Use of Health and Demographic Surveillance Site data for secondary analyses: guidance for researchers using examples from existing analyses

Estelle McLean[1,2#], Emma Slaymaker[2], Rebecca Sear[2]

1. *Malawi Epidemiology and Intervention Research Unit, Lilongwe, Malawi.*
2. *London School of Hygiene and Tropical Medicine, Faculty of Epidemiology and Population Health, London, United Kingdom*

Introduction

Health and Demographic Surveillance sites (HDSS) are open cohorts which operate in specific geographical areas. Everyone living within the defined area is eligible to take part. Participants may enter the HDSS system through the full census carried out at the start of operations, birth or moving into the area, and may leave the system through death, moving out of the area, or if the HDSS stops operations. HDSSs are designed to monitor trends in mortality, fertility and migration: they have been described as a short to medium-term solution to filling the gap of vital registration systems (Ye et al. 2012). Readers are directed to other recent articles which give a useful and detailed descriptions of HDSSs and important ethical issues (Ghafur et al. 2020; Hinga et al. 2021; Herbst et al. 2021).

HDSSs capture data for many years and usually also regularly gather other health and socio-demographic data from the participants; meaning that they are rich sources of data for secondary longitudinal analyses. They also collect data continuously, meaning that data are available for all years that the HDSS is in operation. Participants of all ages are included in HDSSs and they are grouped according to their household, which often have GPS coordinates recorded. Often linkages between parents and children, and between spouses are available. This means that powerful analyses taking environmental, household and family contexts into account can be carried out. The nature of HDSSs, however, means that firstly, data are only captured on participants when they are living in the area, and secondly that some data are collected repeatedly. These factors may introduce bias to analyses, and mean that HDSS data can be complex to understand and manage. It is important that the complexities of HDSS data and analyses are reported appropriately, assuming that readers may not be aware of the platform-specific issues. It was noted in a meta-analysis on HIV incidence, that sometimes the potential impact of migration was not totally clear in some HDSS analyses, making interpretation challenging (Birdthistle et al. 2019).

The objectives of this paper are to demonstrate the utility of using HDSS data for secondary analyses, and to provide guidance on the conduct and reporting of these analyses using published examples from the literature.

Methods

Papers were identified through a systematic search using Google Scholar. The criteria for inclusion were firstly that the analysis must be using the HDSS data longitudinally and secondly that it must use some level of data manipulation which leverages linkages within each individual's HDSS repeated surveys, and/or linkages between individuals within households or family groups. The in-depth review of each selected paper focused on the

methods and results sections, plus the limitations sections of the discussion to assess them according to 5 key aspects of using HDSS data: data manipulation methods and resulting dataset structure; statistical methods; how repeated measures were used or accounted for; how in/out-migration was dealt with; and how other missing data were dealt with.

Results

A total of 46 papers from 14 HDSSs were reviewed. The subjects under analysis included mortality, fertility, migration, household composition, adolescent transitions, HIV diagnosis and care, childhood growth plus one paper analysing participation in surveys. Three examples to demonstrate the breadth of analyses are described below:

1. Data from uMkhanyakude HDSS (South Africa) were used to create a time-to-event dataset from the date the participant was first known to be HIV negative ending at the estimated date of sero-conversion, censor dates were added from the HDSS residency dataset (dates of migration/death) and times of potential exposure identified through an annual survey where sexual partners were reported (Harling et al. 2014).

2. Data from Agincourt HDSS (South Africa) were used to identified women with a live birth and linked clinic data to identify date of HIV sero-conversion, date of ART initiation and any time they were classed as disengaged from care to construct a dataset with the HIV/ART status for each day of the 1000-day period from estimated conception of a live-born child to estimated end of breast-feeding, and used sequence analysis and cluster analysis to assess different patterns of engagement with care (Etoori et al. 2021).

3. Data from 29 HDSSs were used, linking data from children aged under 5 with their mothers and siblings' dates of birth, migration and death which were used to create time-varying variables on a time to event dataset with the outcome of child mortality. (Bocquier et al. 2021).

Data manipulation techniques and dataset structures
10 distinctive data manipulation techniques were identified, either linking data from multiple time-points from one individual or linking data between individuals from 1 or many time points. The resulting datasets were commonly 'time-to-event' with an end point outcome under analysis. Others reduced the multiple records to one record per person and most of the remainder ended up with other multi-record datasets. Most datasets used individuals as the unit of analysis, but there were a few that used the household.

Statistical methods
The majority of statistical techniques used were fairly standard for epidemiology and demography data, survival (time-to-event) analyses, sometimes allowing for competing events, and logistic or linear regression which may be multinomial. A few used more sophisticated techniques, including sequence analysis, and multi-state transition models.

Repeated measures
The studies in the review dealt with data inconsistencies in different ways, either creating a rule (i.e. first report) or attempting to clean the data. A few did not mention any issue with

inconsistencies though it seems likely that there were some. In the studies which used repeatedly reported dates to identify transitions, the earliest date tended to be used and often the assumption was made that the person could only experience the event or transition once. The majority of analyses accounted for repeated records or clustering in the statistical methods used, either by introducing fixed or random effects, or using a method that intrinsically accounts for multiple records, i.e. survival analysis or multi-level modelling.

Migrations

Many of the papers in the review used 'typical' HDSS time-to-event analyses where participants contribute time 'at risk' when they are present, i.e. no-one is excluded for not being present the whole time. Some however restricted their analysis to children who were born in the area (usually because exposure data were not available for those who in-migrated) but did not exclude those who out-migrated before experiencing the event. Other studies used migration as an outcome or exposure so bias was reduced. Some of the studies using longitudinally linked data excluded all participants who did not have all data points. Other studies tried to mitigate the effect of migration by a. trying to keep the inclusion criteria as wide as possible, i.e. only requiring data from 2 time-points, or b. treating each transition/event as a separate analysis so as many participants can be included in at least one, or c. include time spent in the HDSS as a control measure, d. triangulate the findings with a slightly different analytical approach. Surprisingly, many of the reviewed papers made no mention of migration or attrition in the discussion, even when the analysis design makes it seem like migration might have caused some bias. Others did discuss the issues and a few attempted to examine the effect of attrition, by comparing attributes of those included and not, or assessing the effect of the exposure on the outcome of out-migration as well as the outcome under investigation.

Missing data

All studies, regardless of whether the data come from an HDSS or not, may be prone to issues from missing data and the majority of studies either dealt with missing data in standard ways, i.e. excluding those missing data from the model, retaining them under a category of unknown/missing, with a few attempting some multiple imputation. There were a few instances of data from previous or later time points being used to impute missing data, though no discussion over the pros and cons of this approach: while the power of the study is increased, bias may be introduced as older people, or those in the area for longer may be more likely to have earlier or later data points to use. In the analyses which utilised linkage between participants (i.e. mother-child, or within households) to generate exposure or outcome variables, there was never any discussion of what was done if there were any missing data when creating variables.

Conclusions and recommendations

The reviewed HDSS data analyses demonstrated the flexibility of HDSS data. Some quite complex statistical techniques were used, however the majority usage of more 'standard' techniques show that advanced statistical skills are not a requirement for conducting valuable analyses with HDSS data. While there were some interesting and novel ways used to approach the issues of repeated data, migration and missing data it was most striking how often they were not discussed. The nature of HDSS data means that migration is

something that will always need to be considered in almost all analyses beyond simple cross-sectional analyses. Careful use of the data may reduce the impact on results and conclusions, however it will never be possible to mitigate the effect entirely.

Users of HDSS data should be aware of the issues of repeated data capture, of in- or out-migration, and of missing data, and consider the most appropriate way of dealing with it. This may require close consultation with HDSS data producers. Researchers writing up HDSS analyses for publication should consider whether their approach to dealing with repeated data, migrations and missing data have been appropriately detailed in the methods section. Equally, researchers should ensure to include adequate discussion of these factors in the limitations section, to enable readers who are not experts in HDSS data to fully understand any potential issues in interpretation of findings.

References

Birdthistle, I. et al. 2019. Recent levels and trends in HIV incidence rates among adolescent girls and young women in ten high-prevalence African countries: a systematic review and meta-analysis. *The Lancet Global Health* 7(11): p.e1521–e1540. Available at: www.thelancet.com/lancetgh [Accessed July 20, 2023].

Bocquier, P., Ginsburg, C., Menashe-Oren, A., Compaoré, Y., & Collinson, M. 2021. The crucial role of mothers and siblings in child survival: Evidence from 29 health and demographic surveillance systems in sub-saharan africa. *Demography* 58(5): p.1687–1713. Available at: http://read.dukeupress.edu/demography/article-pdf/58/5/1687/1167556/1687bocquier.pdf [Accessed December 1, 2022].

Etoori, D. et al. 2021. Patterns of engagement in HIV care during pregnancy and breastfeeding: findings from a cohort study in North-Eastern South Africa. *BMC Public Health* 21(1): p.1–12. Available at: https://doi.org/10.1186/s12889-021-11742-4 [Accessed December 1, 2022].

Ghafur, T., Islam, M.M., Alam, N., & Hasan, M.S. 2020. Health and Demographic Surveillance System Sites: Reflections on Global Health Research Ethics. *Journal of Population and Social Studies* 28: p.265–275. Available at: https://so03.tci-thaijo.org/index.php/jpss/article/view/207700.

Harling, G. et al. 2014. Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in Rural KwaZulu-Natal, South Africa. *Journal of Acquired Immune Deficiency Syndromes* 66(4): p.443–451. Available at: /pmc/articles/PMC4097949/ [Accessed December 16, 2022].

Herbst, K. et al. 2021. Health and demographic surveillance systems in low- and middle-income countries: history, state of the art and future prospects. *Global Health Action* 14(sup1). Available at: /pmc/articles/PMC8986235/ [Accessed January 4, 2023].

Hinga, A.N., Molyneux, S., & Marsh, V. 2021. Towards an appropriate ethics framework for Health and Demographic Surveillance Systems (HDSS): Learning from issues faced in diverse HDSS in sub-Saharan Africa. *BMJ Global Health* 6(1): p.4008. Available at: http://dx.doi.org/10.1136/bmjgh-2020-004008 [Accessed January 4, 2023].

Ye, Y., Wamukoya, M., Ezeh, A., Emina, J.B.O., & Sankoh, O. 2012. Health and demographic surveillance systems: A step towards full civil registration and vital statistics system in sub-Sahara Africa? *BMC Public Health* 12(1): p.741. Available at: https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-12-741 [Accessed January 4, 2023].