaths. We defined the sensitivity of FBH data as the proportion of respondents with neonatal/post-neonatal deaths in the previous 5 years according to the HDSS, who reported experiencing such events in the same timeframe during FBH. We explored whether the sensitivity of FBH in recording neonatal/post-neonatal deaths varied by time since the death. To do so, we binned neonatal/post-neonatal deaths in 6 months intervals in HDSS datasets, and we recalculated the sensitivity of FBH data in each of these bins.

We defined the specificity of FBH data in recording neonatal/post-neonatal deaths as the proportion of respondents without neonatal/post-neonatal deaths in the previous 5 years according to the HDSS, who did not report such events in the same timeframe during FBH. We investigated whether the specificity of FBH data in recording neonatal deaths varied across the three sampling strata in which women had not experienced a neonatal death according to the HDSS.

In this context, a “false negative” occurred when a respondent with a neonatal/post-neonatal death during the reference period according to HDSS did not report such a death during FBH. A false negative was attributed to 1) a date error if the respondent reported a neonatal/post-neonatal death *before* the reference period; 2) an age error if the respondent reported a death in another age group during the reference period; 3) an omission if the respondent did not report any under-5 death or pregnancy termination during the reference period, and 4) a misclassification if the respondent did not report any under-5 death but reported a pregnancy termination during the reference period.

Similarly, a “false positive” (i.e., a respondent without a neonatal/post-neonatal death according to HDSS who nonetheless reported such a death during FBH) was attributed to 1) a date error if the respondent reported a neonatal/post-neonatal death that had occurred prior to the reference period as having occurred during the reference period; 2) an age error if the respondent had experienced a neonatal, post-neonatal or child death according to HDSS but mistakenly reported it as having occurred in another age group during the reference period; and 3) a misclassification if the respondent had experienced a pregnancy termination according to HDSS but reported it as a neonatal death during the reference period of the FBH. All analyses used survey weights to account for differences in the probabilities of selection and participation across sampling strata.

Effects of FBH reporting errors on NMR estimates: We devised a mathematical model to investigate the effects of observed reporting errors in FBH data on the accuracy of NMR estimates. This model links the true NMR in a population to a survey estimate of the NMR, via a set of conditional probabilities describing reporting errors in FBH (e.g., sensitivity, specificity). A complete description of the model, hypotheses and parameters is given in appendix A1.

Following a common approach in assessments of neonatal mortality data (Hill and Choi 2006), we applied this model to data series from England and Wales, where live births, stillbirths, neonatal deaths, infant and under-5 deaths have been reported annually since the 1920’s (appendix A2). For each 5-year time period in the England and Wales series, we used our model to calculate the estimate of the NMR that would have been obtained if we had conducted a FBH survey with reporting parameters (e.g., sensitivity) similar to those observed in Bissau. We accounted for sampling uncertainty in our assessments of reporting errors in Bissau by generating 1,000 survey estimates of the NMR through

random draws from the distributions of each reporting parameter in conjunction with model equations. We calculated the 2.5th and 97.5th percentiles of the distribution of these estimates. Finally, we compared these counterfactual estimates to the NMR that was recorded through vital statistics in England and Wales in each 5-year period.

Robustness tests: Because our reference HDSS data do not constitute a gold standard measure of neonatal mortality, we performed additional tests to evaluate the robustness of our findings to possible errors in HDSS data. We re-evaluated the FBH data reported during the validation study after 1) randomly reclassifying varying proportions of events recorded by the HDSS datasets (e.g., some stillbirths were recoded as neonatal deaths, and vice-versa), and 2) introducing additional stillbirths, live births and (post-)neonatal deaths in the HDSS dataset to account for the potential incompleteness of HDSS registration of pregnancies. These robustness tests and their results are described in Appendix A3. All analyses were conducted using Stata, version 14.1.

RESULTS

We selected 599 records of women aged 15-49 years old among HDSS rosters (figure 3). After reviewing HDSS data and linked health facility records, we excluded 13 women who had died prior to the study, 3 who were duplicate registrations, and 1 who was older than 49 years old. We contacted 582 women: among those, 7 had died prior to our visit, and 80 had permanently migrated outside of the HDSS area since the last HDSS visit. They were thus ineligible for study participation. Among the 495 eligible women, 66 were absent at all 3 study visits, and 3 were unknown. In two instances, we interviewed someone who was not the target HDSS resident. This was detected by large discrepancies (>20 years) between the age reported by the respondent and her birth

record in the HDSS. It was then confirmed by household inquiries made by study supervisors. In two other instances the respondent was registered by the HDSS only after the birth of (one of) her child(ren). As a result, reference prospective data on neonatal mortality were not available for these two respondents and they were excluded.

[FIGURE 3 ABOUT HERE]

Four hundred and twenty-two women were included in our analytical sample, yielding a participation rate of 85.3% (422/495). In total, 178 respondents had experienced at least one neonatal death over the past 5 years according to the HDSS, 86 had experienced a post-neonatal death and 89 had experienced at least one stillbirth (figure 3). In HDSS datasets, over the past 5 years prior to our study, there were 866 records of pregnancies pertaining to the 422 respondents we interviewed. Among those, 734 resulted in a live birth.

Women who were interviewed were more educated, and were less likely to belong to the Fulani and Manjaco ethnic groups than those who could not be interviewed (table 1). They were also less likely to have experienced multiple post-neonatal deaths among their children in the past 5 years ($p=0.009$).

[TABLE 1 ABOUT HERE]

Close to one in three women interviewed were aged 15-24 years old, whereas slightly less than one in five were 35 years or older (table 1). The educational level of participants was low, with 24.2% of respondents never having attended school according to HDSS records. Only 12.8% of respondents reported working outside of their home.

The socio-economic status of participants was also low, with only 43.1% residing in a house with electricity, and 17.5% in a house with indoor toilets. The large majority of respondents (96.6%) sought antenatal care at least once during their index pregnancy, and most had also been vaccinated against tetanus (87.3%) prior to delivery.

Approximately one in five index deliveries occurred at home, whereas 42.2% of index deliveries occurred at the national hospital. More than two thirds of respondents had never had a live birth prior to their index pregnancy (69.1%). According to HDSS data, close to one in eight index deliveries (12.3%) occurred via cesarean section.

The sensitivity of FBH in recording neonatal deaths was 0.791 (table 2). The sensitivity of FBH in recording post-neonatal deaths was lower (0.691). The highest levels of sensitivity of FBH in recording both neonatal and post-neonatal deaths were observed in deaths that had occurred less than 6 months before the survey according to the HDSS (appendix A4). The sensitivity of FBH in recording neonatal deaths did not appear to decline further with the amount of time elapsed between the survey and the death, whereas there was more evidence of such a continued decline in sensitivity among post-neonatal deaths (appendix A4).

[TABLE 2 ABOUT HERE]

False negative reports of neonatal deaths were due in large part to misclassifications of neonatal deaths as stillbirths (43.4%, figure 4). Omissions accounted for close to a third of false negatives (30.4%). Date displacements and age errors accounted for smaller proportions of false negatives (18.0% and 8.2%, respectively). For post-neonatal deaths, the majority of false negatives stemmed from age under-statement and omissions. The

complete distributions of age at death during the neonatal and post-neonatal periods according to HDSS and FBH data are shown in appendix A5.

[FIGURE 4 ABOUT HERE]

Among false negative reports of neonatal deaths, the sources of error differed between deaths that had occurred in the early neonatal period (0-6 days) and those that occurred in the late neonatal period (7-27 days) according to the HDSS (Appendix A6). Among early neonatal deaths that were not reported during FBH, close to two thirds were due to misclassifications. All of those deaths were neonatal deaths that had occurred on the day of birth according to the HDSS. On the other hand, there were no misclassifications among the deaths that occurred in the late neonatal period according to the HDSS. Instead, errors in that period originated in roughly equal proportion from date errors, age over-estimates and omissions.

The specificity of FBH data in recording neonatal deaths was high (0.993, table 2), but it varied depending on the respondent's pregnancy history, as recorded by the HDSS. Among respondents with at least one stillbirth in the past 5 years, specificity was 0.866, whereas it was 0.840 among respondents with a post-neonatal death in the past 5 years, and 1.000 (i.e., no false positive reports) among other respondents. The specificity of FBH in recording post-neonatal deaths was 0.999.

False positive reports of neonatal deaths in FBH data were primarily due to misclassifications of stillbirths as neonatal deaths (62.7%, figure 4). Other sources of false positive reports of neonatal deaths included age under-statement (31.4%) and date displacement (5.9%). All the false positive reports of neonatal deaths attributable to

misclassifications were reported to have occurred within the first two days after birth according to the FBH respondent. There were few false positive reports of post-neonatal deaths, but all of them were due to age over-statement (figure 4).

When inserted into our mathematical model, these patterns of reporting errors implied that FBH data might over-estimate the NMR. When applied to data from England & Wales between 1927 and 2017 (figure 5), our model predicted a survey estimate of the NMR that would be 4.7% to 19.8% higher than the NMR reported from vital statistics. This upward bias was highest for time periods when the “true” NMR was either high (approximately 30 per 1,000), or low (< 5 neonatal deaths per 1,000) in England and Wales. The confidence intervals associated with our estimates of bias in survey estimates of the NMR were however wide, and frequently included the possibility that a FBH survey would result in a NMR that was either unbiased or too low (i.e., bias < 1).

[FIGURE 5 ABOUT HERE]

DISCUSSION

Using a high-quality reference dataset from an urban HDSS in Guinea-Bissau, we found that FBH data collected during household surveys had high specificity (0.993) but moderate sensitivity (0.791) in recording neonatal deaths. FBH data also had high specificity, but lower specificity (0.691) in recording post-neonatal deaths. This validation study adds to prior work on the accuracy of FBH data in several ways. On the one hand, we provide estimates of the specificity of FBH in recording neonatal and post-neonatal deaths. Such estimates were not available from the other validation study of FBH data that was conducted in the 1990’s in Bangladesh (Espeut and Becker 2015). Measuring

the specificity of FBH data is however crucial in evaluating the potential effects of reporting errors on the accuracy of mortality estimates. Indeed, since neonatal deaths are rare events, large biases can result from even small numbers of false positive reports of neonatal deaths (e.g., misclassified stillbirths, post-neonatal deaths wrongly stated as having occurred during the neonatal period).

On the other hand, we also investigated the sources of errors in FBH data. We showed that FBH data on neonatal mortality were vulnerable to all 4 potential sources of reporting errors: date displacement, age errors, omissions and misclassifications. We found that stillbirths were often misclassified as neonatal deaths, whereas a number of post-neonatal deaths were transferred to the neonatal period due to age errors. This information might play a key role in guiding data improvement strategies and the development of new statistical models of the NMR that account for reporting errors in FBH data.

Finally, whereas the prior validation study in Bangladesh focused on FBH data accuracy on the total fertility rate, we investigated the net effects of reporting errors in FBH data on estimates of the NMR. Using a simple model applied to historical data on perinatal mortality from England and Wales, we found that FBH data might lead to estimates of the NMR that are too high. Across contexts characterized by neonatal mortality rates ranging from 2.5 to >30 per 1,000 live births, the reporting errors in FBH data observed in urban areas of Guinea-Bissau yielded over-estimates of the NMR ranging from 5 to close to 20% (figure 5).

Our study has several limitations. First, the size of our validation sample was too limited to estimate various reporting parameters precisely. For example, our estimates of the

extent of age transfers between neonatal and post-neonatal groups were based on a few cases, thus resulting in high levels of uncertainty. This affected our assessment of the impact of reporting errors on the accuracy of NMR estimates, leading to wide confidence intervals that included the possibility that the survey estimate was unbiased. We explored whether the sensitivity of FBH data declined with the amount of time elapsed between the death and the survey (appendix A4), but we could not ascertain whether the types of reporting errors leading to reduced sensitivity among less recent deaths (e.g., omissions vs. misclassifications) varied with time since the death. Our limited sample size also prevented the investigation of the covariates of reporting errors, e.g., whether such errors were related to characteristics of the mother, or to various beliefs and attitudes that she may hold. As a result, we are not able to characterize the mechanisms through which reporting errors emerged in FBH data in this setting (e.g., recall errors, interviewer behaviors and/or social desirability bias).

Second, we used the FBH instrument that was implemented in the 2014 Guinea-Bissau MICS, but that instrument only inquired about pregnancy terminations, without asking respondents to further specify the type of pregnancy termination(s) they experienced (e.g., stillbirths, miscarriages or abortions). We thus could not investigate the accuracy of survey measures of the stillbirth rate. This might also have slightly affected our assessment of the sources of errors in FBH data. Indeed, when a neonatal death was not reported, we attributed this error to misclassifications between stillbirths and neonatal deaths as long as the respondent reported at least one pregnancy termination during the reference period. But if that pregnancy termination was not a stillbirth (e.g, if it was a miscarriage), then the error might instead have been attributed to an omission instead.

Third, our reference HDSS dataset might have been affected by errors in the recording of neonatal deaths. The HDSS might have included a number of misclassifications between stillbirths and neonatal deaths. It might also have failed to record a number of pregnancies that have occurred in the target population. As a result, some participants in our study might have experienced a neonatal death among their children, even though the HDSS did not record it. In robustness tests (appendix A3), we found that errors in HDSS data might lead to estimates of the specificity and sensitivity of FBH data that are lower than the ones reported in table 2. For example, in one of the HDSS data scenarios we considered, sensitivity declined from 0.79 to <0.70. However, this seldom affected our assessment of the net effects of the reporting errors observed in Bissau on estimates of the NMR. In most data scenarios, FBH data still led to estimates of the NMR that were too high. We found a few exceptions in scenarios in which we assumed highly prevalent and selective patterns of errors in HDSS data. In such scenarios, FBH data might lead to *under*-estimates of the NMR in settings where neonatal deaths are common, relative to stillbirths and post-neonatal deaths (see appendix A3).

Fourth, our results may not be representative of some of the settings where FBH data are collected in LMICs. Our study population was urban, and our results might not extend to rural areas. Our study also did not include a number of potential participants who had either migrated permanently or were temporarily absent from the HDSS area. These women were significantly less educated than women who participated in our study. If the accuracy of FBH reporting is correlated with educational level, then our results might misrepresent the accuracy of FBH. We found that the women who were included in our analytical sample experienced fewer post-neonatal deaths among their children than women who could not be included. Future validation studies of FBH data should seek to track migrants or mobile individuals (Helleringer et al. 2014b).

Fifth, our study did not measure the full extent of reporting errors among live children and among children who died at ages one year and above. We only documented some age errors that transferred those older children to the post-neonatal period. We were not able to measure the prevalence of omissions and date displacements among live children, nor were we able to measure the extent of age errors that transfer deaths among 1-4 year olds to the 5-9 years old age group. Such errors are important because they exclude births from the denominator of the NMR while leaving its numerator largely unaffected. They might thus lead to a large upward bias in survey estimates of the NMR (appendix A1). In our analyses, we used estimates of the prevalence of omissions and date displacements among live children and among deaths that occurred at ages 1-4 years old computed by Pullum and Becker (2014). We used their median estimates of the extent of these errors, whereas in some surveys they found that age transfers in particular could be much more prevalent. If errors in the reporting of FBH data about live children and about deaths among 1-4 year olds are more common than we assumed in this study (appendix A1), then the upward bias in survey estimates of the NMR documented in figure 5 might be larger than we estimated.

Sixth, HDSS participants in Bandim are frequently interviewed and followed-up about their reproductive health and the survival of their children. They may thus report their FBH more accurately during a survey interview than other women who are not in such frequent contact with data collectors. As a result, the extent and magnitude of reporting errors in FBH data might be larger in settings without similar follow-up. Additional validation studies of FBH data should be conducted in diverse LMIC settings with less intensive follow-up, for example through record linkages with antenatal care clinics.

Finally, we did not measure potential errors in the reporting of clusters of neonatal or post-neonatal deaths among mothers (i.e., women experiencing multiple such events over a period of time). This was not possible because of the limited extent of clustering in areas covered by the Bandim HDSS (table 1): only 7 participants (<2%) had experienced more than 1 neonatal death among their children over the past 5 years according to the HDSS. Among those, only one did not report any neonatal death, whereas the other 6 participants reported all the neonatal deaths that were also recorded by the HDSS.

Despite these limitations, our study has important implications. It suggests that in order to improve the accuracy of FBH data collected during surveys, a combination of strategies might be needed. To avoid omissions of neonatal deaths, FBH questionnaires might be supplemented by probing questions or prompts asking respondents about events that might be more likely to be left out of full birth histories (e.g., a death having occurred shortly after birth). To reduce the prevalence of misclassifications between stillbirths and neonatal deaths, FBH questionnaires might be supplemented by questions about the (lack of) vital signs displayed by newborns, e.g., what their skin condition looked like, and whether they moved, cried or breathed. Such questions are often included in verbal autopsy questionnaires, and might allow reclassifying a number of live births as stillbirths if the reported newborn did not display signs of life (Liu et al. 2016). Since misclassifications only appear to affect newborns that have died in the early neonatal period (appendix A5), these questions might be limited to that sub-group in order to save time and resources during surveys.

Survey questionnaires might also be expanded to include a full pregnancy history, during which respondents are asked not only to list all of their live births as in FBH, but also the

pregnancies that did not result in a live birth.⁷ This approach might help collect more accurate data on infant survival (Espeut and Becker 2015). It might also create additional opportunities to probe for the proper classification of events as stillbirths or neonatal deaths. Large-scale trials comparing FBH data to data from full pregnancy histories are currently under way (Baschieri et al. 2019). Data from these trials should be analyzed using our model (appendix A1) to better understand how reporting errors in full pregnancy histories might affect survey estimates of the NMR. Finally, to reduce age errors and date displacements, survey questionnaires might incorporate event history calendars, which have helped improve the quality of adult mortality data (Helleringer et al. 2014b; Helleringer et al. 2015).

Our study also indicates that FBH data should be used cautiously in the development and evaluation of new methods to estimate neonatal mortality. FBH have recently served as a benchmark against which new data collection platforms or statistical models to measure NMR are evaluated (Burstein et al. 2018; Helleringer et al. 2018). While they represent the current best practice in mortality measurement in LMICs with limited vital records, FBH data also misrepresent the true extent of neonatal mortality in potentially complex ways. As a result, biases in NMR estimates obtained from FBH data might be transferred to these new methods of data collection or analysis. When possible, evaluations of new measurement tools should account for bias and errors in FBH data.

Finally, our study highlights that survey data collected through FBH might misrepresent progress towards the achievement of global health objectives related to newborn health.

⁷ In its 8th wave of data collection, the DHS program will begin collecting full pregnancy histories instead of full birth histories in the LMICs where it is implemented (<https://www.dhsprogram.com/pubs/pdf/DHSM11/DHSM11.pdf>)

In order for FBH data to adequately play their expected role in the stewardship of global health programs targeting newborn and child health, they require statistical adjustments that properly account for the added bias and uncertainty linked to reporting errors. The development of such adjustments will likely require additional validation studies in other LMICs, to better understand whether the patterns of sensitivity and specificity of FBH data might differ in other settings, relative to what we observed in urban Guinea-Bissau.⁸ It will also require the development of statistical models that adequately reflect these properties of FBH data in their estimates of the NMR and associated assessments of uncertainty. This should constitute a priority for institutes and organizations currently engaged in assessing progress towards the achievement of the neonatal mortality target of the 3rd sustainable development goal.

⁸ For example, our study yielded an estimate of the sensitivity of FBH data in recording neonatal deaths that was comparable to the estimate obtained in Bangladesh by Espeut and Becker (2015): 0.79 vs. 0.82, respectively. However our study yielded an estimate of the sensitivity of FBH in recording post-neonatal deaths that was much lower than the estimate obtained in Bangladesh: 0.69 vs. 0.87, respectively.

ACKNOWLEDGMENTS:

This work was supported by grants R21HD087811 and R01HD088516 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (PI: Helleringer). We thank Stan Becker and Tom Pullum for comments on a previous version of this manuscript.

Figure 1: an example of reporting errors in full birth history data

Notes: DOB = Date of birth, AAD = Age at death. In this example, baby #1 and baby #2 were affected by date displacement. Baby #3 was affected by age over-statement, whereas baby #4 was an omission.

Figure 2: trends in neonatal mortality rate in urban areas of Guinea-Bissau

Notes: the MICS estimates pertain to all urban areas nationwide. The horizontal dashed lines are not confidence intervals, instead they reflect the time periods for which the estimate was calculated (i.e., the past 5 years prior to the survey). The solid markers represent the mid-point of that interval. Data from the HDSS (hollow circles), on the other hand, are annual estimates and only pertain to the areas of Bissau covered by the HDSS. The shaded grey area represents the 95% confidence interval around the time trend estimated from HDSS data.

Figure 3: flow chart of the study enrollment process

	Not interviewed (N = 177)	Interviewed (N=422)	P-value ^e
Characteristics of the participant			
<i>Age of the mother</i>			0.210
15-19y	17 (9.6)	19 (4.5)	
20-24y	53 (29.9)	115 (27.3)	
25-29y	39 (22.0)	106 (25.1)	
30-34y	38 (21.5)	108 (25.6)	
35-39y	23 (13.0)	60 (14.2)	
≥ 40y	7 (4.0)	14 (3.3)	
<i>Socio-economic status</i>			
Roof material of the house			0.579
Zinc	6 (3.4)	10 (2.4)	
Other	171 (96.6)	412 (97.6)	
Electricity available at the house			0.588
No	105 (59.3)	240 (56.9)	
Yes	72 (40.7)	182 (43.1)	
Toilet facilities at the house			0.551
Outside latrine	150 (84.8)	348 (82.5)	
Inside the house with plumbing	27 (15.2)	74 (17.5)	
<i>Educational level</i>			0.001
None	67 (40.6)	100 (24.2)	
Primary schooling	53 (32.1)	177 (42.9)	
Secondary and higher	45 (27.3)	136 (32.9)	
<i>Activity</i>			0.253
Does not work outside home	150 (90.9)	361 (87.2)	
Works outside home	15 (9.1)	53 (12.8)	
<i>Ethnic group^a</i>			0.021
Fula	59 (33.3)	102 (24.2)	
Pepel	38 (21.5)	112 (26.5)	
Manjaco	13 (7.3)	58 (13.7)	
Other	67 (37.9)	150 (35.6)	
Birth history of the participant			
<i>Neonatal deaths over past 5 years</i>			0.328
0	91 (51.4)	244 (57.8)	
1	83 (46.9)	171 (40.5)	
2	3 (1.7)	7 (1.7)	
<i>Post-neonatal deaths over past 5 years</i>			0.009
0	133 (75.1)	337 (79.9)	
1	40 (22.6)	85 (20.1)	
2	4 (2.3)	--	
<i>Stillbirths over past 5 years</i>			0.996
0	140 (79.1)	333 (78.9)	
1	36 (20.4)	85 (20.1)	
2+	1 (0.5)	4 (1.0)	
<i>Twin births over past 5 years</i>			0.859
None	166 (93.8)	392 (92.9)	
At least one	11 (6.2)	30 (7.1)	

Characteristics of the index pregnancy^b			
<i>Antenatal care (ANC)</i>			0.420
Did not seek ANC	3 (1.8)	14 (3.4)	
Sought ANC at least once	161 (98.2)	401 (96.6)	
<i>Maternal Tetanus Vaccination</i>			0.491
Never vaccinated	24 (15.2)	51 (12.7)	
Vaccinated prior or during pregnancy	134 (84.8)	351 (87.3)	
<i>Place of delivery</i>			0.003
Home	33 (18.6)	81 (19.2)	
National Hospital (Simao Mendes)	81 (45.8)	179 (42.2)	
Health center	25 (14.1)	108 (25.8)	
Other (e.g., private clinic)	38 (21.5)	54 (12.8)	
<i>Number of live births prior to index pregnancy</i>			0.696
None	56 (32.7)	130 (30.9)	
One or more	115 (67.3)	290 (69.1)	
<i>Mode of Delivery^c</i>			0.878
Vaginal	122 (88.4)	293 (87.7)	
C-section	16 (11.6)	41 (12.3)	
<i>Birth weight^{c, d}</i>			0.033
< 1,500 grams	6 (6.9)	14 (6.8)	
1,500–2,499 grams	25 (28.7)	32 (15.5)	
≥ 2,500 grams	56 (64.4)	161 (77.7)	

Table 1: Selectivity of the study sample and characteristics of participants, according to HDSS data

Notes: All data reported in these tables were extracted from HDSS records. Figures in parentheses are column percentages. Data are missing on some covariates, and thus counts reported in cells do not sum to column totals. Missing data occur for various reasons including incomplete HDSS records, or imperfect linkages with data from health facilities (e.g., mode of delivery and birth weight).

^a Fulas, Pepels and Manjacos are the largest ethnic groups. Other ethnic groups include primarily Mandingas, Bijagos, Saracole and Balantas.

^b These variables refer to the pregnancy, which prompted inclusion of the respondent in her sampling stratum. For example, if a respondent had 3 live births over the past 5 years including one that resulted in a neonatal death, and two that are still alive, according to the HDSS, then these variables refer to the neonatal death.

^c These variables were measured through record linkages, for deliveries that occurred at a health facility where the HDSS also collects data from clinical registers (i.e., national hospital and local health center). Missing data might thus also result from limited recording of clinical information in the registers of health facilities.

^d Data on birth weight was missing for approximately 40% of index pregnancies that were delivered in health facilities. The proportion of deliveries with missing birth weight information was larger in health centers and other health facilities, than in the national hospital (Simao Mendes).

^e Due to small sample sizes in some cells, all p-values were calculated using Fisher's two-sided exact test.

	Neonatal deaths		Post-neonatal deaths	
	Sensitivity	Specificity	Sensitivity	Specificity
Overall	0.791 (0.720 to 0.848)	0.993 (0.989 to 0.996)	0.691 (0.577 to 0.786)	0.999 (0.998 to 1.000)
Sub-groups				
Respondent w/ stillbirth	--	0.866 (0.748 to 0.934)	--	--
Respondent w/ post-neonatal death	--	0.840 (0.732 to 0.909)	--	--
Other respondents	--	1.000 (--)	--	--

Table 2: sensitivity and specificity of FBH data in recording neonatal and post-neonatal deaths

Notes: specificity is the proportion of respondents who did not experience a neonatal death during the reference period according to the HDSS, who also did not report such a death during the FBH survey; sensitivity is the proportion of respondents who experienced a neonatal death according to the HDSS during the reference period who also reported such a death during the survey. All figures are weighted to account for sampling design and non-response.

Figure 4: sources of error in FBH data

Notes: Date displacement refers to events reported to have occurred earlier/later than recorded by the HDSS, leading to erroneous exclusion/inclusion from the reference period. Age under-statement refer to errors resulting from the fact that a respondent reported her child's age at death too low compared to the age at death recorded in the HDSS dataset. Age over-statement refers to the opposite situation. Misclassification refers to events that were reported in a manner that does not correspond to HDSS records. For example, a respondent with a neonatal death according to HDSS who did not report such a death during the survey (false negative) might have reported a stillbirth instead. Omission refers to respondents with a neonatal death according to the HDSS who did not report such a death during the survey, and who also did not report a stillbirth nor a post-neonatal death.

Figure 5: Effects of reporting errors on survey estimates of the neonatal mortality rate

Notes: We used the mathematical model described in appendix A1, along with data from England & Wales (described in appendix A2); to calculate the survey estimates of the NMR that would have been obtained in a population if a FBH survey had been conducted with the reporting parameters observed in Bissau. The bias is obtained by dividing the calculated (counterfactual) survey estimate of the NMR by the true level of the NMR in the population of interest. A bias above 1 indicates that FBH data would over-estimate the level of NMR in, whereas a bias under 1 indicates that survey data would under-estimate the level of the NMR. The 95% confidence intervals represented in this figure are obtained from 1,000 draws from the distributions of the reporting parameters. Annual data from England and Wales were aggregated into consecutive 5-year periods (e.g., 1927-1931, 1932-1936) to ensure that our counterfactual survey estimates did not include overlapping sets of events.

REFERENCES:

- AbouZahr, C., D. de Savigny, L. Mikkelsen, P.W. Setel, R. Lozano, and A.D. Lopez. 2015. "Towards universal civil registration and vital statistics systems: the time is now." *Lancet* 386(10001):1407-1418.
- Agresti, A. 1992. "A survey of exact inference for contingency tables." *Statistical science* 7(1):131-153.
- Akuze, J., A. Baschieri, K. Kerber, V. Gordeev, P. Waiswa, H. Blencowe, D. Kwesiga, S. Floyd, and J. Lawn. 2017. "Do different HDSS surveillance systems result in different quality of pregnancy outcome data?".
- Alam, N., T. Ali, A. Razzaque, M. Rahman, M. Zahirul Haq, S.K. Saha, A. Ahmed, A.M. Sarder, M. Moinuddin Haider, M. Yunus, Q. Nahar, and P. Kim Streatfield. 2017. "Health and Demographic Surveillance System (HDSS) in Matlab, Bangladesh." *Int J Epidemiol* 46(3):809-816.
- Amo-Adjei, J. and S.K. Annum. 2015. "Socioeconomic determinants of birth registration in Ghana." *BMC Int Health Hum Rights* 15:14.
- Armstrong Schellenberg, J.R., R. Nathan, S. Abdulla, O. Mukasa, T.J. Marchant, M. Tanner, and C. Lengeler. 2002. "Risk factors for child mortality in rural Tanzania." *Tropical Medicine & International Health* 7(6):506-511.
- Baschieri, A., V. Gordeev, J. Akuze, D. Kwesiga, H. Blencowe, S. Cousens, P. Waiswa, A.B. Fisker, S.M. Thyssen, A. Rodrigues, S. Kasasa, E. Galiwango, D. Natukwatsa, D. Kajungu, Y.A. Enuameh, C. Tawiah, K.P. Asante, S. Owusu-Agyei, N. Alam, M.M. Haider, S.S. Alam, F. Arnold, P. Byass, T.N. Croft, A.J. Herbst, S. Kishor, F. Serbanescu, and J. Lawn. 2019. "Every Newborn-INDEPTH" (EN-INDEPTH) study protocol for a randomised comparison of household survey modules for measuring stillbirths and neonatal deaths in five Health and Demographic Surveillance site." *Journal of Global Health*, 9(1): 010901.
- Bjerregaard-Andersen, M., N. Lund, A.S. Pinstrup Joergensen, F. Starup Jepsen, H. Werner Unger, M. Mane, A. Rodrigues, S. Bergstrom, and C. Stabell Benn. 2018. "Stillbirths in urban Guinea-Bissau: A hospital- and community-based study." *PLoS One* 13(5):e0197680.
- Burstein, R., H. Wang, R.C. Reiner, Jr., and S.I. Hay. 2018. "Development and validation of a new method for indirect estimation of neonatal, infant, and child mortality trends using summary birth histories." *PLoS Med* 15(10):e1002687.
- Davey, G. and K. Deribe. 2017. "Precision public health: mapping child mortality in Africa." *Lancet* 390(10108):2126-2128.
- Espeut, D. and S. Becker. 2015. "The validity of birth and pregnancy histories in rural Bangladesh." *J Health Popul Nutr* 33:17.

- Finlay, J.E., C. Moucheraud, S. Goshev, F. Levira, S. Mrema, D. Canning, H. Masanja, and A.E. Yamin. 2015. "The effects of maternal mortality on infant and child survival in rural Tanzania: a cohort study." *Maternal and child health journal* 19(11):2393-2402.
- Fisker, A.B., A. Rodrigues, and S. HELLERINGER. 2019. "Differences in barriers to birth and death registration in Guinea-Bissau: implications for monitoring national and global health objectives." *Trop Med Int Health*, 24(2): 166–174.
- Guillot, M., P. Gerland, F. Pelletier, and A. Saabneh. 2012. "Child mortality estimation: a global overview of infant and child mortality age patterns in light of new empirical data." *PLoS Med* 9(8):e1001299.
- Hallett, T.B., S. Gregson, F. Kurwa, G.P. Garnett, S. Dube, G. Chawira, P.R. Mason, and C.A. Nyamukapa. 2010. "Measuring and correcting biased child mortality statistics in countries with generalized epidemics of HIV infection." *Bull World Health Organ* 88(10):761-768.
- Haws, R.A., I. Mashasi, M. Mrisho, J.A. Schellenberg, G.L. Darmstadt, and P.J. Winch. 2010. ""These are not good things for other people to know": how rural Tanzanian women's experiences of pregnancy loss and early neonatal death may impact survey data quality." *Soc Sci Med* 71(10):1764-1772.
- Helleringer, S., D. Arhinful, B. Abuaku, M. Humes, E. Wilson, A. Marsh, A. Clermont, R.E. Black, J. Bryce, and A. Amouzou. 2018. "Using community-based reporting of vital events to monitor child mortality: Lessons from rural Ghana." *PLoS One* 13(1):e0192034.
- Helleringer, S., G. Pison, A.M. Kante, G. Duthe, and A. Andro. 2014a. "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.
- Helleringer, S., G. Pison, B. Masquelier, A.M. Kante, L. Douillot, G. Duthe, C. Sokhna, and V. Delaunay. 2014b. "Improving the quality of adult mortality data collected in demographic surveys: validation study of a new siblings' survival questionnaire in Niakhar, Senegal." *PLoS Med* 11(5):e1001652.
- Helleringer, S., G. Pison, B. Masquelier, A.M. Kante, L. Douillot, C.T. Ndiaye, G. Duthe, C. Sokhna, and V. Delaunay. 2015. "Improving survey data on pregnancy-related deaths in low-and middle-income countries: a validation study in Senegal." *Trop Med Int Health* 20(11):1415-1423.
- Hill, K. and Y. Choi. 2006. "Neonatal mortality in the developing world." *Demographic Research* 14:429-452.
- UN Inter-Agency Group for Child Mortality Estimate (IGME). 2018. "Levels & Trends in Child Mortality: Report 2018, Estimates developed by the United Nations Inter-agency Group for Child Mortality Estimation." New York: United Nations Children's Fund.
- Instituto Nacional de Estatística (INE). 2014. "Inquérito aos Indicadores Múltiplos (MICS5) 2014, Relatório Final." Bissau, Guiné-Bissau.: Ministério da Economia e Finanças, Direção Geral do Plano, Instituto Nacional de Estatística (INE).

Khan, S. and A. Hancioglu. 2019. "Multiple Indicator Cluster Surveys: Delivering Robust Data on Children and Women across the Globe." *Stud Fam Plann* 50(3):279-286.

Lawn, J.E., H. Blencowe, S. Oza, D. You, A.C. Lee, P. Waiswa, M. Lalli, Z. Bhutta, A.J. Barros, P. Christian, C. Mathers, S.N. Cousens, and G. Lancet Every Newborn Study. 2014. "Every Newborn: progress, priorities, and potential beyond survival." *Lancet* 384(9938):189-205.

Lawn, J.E., D. Osrin, A. Adler, and S. Cousens. 2008. "Four million neonatal deaths: counting and attribution of cause of death." *Paediatr Perinat Epidemiol* 22(5):410-416.

Liu, L., H.D. Kalter, Y. Chu, N. Kazmi, A.K. Koffi, A. Amouzou, O. Joos, M. Munos, and R.E. Black. 2016. "Understanding Misclassification between Neonatal Deaths and Stillbirths: Empirical Evidence from Malawi." *PLoS One* 11(12):e0168743.

Liu, L., S. Oza, D. Hogan, J. Perin, I. Rudan, J.E. Lawn, S. Cousens, C. Mathers, and R.E. Black. 2015. "Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis." *Lancet* 385(9966):430-440.

Malqvist, M. 2011. "Neonatal mortality: an invisible and marginalised trauma." *Glob Health Action* 4(10).

Masquelier, B., G. Reniers, and G. Pison. 2014. "Divergences in trends in child and adult mortality in sub-Saharan Africa: survey evidence on the survival of children and siblings." *Popul Stud (Camb)* 68(2):161-177.

Mikkelsen, L., D.E. Phillips, C. AbouZahr, P.W. Setel, D. de Savigny, R. Lozano, and A.D. Lopez. 2015. "A global assessment of civil registration and vital statistics systems: monitoring data quality and progress." *Lancet* 386(10001):1395-1406.

Moultrie, T., R. Dorrington, A. Hill, K. Hill, I. Timæus, and B. Zaba. 2013. *Tools for demographic estimation*: International Union for the Scientific Study of Population.

Moxon, S.G., H. Ruysen, K.J. Kerber, A. Amouzou, S. Fournier, J. Grove, A.C. Moran, L.M. Vaz, H. Blencowe, N. Conroy, A. Gulmezoglu, J.P. Vogel, B. Rawlins, R. Sayed, K. Hill, D. Vivio, S.A. Qazi, D. Sitrin, A.C. Seale, S. Wall, T. Jacobs, J. Ruiz Pelaez, T. Guenther, P.S. Coffey, P. Dawson, T. Marchant, P. Waiswa, A. Deorari, C. Enweronu-Laryea, S. Arifeen, A.C. Lee, M. Mathai, and J.E. Lawn. 2015. "Count every newborn; a measurement improvement roadmap for coverage data." *BMC Pregnancy Childbirth* 15 Suppl 2:S8.

United Nations. 2017. "United Nations Department of Economic and Social Affairs/Population Division (2017): World Population Prospects: The 2017 Revision." New York, USA.

Obermeyer, Z., J.K. Rajaratnam, C.H. Park, E. Gakidou, M.C. Hogan, A.D. Lopez, and C.J. Murray. 2010. "Measuring adult mortality using sibling survival: a new analytical method and new results for 44 countries, 1974-2006." *PLoS Med* 7(4):e1000260.

- Pullum, T. and S. Becker. 2014. "Evidence of Omission and Displacement in DHS Birth Histories." in *DHS Methodological Reports No. 11*. Rockville, Maryland, USA: ICF International.
- Pullum, T. and S. Staveteig. 2017. "An assessment of the quality and consistency of age and date reporting in DHS Surveys, 2000-2015." in *DHS Methodological Report No. 19*. Rockville, Maryland, USA: ICF International.
- Pullum, T.W. 2006. "An assessment of age and date reporting in the DHS Surveys 1985-2003." in *DHS Methodological Reports No. 5*. Rockville, Maryland, USA: ICF International.
- Rodrigues, A., J.A. Schellenberg, P.E. Kofoed, P. Aaby, and B. Greenwood. 2008. "Changing pattern of malaria in Bissau, Guinea Bissau." *Trop Med Int Health* 13(3):410-417.
- Setel, P.W., S.B. Macfarlane, S. Szreter, L. Mikkelsen, P. Jha, S. Stout, C. AbouZahr, and E. Monitoring of Vital. 2007. "A scandal of invisibility: making everyone count by counting everyone." *Lancet* 370(9598):1569-1577.
- Snow, J. and M. Mann. 2013. "Qualtrics survey software: handbook for research professionals." *Qualtrics Labs, Inc.*
- Spector, J.M. and S. Daga. 2008. "Preventing those so-called stillbirths." *Bull World Health Organ* 86(4):315-316.
- United Nations. 2018. "Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development." New York: United Nations.
- Walker, N., K. Hill, and F. Zhao. 2012. "Child mortality estimation: methods used to adjust for bias due to AIDS in estimating trends in under-five mortality." *PLoS Med* 9(8):e1001298.
- Wixted, J.T. 2004. "The psychology and neuroscience of forgetting." *Annu Rev Psychol* 55:235-269.
- Zaba, B., J. Whitworth, M. Marston, J. Nakiyingi, A. Ruberantwari, M. Urassa, R. Issingo, G. Mwaluko, S. Floyd, and A. Nyondo. 2005. "HIV and mortality of mothers and children: evidence from cohort studies in Uganda, Tanzania, and Malawi." *Epidemiology*:275-280.

APPENDIX A1: A MODEL OF THE EFFECTS OF REPORTING ERRORS IN FBH DATA ON SURVEY ESTIMATES OF THE NMR

We developed a mathematical model to understand the net effects of reporting errors in FBH data on survey-based estimates of the NMR.

Consider a hypothetical population of W women, who have had B live births over a recent period of reference (e.g., the past 5 years). Among these B live births, N resulted in a neonatal death; P resulted in a post-neonatal death; and C resulted in a death at ages 1 to 4 years old. We note L the children who are still alive at time t , and who may be aged between 0 and 4 years old. Then we have $L = B - N - P - C$. In addition, these W women experienced a number of stillbirths, noted S .

In this hypothetical population, we conduct a FBH survey. We used upper-case letters to denote “true” quantities pertaining to the hypothetical population of interest, but we now use lower-case letters to describe the data obtained during this FBH survey. Thus, b is the number of live births reported by survey respondents; n is the reported number of neonatal deaths; p is the reported number of post-neonatal deaths; and c is the reported number of deaths among children 1-4 years old. As in the population, the children 0-4 years old who are reported as still alive at time t are noted l , and we have $l = b - n - p - c$. The true NMR in the population of interest is thus $\frac{N}{B}$, whereas the estimate of the NMR obtained from data generated by the FBH survey is given by $\frac{n}{b}$. We seek to describe the relations between those two quantities.

We define conditional probabilities that describe the pattern of reporting errors during the FBH survey. In our notation, the upper-case subscript indicates the “true” nature of an event in the population, whereas the lower subscript indicates how it was reported during FBH. We thus call $R_{N \rightarrow n}$ the probability that a neonatal death is reported as such during the FBH survey, $R_{P \rightarrow p}$ the probability that a post-neonatal death is reported as such in the FBH survey, and $R_{C \rightarrow c}$ the probability that the death of a child aged 1-4 years old is reported as such during the FBH survey. Similarly, $R_{S \rightarrow n}$ is the probability that a stillbirth is misclassified as a neonatal death during the survey; $R_{P \rightarrow n}$ is the probability that a post-neonatal death is reported as a neonatal death during the survey and $R_{C \rightarrow n}$ is the probability that a child who died at ages 1-4 years old is reported as a neonatal death during the survey. In this model, $R_{P \rightarrow n}$ and $R_{C \rightarrow n}$ represent the extent of age under-statement in FBH data that affects the reporting of neonatal deaths. For completeness, we also define $R_{L \rightarrow n}$ as the probability that a surviving child aged 0-4 years old is reported as a neonatal death during the FBH survey.

We also define conditional probabilities indicating the likelihood that an event that is incorrectly reported during the survey is nonetheless reported as a live birth that should be included in the denominator of the NMR. We thus call $R_{N \rightarrow o}$ the proportion of neonatal deaths that are not reported as neonatal deaths during the FBH, but are nonetheless listed as live births having occurred in the reference period (e.g., they may be reported as post-neonatal deaths); $R_{P \rightarrow o}$ is the proportion of post-neonatal deaths that are not reported as such but are nonetheless listed as live births having occurred during the reference period (e.g., they are reported as neonatal deaths or as deaths above age 1); and $R_{C \rightarrow o}$ is the proportion of deaths at ages 1-4 that are not reported as such during the

FBH, but are nonetheless listed live births having occurred during the reference period. Finally, $R_{S \rightarrow o}$ is the proportion of stillbirths that are reported as live births during the FBH.

Then, we can express the survey estimate of the NMR as a function of 1) the true frequency of various events in the population (e.g., N , S or P) and 2) the conditional probabilities that these events are reported during the FBH survey. We thus have:

$$n = N \times R_{N \rightarrow n} + S \times R_{S \rightarrow n} + P \times R_{P \rightarrow n} + C \times R_{C \rightarrow n} + L \times R_{L \rightarrow n} \quad (1)$$

and

$$b = N \times [R_{N \rightarrow n} + (1 - R_{N \rightarrow n}) \times R_{N \rightarrow o}] + S \times R_{S \rightarrow o} + P \times [R_{P \rightarrow p} + (1 - R_{P \rightarrow p}) \times R_{P \rightarrow o}] + C \times [R_{C \rightarrow c} + (1 - R_{C \rightarrow c}) \times R_{C \rightarrow o}] + L \times [R_{L \rightarrow l} + (1 - R_{L \rightarrow l}) \times R_{L \rightarrow o}] \quad (2)$$

In this framework, $R_{N \rightarrow n}$ and $R_{P \rightarrow p}$ are the sensitivity of FBH in recording neonatal and post-neonatal deaths, respectively. The specificity of FBH in recording neonatal deaths, on the other hand, is a weighted function of $R_{S \rightarrow n}$, $R_{P \rightarrow n}$, $R_{C \rightarrow n}$ and $R_{L \rightarrow n}$ where the weights are supplied by the population proportions of women in each of those strata. In table A1.1 below, we summarize the various parameters included in this model, and the source of information on their values and variation.

To analyze this model, we apply it to data on the incidence of stillbirths, neonatal deaths post-neonatal deaths, and child deaths (i.e., S , N , P and C) available for England and

Wales (1927 – 2017, see appendix A2). We obtain predicted values of NMR estimates that would have been obtained if a survey had been conducted in these populations according to the reporting parameters observed in Bandim (see table A1.1).

We make several simplifying hypotheses in analyzing this model:

1. We assume that $R_{S \rightarrow n} = R_{S \rightarrow o}$, i.e., all the misclassified stillbirths are reported as neonatal deaths. This is supported by our results in table 2 of the main text: we did not find any misclassifications in which a stillbirth was reported as a post-neonatal death or child death, or as a live child. Instead, we found that all the misclassifications of stillbirths were reported as early neonatal deaths (appendix A6).
2. We assume that $R_{C \rightarrow n} = 0$, i.e., no deaths that occurred at ages 1-4 years old are transferred to the neonatal period due to age errors. This assumption is supported by our findings in table 2, where we found a specificity of 100% for mothers whose child had either died at ages 1-4 or were still alive.
3. We assume that $R_{L \rightarrow n} = 0$, i.e., no children who are still alive are being reported as having died during the neonatal period. This assumption is supported by our findings in table 2, where we found a specificity of 100% for mothers whose child had either died at ages 1-4 or were still alive.

There are also several parameters that our validation study does not allow estimating.

These include $R_{C \rightarrow c}$, $R_{C \rightarrow o}$, $R_{L \rightarrow l}$ and $R_{L \rightarrow o}$. This is the case because we did not survey a sufficient number of women whose children had reached age 1, and because the

Bandim HDSS only follows up children up to age 3. To remedy this limitation, we first explored how these parameters affect the extent of bias in survey estimates of the NMR. We conducted one-way analyses: we set all parameters in equations (1) and (2) at their perfect value, e.g., $R_{N \rightarrow n} = 1$ and we let the parameters of interest vary over a relevant range. We present the results of these analyses below (figures A1.1 and A1.2).

These one-way analyses indicate that $R_{C \rightarrow c}$, $R_{C \rightarrow o}$ and $R_{L \rightarrow o}$ have very limited impact on the extent of bias in survey estimates of the NMR (Figure A1.1). This is so for different reasons. For $R_{C \rightarrow c}$ and $R_{C \rightarrow o}$, this is the case because these reporting errors generally apply to a small number of events, i.e., C . In England and Wales, as in other settings, mortality at ages 1-4 has declined more rapidly than mortality during infancy (appendix A2). The resulting bias is thus less than 1% in the settings analyzed here. For $R_{L \rightarrow o}$, this is the case because errors on the vital status of children (i.e., reporting a live child as deceased and vice-versa) are exceedingly rare. They might result from occasional interviewers' mistakes, but these are often corrected when different sets of follow-up questions must be asked about live vs. deceased children. In our analyses, we thus set $R_{L \rightarrow o} = 0$.

On the other hand, small variations in $R_{L \rightarrow l}$ have a fairly large impact on the bias in survey estimates of the NMR. This is so because a) $R_{L \rightarrow l}$ is applied to a large number of surviving children L , and 2) variations in $R_{L \rightarrow l}$ only affect the denominator of the NMR without (possibly) compensatory impact on its numerator. Indeed, we did not document any false positive reports of neonatal deaths among mothers whose children were all alive at the time of the survey (table 2 of the main text). In our analyses (figure A1.2), $R_{L \rightarrow l}$ was inversely related to the extent of bias in NMR estimate. For example $R_{L \rightarrow l} = 0.8$

would lead to overestimating the NMR by more than 20%, whereas $R_{L \rightarrow l} = 0.95$ would lead to a 5% overestimate. The effects of $R_{L \rightarrow l}$ on bias in survey estimates vary only slight with the underlying level of neonatal mortality (figure A1.1.)

Figure A1.1: one-way analysis of the effects of $R_{C \rightarrow c}$ and $R_{C \rightarrow o}$ on bias in survey estimates of the NMR.

Notes: in this analysis, all other reporting parameters in equations (1) and (2) were set at their perfect values, e.g., $R_{N \rightarrow n} = 1$. We let $R_{C \rightarrow o}$ vary between 0.1 and 0.3 because we have already set $R_{C \rightarrow n} = 0$ (see above), and $R_{C \rightarrow p}$ is limited (see figure 4 in the main text). As a result, the large majority of deaths at ages 1-4 years old that are not reported as such during a FBH are either transferred to earlier time periods, or shifted to older age groups, and thus do not enter the calculation of the denominator of the NMR.

In Table A1.1, we summarize all the values used in our analyses of the model in equations (1) and (2). Most of the parameters are estimated from the study data, and we report their standard errors. For the parameters whose value could not be estimated

during this study, we reviewed available literature on their estimation. Specifically, we relied on the analyses of displacements and omissions in FBH data conducted by Pullum and Becker (2014)¹. They analyzed data collected during close to 200 Demographic and Health Surveys conducted in low- and middle-income countries. Using a series of internal consistency checks, they estimated that omissions and displacements of live births concerned generally <2% of events each, except in some surveys where they might have reached values as high as 10%. Surveys with high values for displacement and omissions were predominantly conducted during the earlier phases of the DHS program, e.g., in the 1990's or early 2000's. Conservatively, we adopt the median values reported by Pullum and Becker to inform our calculations. We also considered that children who are still alive at the time of the survey are not reported as deceased. Thus, omissions and displacements of live births are the only sources of errors. We use this information conservatively to set $R_{L \rightarrow l} = 1 - 0.02 - 0.02 = 0.96$.

In selecting a realistic value of $R_{C \rightarrow c}$, we consider that a death between age 1-4 might a) be misreported as a death at ages <1 year old, b) be misreported as a death at ages 5-9 years old, c) be misreported as a death having occurred prior to the reference period, and d) might be omitted altogether from FBH. Pullum and Becker (2014) provide some guidance about these various aspects. They suggest that displacements and omissions of deaths are more common than for live children, and may routinely reach 10% for each of these aspects in surveys. We thus set $R_{C \rightarrow c} = 1 - 0.1 - 0.1 = 0.8$. Finally, we set $R_{C \rightarrow o} = 0.1$ based on information about the prevalence of age under-statement in

¹ Pullum, T. and S. Becker, "Evidence of omission and displacement in DHS birth histories", *DHS Methodological Reports No. 11*, Rockville, Maryland. Available at: <https://dhsprogram.com/publications/publication-MR11-Methodological-Reports.cfm>

misreport post-neonatal deaths (see figure 4 of the main text), and the relative frequency of displacements and omissions according to Pullum and Becker (2014).

Figure A1.2: one-way analysis of the effects of $R_{L \rightarrow l}$ on bias in survey estimates of the NMR.

Notes: in this analysis, all other reporting parameters in equations (1) and (2) were set at their perfect values, e.g., $R_{N \rightarrow n} = 1$.

Parameter	Value	Standard error	Source	Assumptions
$R_{N \rightarrow n}$	0.7912	0.0326	Table 2	--
$R_{S \rightarrow n}$	1-0.8662	0.0461	Table 2	--
$R_{P \rightarrow n}$	1-0.8396	0.0448	Table 2	--
$R_{C \rightarrow n}$	0	N/A	Table 2	No deaths at ages 1-4 reported as neonatal deaths
$R_{L \rightarrow n}$	0	N/A	Table 2	No live children reported as neonatal deaths
$R_{N \rightarrow o}$	0.08182	0.0492	Figure 3	--
$R_{S \rightarrow o}$	1-0.8662	0.0461	Table 2	All misclassified stillbirths reported as neonatal deaths
$R_{P \rightarrow p}$	0.6908	0.0538	Table 2	--
$R_{P \rightarrow o}$	0.4854	0.1021	Figure 3	--
$R_{C \rightarrow c}$	0.8000	N/A	Pullum and Becker (2014)	--
$R_{C \rightarrow o}$	0.1000	N/A	Pullum and Becker (2014) and figure 3	--
$R_{L \rightarrow l}$	0.9600	N/A	Pullum and Becker (2014)	--
$R_{L \rightarrow o}$	0	N/A	N/A	No live children misreported as children who have died in 0–4 age group

Table A1.1: summary of parameters and values used in modeling of bias in survey estimates of the NMR

Notes: "N/A" = Not applicable. For the parameters drawn from the literature, we use a single value and ignore potential uncertainty in the value of these parameters. All other parameters take into account sampling variability as indicated by estimates of the standard error.

APPENDIX A2: TIMES SERIES OF NEONATAL MORTALITY, POST-NEONATAL MORTALITY, CHILD MORTALITY AND STILLBIRTH RATES IN ENGLAND AND WALES, 1927-2017.

Notes: from these data, we obtain the values of S , N , P and C used to calculate survey estimates of the NMR that are reported in figure 5 of the main text.

APPENDIX A3: ROBUSTNESS TESTS

We explore how our findings might be affected by errors in reference HDSS data. In doing so, we follow a counterfactual approach: how would our assessment of the extent of bias in NMR estimates (figure 5 of the main text) change if we modified the reference HDSS data in pre-defined ways that reflect plausible patterns of errors?

The errors in HDSS data that we considered in our robustness tests include:

1. Misclassifications of stillbirths as neonatal deaths, and vice-versa;
2. Omissions of neonatal deaths due to incomplete pregnancy records.

These errors might emerge because of incomplete coverage of household visits by HDSS workers, limited linkages with data from health facilities or difficulties in adequately classifying events that occur at home, or even in health facilities (e.g., stillbirths vs. neonatal deaths). We describe below the available studies on errors in HDSS data, and how they might inform our robustness tests. Then we describe how we carried out those tests. Finally, we describe the results of those tests.

Misclassifications of stillbirths recorded in HDSS data: There is limited guidance about the prevalence of misclassifications between stillbirths and neonatal deaths in HDSS data collected in Bandim HDSS (as in other HDSS). Existing studies of the extent of misclassifications between stillbirths and neonatal deaths² have primarily been focused on FBH data. Such FBH data are often collected long after the delivery/death, and only rarely rely on clinical records to classify events and pregnancy outcomes. In contrast to FBH data, HDSS

² See for example Liu, L. et al., “Understanding misclassification between neonatal deaths and stillbirths: empirical evidence from Malawi”, *PLOS One*, 2016:10.1371/journal.pone.0168743.

data collected in Bissau are systematically linked to clinical records for deliveries that occurred in major health facilities when feasible. They are also collected within a few days of the delivery. We thus expect fewer misclassifications between stillbirths and neonatal deaths in the HDSS datasets than in these studies having evaluated FBH data.

A recent analysis of medical records from the National Hospital in Bissau³ indicated that <1% of pregnancy outcomes that were recorded as stillbirths by hospital workers had a positive Apgar score, i.e., they showed signs of life.⁴ They might thus have been “true” neonatal deaths that were misclassified as stillbirths by health workers in medical records. Additional stillbirths recorded at this hospital might also have been misclassified, even though they received an Apgar score of 0 (or had missing Apgar). In other, less equipped, health facilities (e.g., health centers), or in deliveries that occurred at home, the likelihood of misclassifying neonatal deaths as stillbirths might be higher, e.g., because attendants are not properly trained in detecting such signs, for example.

Conservatively, we thus considered two potential levels of misclassification among stillbirths recorded by the HDSS: one in which 5% of all stillbirths recorded by the HDSS were in fact neonatal deaths, and one in which 10% of all stillbirths recorded by the HDSS were in fact neonatal deaths.

³ Bjerregaard-Andersen, M. et al., “Stillbirths in Urban Guinea-Bissau: a hospital and community-based study”, *PLOS One*, 2018: doi.org/10.1371/journal.pone.0197680.

⁴ The Apgar score is a measure ranging from 0 to 10, which summarizes the health of a newborn one-minute and (often) 5 minutes after birth. It is based on assessments of a newborn’s color, heart rate, reflexes, muscle tone and respiration. An Apgar score greater than zero is considered an indication that the newborn showed “signs of life” and should not be classified as a stillbirth.

Misclassifications of neonatal deaths recorded in HDSS data: We do not have similar information on the likelihood that a recorded neonatal death in HDSS datasets was in fact a stillbirth. However, there are indications, in the global health literature and in the literature on health systems,⁵ that this occurrence might be less common than the opposite (i.e, recorded stillbirths are in fact neonatal deaths). This is so for several reasons. In many settings, the equipment required to detect a heartbeat (e.g., stethoscope, fetal monitors) might be unavailable. As a result, clinicians might not be able to detect faint heartbeats and may thus be more likely to classify an adverse pregnancy outcome as a stillbirth rather than as a neonatal death. Similarly, neonatal deaths may prompt audits and inquiries from various actors of the health system, whereas supervising clinicians and health systems managers less frequently investigate stillbirths. In such contexts, health workers involved in the delivery have incentives to (mis)-record the event as a stillbirth rather than accurately record it as a neonatal death, in order to avoid such supervision and audits. Finally, the stigma associated with a neonatal death might be larger than the stigma associated with a stillbirth.⁶ Parents and families might thus also be more likely to report that a neonatal death was a stillbirth, rather than the opposite.

We thus considered two patterns of misclassifications in our robustness tests: a “symmetric” pattern in which the proportion of misclassified events in HDSS data is the same for stillbirths and neonatal deaths, and an “asymmetric” pattern in which the proportion of stillbirths that have been misclassified in HDSS datasets is larger than the proportion of neonatal deaths that have been misclassified. In our asymmetric pattern, we consider that the proportion of

⁵ See for example Spector, J. M. and S. Daga, “Preventing these so-called stillbirths”, *Bulletin of the World Health Organization*, 2008; 86(4): 315-316.

⁶ See for example, Haws, R.A. et al. ““These are not good things for other people to know”: How rural Tanzanian women’s experience of pregnancy loss and early neonatal death may impact survey data quality”, *Social Science & Medicine*, 2010, 71(10): 1764-1772.

HDSS-recorded neonatal deaths that were misclassified was 60% lower than the proportion of HDSS-recorded stillbirths that were misclassified. We discuss below how our robustness tests are affected if this relative accuracy of the HDSS in recording stillbirths vs. neonatal deaths is lower/higher.

Incomplete pregnancy records in HDSS datasets: HDSS fieldworkers conducting prospective monitoring of pregnancies in Bissau might have failed to record some of the pregnancies that have occurred to members of the HDSS population. Evaluating the frequency of such omissions requires comparing HDSS data on pregnancies with another independent data collection system recording pregnancies. Such a study has not been conducted for the HDSS in Bissau. Instead, we used data from such a study conducted in the Iganga-Mayuge HDSS in Uganda⁷ to inform our robustness tests.

In the study in Uganda, the completeness of pregnancy records in the HDSS datasets varied from year to year. In developing our robustness tests, conservatively, we considered the year during which the completeness of HDSS records on pregnancies was the lowest. In addition, in the study conducted in Iganga-Mayuge, the pregnancies that were not captured by the HDSS were disproportionately pregnancies that had ended in fetal loss or abortion before 28 weeks of gestation (169 out of 314 pregnancies missed by the HDSS, 53.8%). The outcomes of such pregnancies are unlikely to be classified as neonatal deaths during a FBH interview. Among pregnancies that had lasted more than 28 weeks, the estimated completeness of HDSS data on pregnancies in Iganga-Mayuge was 94% (2,226 out of 2,371 pregnancies were recorded by the HDSS).

⁷ See Kadobera, D. et al. “Comparing performance of methods used to identify pregnant women, pregnancy outcomes and child mortality in the Iganga-Mayuge Health and Demographic Surveillance Site, Uganda”, *Global Health Action*, 2016, 10: 1356641.

Based on this information, we consider two completeness levels for the HDSS records of pregnancies: one “low” level of completeness in which the HDSS captures 90% of pregnancies, i.e., slightly less complete than what was observed in Iganga-Mayuge; and one “high” level of completeness in which the HDSS records 95% of pregnancies. The high level of completeness might be more reflective of the completeness of the HDSS pregnancy records in Bissau, because the HDSS uses a protocol in which the frequency of household visits by HDSS fieldworkers is much higher than in Iganga-Mayuge (2 times a year in Iganga-Mayuge vs. monthly in Bissau). In addition, in the Bandim HDSS in Bissau, record linkages with health facilities might allow recording a number of the pregnancies that might have been missed during household visits but have recorded during antenatal care.

Relative risk of neonatal death in unrecorded pregnancies: The pregnancies that were not recorded by the HDSS might also constitute a selective subset of all pregnancies that have occurred in the HDSS population. In particular, pregnancies that have led to an adverse perinatal outcome (e.g., a stillbirth, a neonatal death) might be more prevalent among unrecorded pregnancies than pregnancies that have led to other outcomes. This might be so, for example due to stigma and social desirability bias that might limit women’s willingness to disclose a pregnancy during monthly household visits. It might also happen, for example, if the HDSS is less likely to detect pregnancies in which antenatal care is not sought, and thus matching medical records are not available from health facilities.

We considered two levels of risk of neonatal death in unrecorded pregnancies. Specifically, we considered a “low-risk” context, in which the prevalence of neonatal deaths observed in unrecorded pregnancies was similar to the prevalence observed in recorded pregnancies; and another “high-risk” scenario in which this prevalence was 50% higher than the prevalence of

neonatal deaths observed in recorded pregnancies. We also discuss below how levels of relative risk higher than those considered here might affect our robustness tests.

Scenarios and implementation: Based on the discussions above, we defined 16 scenarios for our robustness tests. The parameters and hypotheses that characterize those scenarios are presented in table A3.1 below.

To implement these scenarios, we proceeded as follows:

1. We randomly selected a percentage of stillbirths and neonatal deaths recorded by the HDSS, and we reclassified them as neonatal deaths and stillbirths, respectively.
2. We randomly selected a subset of the respondents in our study,⁸ and for those who were selected, we created an additional HDSS pregnancy record.⁹
3. We randomly imputed the outcome of this additional pregnancy (stillbirth, neonatal death, post-neonatal death or surviving child), according to the various scenarios of relative risk defined in table A3.1.
4. In this reclassified and augmented HDSS dataset, we recalculated the main study outcomes described in table 2 of the main paper (i.e., sensitivity and specificity).
5. We repeated steps 1 through 4 in 1,000 replications, and we examined the distribution of recalculated study parameters across those replicates. We did so using the `simulate` command in Stata.

⁸ We excluded a small number of respondents ($n = 8$) from this step for whom pregnancy interval data from the HDSS suggested that the HDSS recorded all pregnancies that have lasted > 28 weeks for that person. For example, it would not be possible to add another pregnancy to the records of a woman who had had 4 pregnancies of > 28 weeks in the past 5 years prior to the survey according to the HDSS.

⁹ In our HDSS datasets, there were 866 records of pregnancies for our study respondents. In scenarios where we assumed that the HDSS pregnancy records were 95% complete, we thus randomly selected $(866/0.95) - 866 = 46$ respondents to receive an additional pregnancy record, whereas in scenarios where the completeness of HDSS pregnancy records was assumed at 90%, we randomly selected $(866/0.90) - 866 = 96$ respondents to receive an additional pregnancy record.

In implementing these scenarios, we made a number of simplifying assumptions. First, we did not modify the FBH reports collected during the study; only the HDSS data were augmented and reclassified in these robustness tests. Second, we hypothesized that errors in HDSS data were independent of the errors observed in FBH data, i.e. that the HDSS records of women with false positives or false negatives in FBH data were not more or less likely to be affected by errors than the HDSS records of women with true positives/true negatives. Third, we did not consider the possibility of age and date errors in HDSS datasets, i.e., errors in the recording of the timing of events. Instead, we considered that when the HDSS records an event, it records its date accurately. Fourth, we also did not consider the possibility of under-coverage of the HDSS datasets, i.e., systematic non-inclusion of specific population groups during regular census rounds conducted by the HDSS. Finally, in adding pregnancy records to account for unrecorded pregnancies in HDSS data, we considered that 1) the HDSS missed at most one pregnancy per respondent, and 2) all unrecorded pregnancies led to singleton births.

Scenario type	Scenario number	Re-classified stillbirths	Re-classified neonatal deaths	Completeness of HDSS pregnancy records ^c	Relative risk of neonatal death in unrecorded pregnancies
		(1)	(2)	(3)	(4)
<i>Symmetric^a</i>	1	5%	5%	95%	Equal
	2	5%	5%	95%	50% higher
	3	5%	5%	90%	Equal
	4	5%	5%	90%	50% higher
	5	10%	10%	95%	Equal
	6	10%	10%	95%	50% higher
	7	10%	10%	90%	Equal
	8	10%	10%	90%	50% higher
<i>Asymmetric^b</i>	1	5%	3%	95%	Equal
	2	5%	3%	95%	50% higher
	3	5%	3%	90%	Equal
	4	5%	3%	90%	50% higher
	5	10%	6%	95%	Equal
	6	10%	6%	95%	50% higher
	7	10%	6%	90%	Equal
	8	10%	6%	90%	50% higher

Table A3.1: Scenarios of HDSS data accuracy implemented in robustness tests.

Notes: ^a *Symmetric scenarios* are scenarios in which the proportion of stillbirths that have been misclassified by the HDSS is equal to the proportion of neonatal deaths that have been misclassified by the HDSS; ^b *Asymmetric scenarios* are scenarios in which the proportion of stillbirths that have been misclassified by the HDSS is greater than the proportion of neonatal deaths that have been misclassified by the HDSS; ^c Completeness is defined as the proportion of all pregnancies that have occurred among HDSS residents that have been captured by the HDSS.

Results: Our robustness tests indicated that potential errors in HDSS data would lead to lower levels of estimated sensitivity of FBH data in our validation study (figure A3.1). In the most affected scenario (scenario 8), the median estimate of sensitivity fell below 0.70. It is important to note that the effects of HDSS errors on estimates of the sensitivity of FBH data did not vary between symmetric and asymmetric scenarios. This is so because estimates of the sensitivity of FBH data only depend on the proportion of stillbirths recorded by the HDSS that are re-

classified in each scenario, and this proportion does not vary between symmetric and asymmetric scenarios. Estimates of the sensitivity of FBH data in recording neonatal deaths would decline further if a) the completeness of HDSS records of pregnancies was lower than hypothesized in our scenarios, and/or b) the risk of neonatal death associated with unrecorded pregnancies was more than 50% higher than the risk of neonatal death observed in recorded pregnancies.

Our robustness tests also indicated that estimates of the specificity of FBH data in recording neonatal deaths were less affected by potential errors in HDSS data. Overall, specificity estimates remained >0.99 in all scenarios (not shown). In population sub-groups however, errors in HDSS data have the potential to affect specificity levels to a greater extent. In figure A3.2, we show recalculated specificity levels among women who had experienced a stillbirth in the past 5 years according to the HDSS. After accounting for errors in HDSS data, specificity levels declined below the estimate of 0.866 reported in table 2 of the main text. In the most affected scenarios (5 and 6), the specificity of FBH data is estimated at approximately 0.8. Estimated specificity levels were more robust in asymmetric scenarios, i.e., those in which the likelihood in HDSS recording of stillbirths is higher than the likelihood of errors in HDSS recording of neonatal deaths. This is important because, in estimating the NMR, lower levels of specificity might slightly offset the imperfect sensitivity of FBH data in recording neonatal deaths (see equation 1 in appendix A1).

Figure A3.1: effects of errors in HDSS data on estimates of the sensitivity of FBH data on neonatal mortality.

Notes: the dashed red line represents the sample estimate of sensitivity, under the assumption that the HDSS data are not affected by errors (as presented in table 2).

Finally, we evaluated how our assessment of bias in NMR estimates (figure 5 of the main text) would differ in each HDSS data scenario described above. In doing so, we re-evaluated the results of our mathematical model (described in appendix A1) after inserting the median re-calculated value of sensitivity and specificity obtained in each HDSS data scenario. These results are presented in figures A3.3 (symmetric scenarios) and A3.4 (asymmetric scenarios). They indicate that our qualitative results are largely robust to assumed levels of errors in HDSS data, particularly in symmetric scenarios. In these scenarios, only scenario 8 includes a small cluster of data points from the England and Wales series for which a FBH survey would yield an *under*-estimate of the NMR, contrary to our main findings in figure 5. In asymmetric

scenarios, a larger proportion of scenarios include such clusters. These clusters are characterized by relatively low ratios of stillbirths to neonatal deaths, and of post-neonatal deaths to neonatal deaths. In such contexts, false negative reports of neonatal deaths (due to low sensitivity) are not offset by false positive reports of such deaths among women with stillbirths and/or post-neonatal deaths in the past 5 years.

Figure A3.2: effects of errors in HDSS data on estimates of the specificity of FBH data on neonatal mortality.

Notes: These calculations are limited to respondents who had experienced a stillbirth in the past 5 years prior to the survey according to the HDSS. The dashed red line represents the sample estimate of sensitivity, under the assumption that the HDSS data are not affected by errors (as presented in table 2).

Symm. = symmetric scenario; Asymm = Asymmetric scenario.

Symmetric Scenarios

Figure A3.3: effects of errors in HDSS data on estimates of the sensitivity of FBH data on neonatal mortality; symmetric scenarios

Notes: in these figures, the dark blue and light blue areas represent the results of our main analyses, in which we ignore potential errors in HDSS data. The dark blue area thus represents settings (defined by the relative prevalence of stillbirths, NN deaths and PNN deaths) in which a FBH survey conducted with reporting parameters (i.e., sensitivity, specificity) observed in Bissau would under-estimate the NMR; whereas the light blue area represents settings in which a FBH survey would over-estimate the NMR. At the boundary between those two areas, the FBH survey yields an unbiased estimate of the NMR. The scatter plot represents the data from the England and Wales series described in appendix A2. The black contour added to each plot represents the new location of the boundary between under- and over-estimates of the NMR that was obtained in each scenario of HDSS data accuracy. For example, in scenario 8, a cluster of points describing data from England and Wales are located below that added black contour. This indicates that, in that HDSS data scenario, a FBH survey conducted in those settings would yield an under-estimate of the NMR.

Asymmetric Scenarios

Figure A3.4: effects of errors in HDSS data on estimates of the sensitivity of FBH data on neonatal mortality; asymmetric scenarios

Notes: in these figures, the dark blue and light blue areas represent the results of our main analyses, in which we ignore potential errors in HDSS data. The dark blue area thus represents settings (defined by the relative prevalence of stillbirths, NN deaths and PNN deaths) in which a FBH survey conducted with reporting parameters (i.e., sensitivity, specificity) observed in Bissau would under-estimate the NMR; whereas the light blue area represents settings in which a FBH survey would over-estimate the NMR. At the boundary between those two areas, the FBH survey yields an unbiased estimate of the NMR. The scatter plot represents the data from the England and Wales series described in appendix A2. The black contour added to each plot represents the new location of the boundary between under- and over-estimates of the NMR that was obtained in each scenario of HDSS data accuracy. For example, in scenario 8, a cluster of points describing data from England and Wales are located below that added black contour. This indicates that, in that HDSS data scenario, a FBH survey conducted in those settings would yield an under-estimate of the NMR.

APPENDIX A4: VARIATION IN THE SENSITIVITY OF FBH DATA BY TIME SINCE THE DEATH.

Figure A4.1: Variation in the sensitivity of FBH data, by timing of the death.

Notes: deaths were binned in 6-month intervals, and we recalculated the sensitivity of FBH data for each of these bins. The solid red lines in each graph represent the overall estimate of sensitivity reported in table 2 of the main text. Within each bin, the placement of the marker is determined by the mean time between the survey and the death among deaths having occurred in that interval. For post-neonatal deaths, there were too few deaths at durations > 54 months before the survey (<5).

APPENDIX A5: DISTRIBUTIONS OF AGE AT DEATH

Figure A5.1:
Distribution of
ages at death
during the
neonatal period
according to
different data
sources

Notes: these distributions are limited to deaths reported to have occurred within 5 years of the survey.

Figure A5.2:
Distribution of
ages at death
during the post-
neonatal period
according to
different data
sources. *Notes:*
these distributions
are limited to
deaths reported
to have occurred
within 5 years of
the survey. There
were no deaths
that occurred
during the first
month but after
the 28th day of
life.

APPENDIX A6: DIFFERENCES IN SOURCES OF ERRORS BETWEEN EARLY AND LATE NEONATAL DEATHS AMONG FALSE NEGATIVES

Notes: the sample size for these analyses was n = 36 false negative reports of neonatal deaths.