

Versatility, value and limitations of using Health and Demographic Surveillance System data for secondary analyses: guidance for researchers, using examples from existing analyses

Introduction

Health and Demographic Surveillance Systems (HDSS) are open cohorts which operate in specific geographical areas, which may be one contiguous area, or multiple separate ones. Everyone living within the defined area is eligible to take part. Participants may enter the HDSS through being identified in the full census carried out at the start of operations, moving into the area, or at birth, if they are born into the HDSS area. They may leave the system through death, by moving out of the area, or if the HDSS stops operations. Historically, HDSSs were set up to monitor the health of communities, by monitoring 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) and exist in multiple low and middle income countries. Readers are directed to other recent articles which give a useful and detailed descriptions of the past, present and future of HDSSs (Herbst et al. 2021) and important ethical issues associated with them (Ghafur et al. 2020; Hinga et al. 2021).

As well as collecting data on demographic events, HDSSs tend to regularly gather other health and socio-demographic data from the participants. They are also used as platforms for specific studies which make use of the existing infrastructure to firstly select and find the specific type of participant they need, and then to find them again, if the study requires multiple interactions (Agarwal et al. 2017). 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 also available. This means that powerful secondary data analyses taking environmental, household and family contexts into account can be carried out.

HDSSs are different from other cohort studies which typically recruit participants over a certain period (such as children born during a particular week or year), and then follow them up for a set length of time, often only collecting certain information at baseline. In comparison, HDSS recruitment is continuous, people may leave and subsequently return to the cohort, follow-up times tend to be longer, and information is collected at regular intervals making them rich sources of data regarding time-changing characteristics. While the constantly changing HDSS population ensures that the study continues to represent the population in the area, the data are complex to analyse because of right- and left-censoring (people who drop out of the HDSS before death, or enter the HDSS after birth, respectively)

and residency gaps (when participants migrate out but later return to the area): most analyses at the least need to take delayed entry (left-censoring) into account. Additionally, participants who leave the area and therefore provide less data may be different to those who stay, so excluding them may introduce bias, depending on the specific research question. The long-running nature of the HDSSs may also result in complications as survey instruments and practices evolve over time. This is, however, also a strength of HDSS systems, as the existing infrastructure can quickly be adapted to include current health priorities. For example, existing and new HDSS processes and data were leveraged to inform HIV policy (Herbst, Law, et al. 2015) and several HDSSs used existing and new systems to respond to the Covid-19 pandemic (Siedner et al. 2021; Romero Prieto et al. 2022). Repeated data capture allows for detailed analysis of time-changing variables, but can lead to complex data management needs if inconsistencies are recorded. HDSS populations can be large and represent the whole local population and as such can be highly powered for some analyses; however if certain sub-groups are less common in the area, power may be reduced for studying these groups.

Like most health and population research, HDSS data are generated with one of the goals being to influence local, national and global policy (Williams et al. 2010; Herbst, Law, et al. 2015). Individual and pooled analyses of HDSS data on various subjects can be found in the literature, for example mortality (Wasko et al. 2022); HIV (Roberts et al. 2022); religiosity (Lynch et al. 2022); education and poverty (Werner et al. 2022); nutrition (Mank et al. 2021) and vaccination (King et al. 2020). HDSS analyses are also commonly used in systematic reviews and meta-analyses, for example, on sexual and reproductive health (Arthur et al. 2013); child mortality in relation to mother's survival (Nguyen et al. 2019); HIV incidence in adolescents and young women in Africa (Birdthistle et al. 2019); child stunting and school performance (Gansaonré et al. 2022); HIV incidence in pregnancy and breastfeeding (Graybill et al. 2020); and migration and HIV acquisition (Dzomba et al. 2019). However, HDSS data are used in published analyses far less than data from Demographic and Health Surveys, which are regular nationally-representative surveys carried out in low and middle income countries.

Objectives and structure of this paper

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 examples from the literature. It is anticipated that this paper will be useful to producers of HDSS data, and new and existing users of the data. Firstly, the positive and negative

aspects of using HDSS data, summarised above, are reviewed fully using Demographic and Health Survey (DHS) data as a comparison, focusing on the areas of representativeness and generalisability, longitudinal analyses, linkages between participants, pooling data across sites and countries, and accessing data. Secondly, examples of HDSS analyses which leverage the key aspects of HDSSs, their longitudinal nature and the linkages between individuals, are identified through a search of the literature. These papers are used as the basis to discuss key aspects related to the conduct and reporting of HDSS analyses, focusing on data manipulation and dataset structures, statistical methods, accounting for repeated measures, dealing with in- and out-migration and issues with missing data. Finally, key recommendations to researchers using HDSS data are provided.

Pros and cons of HDSS data with comparison to DHS data

Demographic and Health Surveys (DHS, <https://www.dhsprogram.com/>) are a useful comparison for HDSS data as both provide data on similar core topics (HDSSs may be used as platforms for studies of various topics, but this section focuses on the core demographic surveillance data), and the two data sources have been used for comparative analyses before (Byass et al. 2007). DHS is a program of nationally-representative, cross-sectional household surveys carried out in low and middle income countries. It started in 1984, developing from World Fertility Surveys and contraceptive prevalence surveys in the 1970s and 1980s (Cleland 2010). The core surveys collect information on a range of population and health topics, and optional modules are available for countries to add if required (Fabic et al. 2012). DHSs are meant to be carried out regularly, usually every 3-6 years, though individual countries decide upon the timing and frequency of the surveys. They are cross-sectional surveys, with a new sample interviewed in each round. Sampling is stratified by geographic area and, depending on the size and needs of the country, 5000-30,000 households are included. For each household, a 'roster' of household members is recorded, along with a household survey from a key informant, and then individual interviews are carried out with specific individuals, usually all women aged 15-49, and often a sub-sample of men aged 15-59 years (Corsi et al. 2012). Pros and cons of using HDSS data for secondary analysis compared to DHS are described below and summarised in table 1, more detail is given on HDSS as that is the focus of this paper.

Representativeness and generalisability

The DHS has advantages over HDSS in terms of national representativeness (Corsi et al. 2012). As HDSSs only exist in small areas, they are not generalisable to the wider population, however they include the full population in a specific area and tend to involve

large numbers of participants (though this varies considerably between sites, from under 50,000 participants to over 500,000), which are advantages for certain analyses (Bocquier, Sankoh, et al. 2017). It has also been argued that the lack of national representativeness can be mitigated to some extent through triangulation with other sources, and that the richness and quality of the data and statistical power are more important when considering their usefulness (Bocquier, Sankoh, et al. 2017).

Longitudinal analyses

Repeated DHS surveys allows for assessment of population-level trends over time, however there is no information in years between DHS surveys, and, as new participants are interviewed for each survey, no possibility of conducting longitudinal analyses of individuals, or following them up for further surveys (Corsi et al. 2012). HDSS data can also be used to examine (local) population trends over time, with the advantage that data from all years of operation are available. For some HDSSs the time spans available are very long: the oldest HDSSs were set up in the 1940s and 50s (Herbst et al. 2021; Ye et al. 2012), however, the populations under observation may become different due to being observed because of the HDSS, and participation fatigue may affect data (Herbst et al. 2021).

Additionally, HDSS participants are linked across all time points allowing individual longitudinal analyses, which are not possible with DHS data. However, there are several important issues that must be considered when contemplating such analyses. Firstly, HDSSs are open cohorts resulting in unbalanced data: some participants may have fewer data points than others if they migrate in or out of the area, and migration may be related to outcomes (Bocquier 2016). Secondly, some participants will have the same items of data collected repeatedly which may lead to more complete and accurate data but may also lead to bias if people living in the area for longer have more reliable data than newcomers, and can introduce inconsistencies which require more complex data manipulation skills. These two issues will be discussed in greater detail later in this paper. Thirdly, while it is tempting with the richness of HDSS data to use it to conduct detailed analysis or draw in-depth conclusions about specific individuals or small groups of individuals, it must be remembered that the system is designed to draw conclusions about the whole population, rather than for individuals. Data are collected over a long period, by multiple (albeit trained) interviewers, often from proxies. Data may be reported differently from one round to another due to lack of knowledge from a proxy, recall errors, different interpretation of events, misunderstandings between the interviewer and interviewee, input errors by the interviewer, or through data entry errors (not all of these are unique to HDSS data). Once the data are collected, data manipulations may take the data used in an analysis further from the 'truth' for any specific

individual. In addition, even if the data is 'correct' participants' life experience may be much more nuanced than could be concluded from HDSS data (Reynolds 2015).

Linkages between participants

The HDSSs are at an advantage in terms of linking between individuals: linkages are usually possible between household members, parents, children and spouses often regardless of co-residence (for example the Karonga HDSS in Malawi (Crampin et al. 2012; McLean et al. 2023)). Within the DHS, women may be linked with their spouses, if they are included, and will report on all their children ever born, however, in a DHS, no data are collected about children living in a household without their mother. Additionally, often only women of child-bearing age are included in DHS (Corsi et al. 2012). This latter point is probably the most important with regards to interpreting DHS data, as infant and child mortality rates can only be calculated based on data from living mothers in that age group, so there is a potential for bias if maternal survival is linked with infant mortality or participation rates. Beyond linkages within the HDSS, in many HDSSs it may be possible to link members to other data sources, such as clinic records (for example Agincourt HDSS in South Africa (Kahn et al. 2012)); however, as this requires participant consent, funding, technical support to conduct the linkage, ethical approval and a supportive legislative framework, these links are often not maintained for a long period.

Pooling data across sites and countries

A key advantage of DHS data over HDSS is the consistency and comparability of methods and data across countries (Corsi et al. 2012). DHS was set up with harmonisation and pooled analyses in mind, and the continuous funding and support allows researchers across the world to access and analyse the data. HDSSs do not have such an over-arching structure of support and funding and the core datasets generated by each site are often quite different, however they can still be used for pooled or comparative analyses. There have been attempts to pool and harmonise the data (i.e. the INDEPTH network (<https://data.indepth-network.org/>) [see the data access section below], the Alpha network (Reniers et al. 2016), SAPRIN (<https://sapr.in.mrc.ac.za/>), and currently the Implementation Network for Sharing Population Information from Research Entities (INSPIRE) (<https://aphrc.org/project/inspire-implementation-network-for-sharing-population-information-from-research-entities/>) is investigating the use of common data models: they are transforming data from East Africa HDSSs into the Observational Medical Outcomes Partnership (OMOP) format to facilitate data discovery, sharing and reuse. This process is complex, and requires new skills from both the data producers and users, and new ways to

convey nuances in the data. The sustainability of this kind of initiative will require on-going resources and training.

While HDSSs cannot compete with the level of geographic coverage of the DHS, they exist in many low and middle income countries across the globe and, considering all HDSS areas combined, coverage of different types of ecological and environmental areas, and socio-economic strata has been found to be relatively high (Tatem et al. 2006; Jia et al. 2015; Edson Utazi et al. 2018).

Data access

One of the great strengths of the DHS data is that it is all freely available via their website, either as individual data, or summary statistics and graphs using web-based analytical tools (<https://dhsprogram.com/>). While it is possible for external researchers to access HDSS data (see below) there are more barriers. Production of HDSS data (involving study design, programming, logistics, field-work, data entry, cleaning and storage) takes a huge amount of resources, and can represent quite a burden for the study population who give a lot of time to provide the data. Along with other public health datasets there have been calls to make HDSS data more accessible to external scientists (Chandramohan et al. 2008). Researchers often struggle to share their data as they need return on their investment of the time spent collecting the data (Anane-Sarpong et al. 2018), and the additional time, resources and skill/expertise needed to generate suitably documented datasets do not tend to be recognised in academic career progression pathways (Chawinga et al. 2019): there have been calls to reduce some of these barriers by increasing recognition of the vital role of data professionals (Pisani et al. 2010). In addition to standard barriers to data sharing, HDSS sites are exclusively operated in low income settings, where both analytical and data management capacity are lower and there is an understandable reluctance to professionals from low income settings to put considerable time and expertise into collecting and curating datasets, with little recognition and career advancement, only for the data to be analysed by researchers from high income settings (Chandramohan et al. 2008; Hinga et al. 2021). With multiple records per participant it is also difficult to share data in an open access form, as suitably anonymised data may become useless for analysis. Work has been done to develop methods to anonymise HDSS data which reduces the chances of individuals being identifiable yet maintaining the utility of the dataset (Templ et al. 2022), and to 'geomask' datasets with GPS data so that privacy is maintained but useful analyses can still be carried out (Hunter et al. 2021).

Despite all the issues detailed above, some HDSS data is available openly. The main platform for HDSS data is the INDEPTH network (<https://www.indepth-ishare.org/index.php/home>). This is a collaboration of multiple HDSSs across the globe providing analytical support and workshops, as well as a platform to share basic harmonised HDSS datasets (Herbst, Juvekar, et al. 2015). These datasets are available to external researchers, and while they do not contain all the possible data from each HDSS, have been used for multiple pooled analyses. Two training manuals are available; firstly describing how to obtain data from the INDEPTH system, how to format it for event history analysis/survival analyses and how to carry these out (Bocquier, Ginsburg, et al. 2017) and secondly specifically on migration and mortality analyses (Bocquier et al. 2019). Additionally the Mekong HDSS which ran from 2001 to 2006, released their full datasets open access in 2016 (<https://www.icpsr.umich.edu/web/DSDR/studies/36601/publications>) (Heuveline et al. 2016), and the Agincourt HDSS in South Africa also makes some of their data openly available: they regularly update their '1 in 10' dataset, which is a 10% sample of the full dataset (<https://www.agincourt.co.za/data>). Anyone is free to download this sample data, to plan analyses and then a full request with individual approval is needed to obtain the full set. Accessing other HDSS data may be possible through contacting the managing organisations individually.

Conclusions of the comparison between HDSS and DHS data

This review has demonstrated how HDSS data can complement DHS data for producing useful analyses on demography and health to fill evidence gaps. While HDSS data may be the better choice for some analyses, it is more complex to analyse to avoid accidentally introducing biases or drawing the wrong conclusions: and a greater level of statistical and data manipulation expertise, and collaboration with the data producers, is required for HDSS analyses compared to DHS ones.

Table 1. Comparison of advantages and disadvantages of DHS and HDSS

Demography and Health Surveys (DHS)	Health and Demographic Surveillance Sites (HDSS)
+ Nationally representative	-/+ Only small area and small portion of national population included, however the entire local population is included
+ Data available on-line	+/- Some data are available, i.e. through the INDEPTH network; most HDSS data is only available through requests to specific sites

Demography and Health Surveys (DHS)	Health and Demographic Surveillance Sites (HDSS)
+ Harmonised data available across countries. Common data standard exists	+/- Each site's data are different and requires additional effort to harmonise for pooled/comparative analyses. Only a limited common data standard/model
+ Comparative data capture techniques across countries	+/- Procedures and definitions may vary across sites, however they are relatively similar
+ High response rates	+ Very high response rates
+ Repeated cross-sectional surveys allow analysis of change at population level	+ Longitudinal data allows analysis of change over time at population level
+ Mother data can be linked to child data and cohabiting couples can be linked if both name the other.	++ Data on individuals can be linked to other household members, usually parents and often other relatives
- Surveys cannot be linked to analyse change over time at individual level	+ Longitudinal data allows for life-course analysis of individuals
- Surveys in different countries happen at different times with different frequencies	+ Data are collected sufficiently often to permit analysis at any time point and as continuous time data as opposed to discrete time or panel data analysis.
- Only certain individuals included, i.e. women aged 15-49	+ All ages included
- It is not possible to identify and follow-up individuals	+ Can act as a sampling frame for follow-up studies
+ Data mostly collected from individuals directly	- Proxies often used for demographic data; health data usual collected directly from individuals
+ Little/no data cleaning required by analysts	- Data cleaning required, especially to deal with inconsistencies with longitudinal data, and expertise with the HDSS required for making appropriate data cleaning decisions.
+ Backwards compatible questionnaires and data structures	- Backwards compatibility constrained by evolution of research priorities and limited funding

Review of existing secondary analyses of HDSS data

The aim of this section was to investigate how HDSS data is used in longitudinal analyses, to demonstrate the use of such data and highlight alternative ways of handling and analysing HDSS data. To identify a broad range of HDSS analyses to review, a search of the literature was conducted using Google Scholar, as this tool tends to find literature from many different disciplines, and also covers the 'grey literature': conference papers, working papers and student theses. The search was not intended to be a comprehensive review of all HDSS analyses, so was restricted to relatively recently papers published from 2012 up to the end of 2022. The following search terms were used, which were based on the types of analyses known to utilise the longitudinal and linkage aspects of HDSSs:

("Demographic Surveillance System" OR "Demographic Surveillance Site") AND ("event history analysis" OR "joint modelling" OR "latent transition analysis" OR "lifecourse analysis" OR "life course analysis" OR "longitudinal analysis" OR "multi-state modelling" OR "sequence analysis" OR "seroconversion" OR "sero-conversion" OR "survival analysis" OR "trajectory analysis" OR "typology")

A total of 1745 results were identified. The titles, abstracts or full text were scanned to identify duplicates, papers that were not in English, were BSc. or MSc. projects or were not analytical (i.e. protocols, review papers etc.) and analyses that were not involving humans, were laboratory or qualitative studies, or were quantitative studies but not carried out in HDSSs: using these criteria, 1250 results were excluded. 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. Thus, papers which just analysed a cross-sectional survey, or reported a standard mortality rate analysis, were not included because those required only commonly used methods for data manipulation and analysis. Analyses of studies that solely used the HDSS as a platform, i.e. collected data from HDSS members but did not link to other HDSS data or other nested studies, were also excluded, unless there were at least 3 linked data points. Using these criteria a further 454 results were excluded.

There were several examples of data reduction techniques and complex data approaches being used which were not included as the data were not used longitudinally, either just using one round of HDSS data or using the latest data available for each participant: for example in Agincourt HDSS, Bayesian belief network models were used to examine

household food security (Eyre et al. 2021) and ‘mixture of factor analyzers for mixed data’ models were used to assign households to clusters based on socio-economic factors (McParland et al. 2014). In Cuatro Santos HDSS in Nicaragua, cluster analysis was used to assign households to groups based on multi-dimensional poverty (Källestål et al. 2020). There were also examples of data linkage within households to generate household composition variables, however the overall analysis was not longitudinal, so they were not included (as any one-off household survey could have done the same thing); for example an analysis from the Mekong HDSS in Cambodia which looked at living arrangements and associated them with school attendance (Heuveline et al. 2017), and a similar analysis from Agincourt in South Africa (Madhavan, Myroniuk, et al. 2017). There were also some examples of data science techniques being used on verbal autopsy data regarding improving cause of death algorithms (Murtaza et al. 2018) and several examples where spatial clustering was assessed, sometimes longitudinally (Becher et al. 2016). These were not included as they do not really harness the individual HDSS linkages. Several HDSSs have carried out multiple rounds of HIV testing enabling estimates of HIV incidence if participants have observable negative and then a positive test, these papers were only included if they harnessed another aspect of HDSS linkage, i.e. linkage between couples or household members, or linkage to another HDSS outcome such as migration or mortality.

A total of 41 papers were included from the main review and a further 5 were added following additional searches undertaken to identify papers where the term ‘demographic surveillance site’ was not used. All screening and selection of papers was carried out solely by the first author EM. A total of 46 papers from 14 HDSSs were reviewed (NB. 7 use data from multiple HDSSs): uMkhanyakude (South Africa) has the largest number with 12, Agincourt (South Africa) had 7, Nairobi (Kenya) had 6, Karonga (Malawi) had 3 and Rufiji (Tanzania) had 2; Cuatro Santos (Nicaragua), Farafenni (Gambia), Kanchanaburi (Thailand), Kyamulibwa (Uganda), Magu (Tanzania), Manicaland (Zimbabwe), Matlab (Bangladesh), Niakhar (Senegal), Rakai (Uganda) all contributed 1 paper. The subjects under analysis included mortality, fertility, migration, household composition, adolescent to adulthood transitions, HIV diagnosis and care, childhood growth plus one paper analysing participation in surveys. It is acknowledged that the search was not exhaustive, however the number of papers was felt to be sufficient, and the coverage of HDSSs, topics and techniques broad enough to conduct the review.

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. Categories within each section were created a priori, and/or during the review process. The findings within each section are used to highlight specific issues to be aware of when conducting secondary analyses of HDSS data and as such, in general, the absolute numbers within each category are not reported. Examples from the reviewed papers are referenced through, and all papers are listed in table 2, along with the categories they were assigned to in each section.

Data manipulation techniques and dataset structures

Raw HDSS datasets can be complex and almost all analyses will require some level of data manipulation. Additionally, HDSS datasets can be very flexible, allowing for many types of dataset structure and resulting analysis. In this section the breadth of data manipulation techniques and dataset structures used in the reviewed papers is described, along with advantages and disadvantages of each method.

The identified techniques were classified first into those which took advantage of linking data from multiple time-points from one individual, this was used in 5 different ways:

A. to create summary exposure or outcome measures which were not linked to a date/time

- e.g. child anthropometry data from multiple time points from Kyamulibwa HDSS in Uganda were used to create a summary variable to indicate child's experience with stunting (i.e. always stunted, never stunted, improved, worsened), this variable was used as an exposure for a later analysis (Asiki et al. 2019).
- This technique has benefits of being relatively simple to carry out (depending on the number of inputs) and providing a usable outcome measure; however, it may be too simplistic if timing of measures is important.

B. to create an event with a date as an end-point

- e.g. multiple rounds of HIV test and clinic data from uMkhanyakude HDSS in South Africa were linked to generate dates of specific transitions along the HIV care pathway which were used in multiple time-to-event analyses (Haber et al. 2017).
- This technique allows synthesis of multiple data points into a single (or multiple in cited example) outcome suitable for longitudinal data analysis; however, care must be taken to consider whether those for whom a transition date could not be derived actually did not experience the transition or do not have enough data.

C. a sequence of events/states

- e.g. in Agincourt HDSS in South Africa, HIV test, clinic and treatment data on women with a live birth were combined to generate a sequence with one variable per day from estimated date of conception to estimated date of cessation of breast-feeding (Etoori et al. 2021).
- This technique was rarely used, it has value in that it takes full advantage of the longitudinal data source; however, as in point B, care must be taken to consider whether excluding participants without enough data to create a complete sequence will introduce bias (see section on migration below).

D. to create time-varying exposure variables

- e.g. data from multiple sexual behaviour surveys from uMkhanyakude HDSS in South Africa were used to create time-varying exposures relating to age of sexual partners in a time-to-event analysis with HIV sero-conversion as the outcome (Harling et al. 2014).
- This technique improves time-to-event analyses by taking into account that people's experiences change over time; however, data manipulation decisions need to be made to ensure that categories are assigned to person-time appropriately, i.e. when deciding how long a certain report should stand, and deciding what to do about overlapping reports.

E. Lagged data used as exposure

- e.g. in uMkhanyakude HDSS in South Africa, data collected at age 11-13 was used as the exposure for outcomes in teen mothers at age 20 (Ardington et al. 2015).
- This technique, which builds on point D, recognises that events and experiences may not have an immediate effect on people; however, care must be taken to consider how those who do not have enough data to generate the lagged variable (i.e. due to age or time in the HDSS) may differ from those who do.

The second main data manipulation method involved linking data between individuals from 1 or many time points, there were also 5 separate techniques:

F. Linkage for matched analysis

- e.g. data from maternal sisters (linked through mother id) were matched for an analysis of teen pregnancy in uMkhanyakude HDSS in South Africa (Ardington et al. 2015).
- Matched analysis can be very useful and powerful; however, in the HDSS context, care must be taken to consider how representative the matched sample are, i.e. people can only be matched to those within the area.

G. Data from linked household members used to create summary exposure or outcome variables

- e.g. in Nairobi HDSS in Kenya, data from household members' education level were used to create a household summary variable used as an exposure variable (Mutisya et al. 2016).
- This technique allows for the wider context to be taken into account, however it must be considered what to do if data from any member are missing, and whether to drop or impute the data. Another consideration is the definition of household membership: assigning participants to one household is essential for logistical reasons; however, this may not represent the actual situation. Definitions of household membership are not straightforward (McLean et al. 2021) and may vary across HDSSs (though often require recognising the same household head and 'eating from the same pot'). For example, some HDSSs allow for non-resident members and others don't (Tanser et al. 2008).

H. As G but the linked unit was something other than the household

- e.g. an example from Cuatro Santos HDSS in Nicaragua summarised data from adolescent girls in small geographical areas to generate background adolescent pregnancy rates to use in models (Pérez et al. 2021).
- This technique also allows for inclusion of wider contextual factors, but care must be taken as above to consider the effects of missing data. Additionally, it must be considered whether the unit in question has real meaning, i.e. the HDSS may be divided into areas for logistical reasons but the areas may not provide enough granularity for meaningful analysis.

I. Data from linked household members used to create household event with date

- e.g. individuals' death and migration data from Rakai HDSS in Uganda were used to generate household events of dissolution or migration (Muniina 2016).
- Household level analysis can be a valuable use of HDSS data, but the same considerations as above regarding missing data and household membership definitions in point G should be taken into account.

J. Data from linked person(s) used to create exposure variables for index

- e.g. an analysis using pooled data from multiple HDSSs used linkage between index children and their mothers and siblings to create time-varying exposure variable relating to mother and sibling survival and migration status (Bocquier et al. 2021).
- This technique builds on point D and E in the above section, recognising that events occurring to other people have an effect on index participants; however, as in point F

care must be taken to consider how index participants without matches may affect the analysis.

The resulting datasets were commonly ‘time-to-event’ with an end point outcome under analysis: this is a very common data structure when analysing longitudinal data, including HDSS data. HDSS data in this structure may be different from other longitudinal cohort data in that entry may be delayed for some participants (those who move into the area after the initial census), and there may be gaps in follow-up for others (those who leave the area and return). It is important that only person-time when participants are resident in the area is included in such a dataset (readers should refer to Bocquier, Ginsburg, et al. 2017 for detailed guidance on constructing such a dataset). The majority of the reviewed papers in this category used data arranged as episodes covering the whole period of their residency in the HDSS, but some simplified the dataset by splitting the data into one record per month (e.g. Houle et al. 2021) or year (e.g. Madhavan et al. 2012) that each participant was present. While simplifying the episodic data to monthly or yearly may seem like information would be lost, sometimes this is the most appropriate method, taking into consideration how frequently the data are captured (Allison 1982); additionally it allows for easier further data manipulation. Others reduced the multiple records to one record per person (e.g. Machiyama et al. 2015). This does not necessarily mean loss of information, as this structure can include longitudinal data in a wide format, such as those used for sequence analysis with one column per time point. Other analyses in this category did result in loss of information, though; for example an analysis which reduced transition to parenthood sequences to a simple variable not taking time between events into account (Del Fava et al. 2016). Most of the remainder ended up with a multi-record dataset which were not typical time-to-event set-ups i.e. only including records from selected years (Schatz et al. 2015). One analysis reduced the data into contingency tables with 1 record per combination of all included variables (Dobra et al. 2019). Most datasets used individuals as the unit of analysis, but there were a few that used the household (Sartorius et al. 2014).

Categorising each paper into the data manipulation techniques and dataset structures for this review was done based on the article text, and in some cases some assumptions were made based on knowledge of the type of data and the analyses used. Many journals and funders now require that datasets and programming files be made available to increase scientific transparency and reproducibility. However, just sharing code is not really sufficient, as knowledge of the coding language is needed to fully understand the manipulations. Additionally, many of the decisions leading to specific commands or sets of commands are contingent on the data structure at the time, on earlier analytical decisions and on data

collection structures; it is therefore hard to understand the implications of specific sections of code in isolation from all the other steps. There are documentation standards for static datasets (<https://ddialliance.org/>) but there is currently not an equivalent for the process used to produce the datasets. In response to this problem (which is not unique to HDSS data) the Continuous Capture of Metadata for Statistical Data Project (C2Metadata) has developed an automated system that ‘reads’ code used to manipulate and transform data (currently Stata, R, SPSS, SAS and Python) and ‘translates’ it into a human-readable form that does not require any knowledge of the source program (Alter et al. 2020). Further developments in automating these processes should help HDSS data producers and users. Firstly, it would help HDSS data producers to be able to combine data from different HDSSs, as it would be easier to assess whether specific variables relate to the same or similar concepts; secondly it would help potential data users (whether experienced with other HDSSs or not) use an HDSS dataset appropriately. However, any ‘human-readable’ documentation for a complex data process whether manually or automatically produced, will still be relatively complicated and require skill and experience to use.

Statistical methods

The flexibility of HDSS data also means that broad ranges of statistical methods may be used, from relatively simple to highly complex. A certain level of statistical expertise is required to appropriately utilise longitudinal HDSS data. In the reviewed papers, the majority of statistical techniques used were fairly standard for epidemiology and demography data: survival (time-to-event) analyses (including Kaplan Meier figures, Cox regression, Poisson regression) (readers are referred to Bocquier et al. 2019 for a manual describing how to carry out this kind of analysis with HDSS data), sometimes allowing for competing events (Haber et al. 2017; Kim et al. 2020), and one using landmark analyses (repeating the model only including people with data from certain ages onwards) (Sunny et al. 2019); also logistic or linear regression which may be multinomial (Korinek et al. 2012). The analyses which had a time-to-event dataset split into monthly records, rather than episodes, used multi-level modelling to allow for the data structure (i.e. Oketch et al. 2012). There were two examples of sequence analysis, which allows the comparison and categorisation of multiple longitudinally linked individual data points (Etoori et al. 2021; Larmarange et al. 2015), one that used Sankey diagrams to display movement between categories at different time points (Larmarange et al. 2015) and two using multi-state transition models (Oduro et al. 2022; Bagayoko et al. 2020). Only one analysis used joint modelling (Risher et al. 2021). Joint modelling is now often preferred over more traditional techniques, such as Cox regression, in clinical trial survival analyses as it allows for more accurate and precise estimated of treatment effects (Gould et al., 2015). It can account for informative participant drop-out (see

section on migration below) and techniques are being developed that allow for left- (Crowther et al. 2016) and interval- (Lovblom et al. 2023) censoring (the Risher paper referred to ignored residency gaps as at the time joint modelling was only possible with one record per participant), making it likely that it will become a useful technique for analyses of HDSS data. Some papers used more complex techniques, one using conditional inference tree analysis (Pérez et al. 2021), and three that used Bayesian models (Sartorius et al. 2014; Dobra et al. 2019; Risher et al. 2021).

Repeated measures

There are a few issues regarding repeated measures in HDSS data which can cause complications for analysis: inconsistencies in data repeatedly collected (e.g. age at first marriage reported differently at different time points), repeated events/outcomes either within one individual (e.g. school drop-out may happen multiple times) or within another binding entity (e.g. mother ID where each mother may have more than one child in the analysis, or household ID where more than one household member may be included), or data formatting resulting in multiple records per unit of analysis (i.e. episodic data changed to 1 record per person per month).

Data inconsistencies are a difficult problem to resolve, and the studies in the review dealt with them in different ways: a study on sexual debut and schooling using data from Karonga HDSS used the youngest reported age of sexual debut if there were inconsistencies (Sunny et al. 2019), this may have introduced bias as young age at debut was the exposure of interest; the earliest report may have been preferable. A study in Manicaland HDSS in Zimbabwe, also looking at sexual behaviour data, excluded participants if inconsistencies could not be resolved even though it reduced their sample size from 28,073 to 11,647 (Del Fava et al. 2016). This would have reduced statistical power but also may have introduced bias if those whose inconsistencies could not be resolved were different from those who could. Many did not mention any issue with inconsistencies though for some of these it seems likely that there would have been, i.e. a study in uMkhanyakude HDSS in South Africa used data from several sources to generate an indicator of teen pregnancy but did not mention how they prioritised the data sources (Ardington et al. 2015). Manual attempts at data cleaning rely on the user to detect all data errors and to create appropriate rules to repair them. This may lead to some data errors being missed, and potentially the introduction of bias, i.e. by unconsciously choosing rules that favour a certain category being chosen for uncertain data. There have been developments regarding automated techniques to detect and repair data errors (Chu et al. 2016), which may be of some interest for use in HDSS data, especially as the datasets can get very large.

In the studies which used 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, i.e. in Rufiji HDSS in Tanzania a time-to-event analysis of predictors of parental absence only used the first instance (Gaydos 2015). Doing this reduces complexity of the analysis and interpretation, but does result in loss of information, which may result in making incorrect inferences about the data especially if experiencing multiple events is common.

Repeated records of clustering can be more easily dealt with using an appropriate statistical method, and the majority of analyses did this, 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.

Migration and attrition

HDSSs only record data on participants if they are living within the area. This can make linked analyses challenging as only using data from people who stay in the area may introduce bias. In 'typical' time-to-event analyses which use HDSS data, participants contribute time 'at risk' when they are present, i.e. no-one is excluded for not being present the whole time. This may be 'right' or 'left' censoring if a participant out-migrates or in-migrates, or 'interval' censoring, if they out-migrate and then return. Many of the papers in the review used this approach. Some however restricted their analysis to children who were born in the area, usually because exposure data collected at birth registration were not available for those who in-migrated) but did not exclude those who out-migrated before experiencing the event. For example a study using data from Farafenni HDSS in the Gambia used a time-to-event analysis starting from birth of children in the area and children were censored if they or their mother left the area (as the exposure was mother's vital status) (Scott et al. 2017). Other studies used migration as an outcome or exposure (Muniina 2016; Ziraba 2013) so bias was reduced. Migration must also be taken into account during data manipulation if retrospectively reported information is used to construct dates of exposures or outcomes (see points B, C, D and J in the data manipulation section above): it is important to cross-reference these dates with the residency data to ensure participants observation time is not artificially extended.

Some of the studies using longitudinally linked data excluded all participants who did not have all data points (which may be due to migration), for example a study using data from Nairobi HDSS in Kenya used anthropometry data during childhood linked to a survey in

adolescence where only 692 of 3419 children had all the data required (Oduro et al. 2022), or a study from Matlab HDSS, Bangladesh which linked fertility intention data from 1990 with data on linked children born up to 7 years later and their survival up to 3 years (Bishai et al. 2015). In all these studies there is a risk of bias if those with and without full data are different. 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 (Harling et al. 2014), or b) treating each transition/event as a separate analysis so as many participants can be included in at least one (Haber et al. 2017), or c) include time spent in the HDSS as a control measure (Fotso et al. 2013), d) triangulate the findings with a slightly different analytical approach, i.e. one study carried out a separate matched analysis of sisters to add to their findings as they felt it may have been affected by attrition (Ardington et al. 2015).

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. For example, a study using data from Cuatro Santos HDSS in Nicaragua included women aged 20-24 who were present in the HDSS in the 2014 survey and used previous studies/data to assign teen pregnancy status. Women experiencing teen pregnancy who left the area would not be included, however there was no mention of this in the discussion (Pérez et al. 2021). Others did discuss the issues: i.e. 1) a study using data from Kyamulibwa HDSS in Uganda needed participants to have data from 4 consecutive surveys, they recognised this as a potential limitation but also reported that 79% of those in the 1st survey had data from all 4, so felt that the data could still be representative (Asiki et al. 2019); 2) a study of changes in household composition in older people using data from Agincourt HDSS, South Africa recognised that not being able to include recent in-migrants may have affected the results as they might be different to those included (Madhavan, Schatz, et al. 2017); and 3) a study in Kanchanaburi HDSS in Thailand did a time-to-event analysis with school drop-out as the exposure, they did not include in-migrants as they were using exposure data from a specific survey, they recognised this as a potential issue though reasoned that, as they only had a relatively short follow-up time, there was not likely to have been a big effect (Korinek et al. 2012). A few attempted to examine the effect of attrition, by comparing attributes of those included and not (e.g. Sunny et al. 2018, which found that children lost to follow-up due to relocation were not too different from those who remained), or assessing the effect of the exposure on the outcome of out-migration as well as the outcome under investigation (Finlay et al. 2015; Ardington et al. 2014): both reported little difference.

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. The effect of migration on conclusions from HDSS analyses will vary by HDSS, as some will have relatively lower levels of migration than others (e.g. urban HDSSs may have much higher levels of migration than rural ones). As this is such a key aspect of HDSS data, it is vital to report on it fully in any publication.

Missing data

The final aspect of information that extracted during the review of the papers was on missing data. All studies, regardless of whether the data come from an HDSS or not, may be prone to missing data; however, there are a couple of HDSS-specific issues. Firstly, due to repeated data capture in HDSSs missingness may be related to time spent in the area, so excluding participants missing data may introduce bias. Secondly, and relatedly, in HDSS datasets some missing data may be imputable from earlier or later data points: while this may increase the power of the study, 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.

The majority of studies in the review dealt with missing data in standard ways, i.e. excluding those missing data from the model (Sunny et al. 2018) or retaining them under a category of unknown/missing (Ginsburg et al. 2016), with a few attempting some multiple imputation (Harling et al. 2014; Marston et al. 2013; Anekwe et al. 2015). There were a few instances of data from previous or later time points being used to impute missing data (Nyirenda 2014; Gaydos 2017; Bagayoko et al. 2020), though there was no discussion over the pros and cons of this approach (detailed above). 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 i.e. in an analysis using data from Nairobi HDSS in Kenya the average number of years of schooling per household was calculated, but there was no mention of what was done if any of the data were missing (Mutisya et al. 2016). In this example, bias may be introduced if those missing data are different to those not. For example if data on schooling is more likely to be missing from people who are out at work and this is not taken into account, households with more workers may be assigned an artificially suppressed average education score.

Conclusions of review of published HDSS analyses

The reviewed HDSS data analyses demonstrated the flexibility of HDSS data. Some quite complex statistical techniques were used; however, the majority of analyses used more 'standard' techniques, which shows that advanced statistical skills are not a requirement for conducting valuable analyses with HDSS data. Rigorous data manipulation skills are important, however, which may represent a barrier to some analyses. 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: while statistical methods may undergo scrutiny during the review process, data manipulations which may impact results are often not treated with the same regard. The review focussed on the more complex secondary analyses which may not be representative of all HDSS data analyses; however, some HDSSs were over-represented in the selected papers, indicating that further analyses may be possible using data from other HDSSs.

Overall conclusions and recommendations to producers and users of HDSS data

Through a detailed comparison with the DHS and then further exploration of HDSS-specific issues using published analyses, this paper has demonstrated the value of HDSS data for secondary analysis, and has laid out important issues to take into account when planning, conducting and reporting on secondary analysis of HDSS data. Recommendations to producers and users of HDSS data may be summarised as follows below.

Producers of HDSS data should assess whether their data are being used to the fullest; whether through more complex analyses as listed in this paper, or more standard analyses. The sections on data manipulation, dataset structures and statistical methods, plus the listing of example analyses may help to inspire further analyses. Many HDSSs do not have the capacity to fully analyse all their data, and could benefit from equitable partnerships with analysts from other institutions: either through analysts working together with the HDSS, or through fees for accessing HDSS data, which would be used to fund local data storage, curation and analysis. There are many barriers to sharing HDSS data open access, the key ones being the complexity of the data, the inequity of analysts often from high income settings analysing data produced by professionals in low income settings, and the lack of time and resources needed to properly document and securely anonymise the data. Many of the issues could be solved through comprehensive funding initiatives which would firstly allow HDSSs to analyse their own data, and also be in a position to support external analysts to use the data.

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, and editors and reviewers, 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. It was noted in a meta-analysis on HIV incidence mentioned above, that sometimes the potential impact of participants moving in and out of the area was not totally clear in some HDSS analyses, making interpretation challenging (Birdthistle et al. 2019)

Overall, what this review shows is that the complexity of HDSS data means that a large number of decisions are made while manipulating data for analysis and then analysing data. A number of studies have recently shown that the analytic choices that researchers make can lead to surprisingly different substantive conclusions, even when the same dataset is used by different researchers to answer the same research question (Brezna et al. 2022; Silberzahn et al. 2018). This is the case even when relatively straightforward datasets are used; this problem is likely to be even more pronounced when complex HDSS data are analysed. It is therefore vital that, at a minimum, researchers clearly state all of the decisions that they made when manipulating and analysing data, and explicitly consider the biases that might have been introduced during their analyses. A more complex approach would be to attempt a multiverse analysis: this technique allows for repeated analyses of the same dataset with multiple different sets of assumptions and data management and analytical decisions (Steegen et al. 2016; Cantone et al. 2023); however such a technique may become too complex with some HDSS analyses as the number of alternative decisions may be very large.

Table 2. Listing of all HDSS analyses reviewed, coded according to the 6 aspects (see footnote for meanings)

Title	HDSS, country	period	Subject	Data manip ulation	Data set	Stati stics	Repe ated data	Migr ation	Miss -ing data	Referenc e
The causal effect of childhood measles vaccination on educational attainment: a mother fixed-effects study in rural South Africa	uMkhanyakude, South Africa	1995-2007	Measles vaccination and education	A; B	D	A	D; E	F	D	(Anekwe et al. 2015)
The economic consequences of AIDS mortality in South Africa	uMkhanyakude, South Africa	2000-2009	Impact of mortality of household SES	I	E	A	D; E	A	C	(Ardington et al. 2014)
Early childbearing, human capital attainment, and mortality risk: Evidence from a longitudinal demographic surveillance area in rural KwaZulu-Natal, South Africa	uMkhanyakude, South Africa	2001-2012	Teen pregnancy and association with schooling and mortality	A; C; E; F	D	A	A; D; E	E	C	(Ardington et al. 2015)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
The effect of childhood stunting and wasting on adolescent cardiovascular diseases risk and educational achievement in rural Uganda: a retrospective cohort study	Kyamulibwa , Uganda	2001-2011	Change in stunting status in childhood used as exposure for NCD risk factor and education status.	A	D	A	D	D	A	(Asiki et al. 2019)
Understanding wealth transitions among households in urban slums of Nairobi: A multi-state transition modelling approach	Nairobi, Kenya	2005-2015	Household wealth transitions	B	C	F	D; E	A	B; E	(Bagayoko et al. 2020)
Selection bias in the link between child wantedness and child survival: theory and data from Matlab, Bangladesh	Matlab, Bangladesh	1990-2000	Child mortality in relation to their mother's pre-conception report of desire for more children	J	A	A; B	D	D	A	(Bishai et al. 2015)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
The Crucial Role of Mothers and Siblings in Child Survival: Evidence From 29 Health and Demographic Surveillance Systems in Sub-Saharan Africa	Multiple, Africa	1990-2016	Effects of mother & sib presence, death, migration and birth intervals on child mortality	J	A	B	D	A	A	(Bocquier et al. 2021)
Improving the validity of mathematical models for HIV elimination by incorporating empirical estimates of progression through the HIV treatment cascade	uMkhanyakude, South Africa	2004-2014	Multiple transitions along HIV care pathway assessed for their effect on life expectancy & potential onward transmission	B	A	B	A	E	A	(Chang et al. 2018)
Young children's probability of dying before and after their mother's death: a rural South African population-based surveillance study	Agincourt, South Africa	1994-2008	Child mortality in relation to timing and cause of mother's death	J	B	C	D	A	A	(Clark et al. 2013)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
HIV-seroconversion among HIV-1 serodiscordant married couples in Tanzania: a cohort study	Magu, Tanzania	2006-2016	Factors associated with HIV sero-conversion in HIV discordant couples	J	A	B	E	A	A	(Colombe et al. 2019)
Transition to Parenthood and HIV Infection in Rural Zimbabwe	Manicaland, Mozambique	1998-2011	Sequences from sexual debut, union formation and child-bearing created and grouped and checked for association with HIV status and birth cohort	A; C	D	A	C	E	D	(Del Fava et al. 2016)
Space-time migration patterns and risk of HIV acquisition in rural South Africa	uMkhanyakude, South Africa	2000-2014	Migration & HIV acquisition	A; B	A	B	E	C; F	B	(Dobra et al. 2017)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
A method for statistical analysis of repeated residential movements to link human mobility and HIV acquisition	uMkhanyakude, South Africa	2004-2016	Migration & HIV acquisition	B	F	H	E	C	B	(Dobra et al. 2019)
Patterns of engagement in HIV care during pregnancy and breastfeeding: findings from a cohort study in North-Eastern South Africa	Agincourt, South Africa	2014-2018	HIV diagnosis and treatment through pregnancy and breast-feeding period examined	B; C	A; D	A; B; E	E	B; C	C	(Etoori et al. 2021)
The effects of maternal mortality on infant and child survival in rural Tanzania: a cohort study	Multiple, Tanzania	1996-2012	Maternal mortality and subsequent child mortality	J	A	B	D	A	C	(Finlay et al. 2015)
Birth spacing and child mortality: an analysis of prospective data from the Nairobi urban health and	Nairobi, Kenya	2003-2009	Birth spacing and child mortality	G; J	A	B	D	B; E	D	(Fotso et al. 2013)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
demographic surveillance system										
Childhood risk of parental absence in Tanzania	Rufiji, Tanzania	2001-2011	Risk of parental absence for any cause from birth to age 10	J	C	B	A	B; C	B; E	(Gaydosch 2015)
Beyond Orphanhood: Parental Nonresidence and Child Well-being in Tanzania	Rufiji, Tanzania	1998-2011	Effect of parental non-residence on child mortality and entry into school at appropriate age	J	A	B	D	B; C	B; E	(Gaydosch 2017)
Healthy or Unhealthy Migrants? Identifying Internal Migration Effects on Mortality in Africa using Health and Demographic Surveillance	Multiple, Africa	1998-2012	Migration status categorised as never, in-migrant and returned migrant and used as exposure for	D	A	B	D	C	C	(Ginsburg et al. 2016)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
Systems of the INDEPTH network			mortality risk in adults							
From HIV infection to therapeutic response: a population-based longitudinal HIV cascade-of-care study in KwaZulu-Natal, South Africa	uMkhanyakude, South Africa	2006-2011	Speed of multiple transitions along HIV care pathway	B	A	B	B	E	B	(Haber et al. 2017)
Do age-disparate relationships drive HIV incidence in young women? Evidence from a population cohort in Rural KwaZulu-Natal, South Africa	uMkhanyakude, South Africa	2003-2012	HIV sero-conversion as outcome, age disparity with sexual partner is exposure	B; D	A	B	D	E	D	(Harling et al. 2014)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status	Agincourt, South Africa	1994-2008	Child mortality & household composition	G; J	B	C	D	A	B	(Houle et al. 2013)
Linking the timing of a mother's and child's death: Comparative evidence from two rural South African population-based surveillance studies, 2000–2015	Multiple, South Africa	2000-2015	Child mortality, association with maternal mortality and presence of relatives	G; J	B	C	D	B	A	(Houle et al. 2021)
HIV seroconcordance among heterosexual couples in rural KwaZulu-Natal, South Africa: a population-based analysis	uMkhanyakude, South Africa	2003-2016	HIV incidence & concordance	B; H; J	A	B	E	A	E	(Kim et al. 2020)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
The effect of household and community on school attrition: An analysis of Thai youth	Kanchanaburi, Thailand	2001-2004	Household composition and school drop-out in adolescents	G; H	C	A	D; E	B; C	C	(Korinek et al. 2012)
Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa	uMkhanyakude, South Africa	2003-2012	Participation in repeated HIV survey rounds examined for patterns	C	D	A; E; I	E; F	E	A	(Larmange et al. 2015)
An assessment of childbearing preferences in northern Malawi	Karonga, Malawi	2008-2015	Fertility intentions and subsequent births	J	D	A	E	A	C	(Machiyama et al. 2015)
Child mobility, maternal status, and household composition in rural South Africa	Agincourt, South Africa	1999-2008	Household level variables and their association with child migration	G; J	C	C	D	A	B	(Madhavan et al. 2012)
Social positioning of older persons in rural South Africa: change or stability?	Agincourt, South Africa	2000-2010	Change in household composition in older people	A	E	D	D	B; D	A	(Madhavan, Schatz, et al. 2017)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
Is the risk of HIV acquisition increased during and immediately after pregnancy? A secondary analysis of pooled HIV community-based studies from the ...	Multiple	1989-2012	HIV incidence in relation to pregnancy	B; J	A	B	E	A; F	D	(Marston et al. 2013)
Household survival and changes in characteristics of households in rural South-western Uganda through the period of 1989 to 2008 [Phd thesis]	Rakai, Uganda	1989-2008	Household level variables were created (composition, presence of person with HIV) and used to test association with household dissolution, migration or change in membership	G; I	C	D	E	A; C	A	(Muniina 2016)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
The effect of education on household food security in two informal urban settlements in Kenya: a longitudinal analysis	Nairobi, Kenya	2007-2012	Summary household measures used to assess association with food security	G	E	A	D	A	A	(Mutisya et al. 2016)
How do migrations affect under-five mortality in rural areas? Evidence from Niakhar, Senegal	Niakhar, Senegal	1998-2013	Summary data on migration of household members tested for association with occurrence of U5 mortality in household	A; E; G	E	A	D	C; F	A	(Nguemdjio et al. 2021)
Ageing with HIV: An investigation of the health and well-being of older people in a rural South African population with a	uMkhanyakude, South Africa	2005-2010	Changes in living arrangements for older people	A; G	D; E	A	E	D	C; E	(Nyirenda 2014)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
severe HIV epidemic [Phd thesis]										
A multi-state transition model for child stunting in two urban slum settlements of Nairobi: a longitudinal analysis, 2011-2014.	Nairobi, Kenya	2010-2014	Transition between stunting categories in young children	B	D	F	E	D	A	(Oduro et al. 2022)
Do poverty dynamics explain the shift to an informal private schooling system in the wake of free public primary education in Nairobi slums?	Nairobi, Kenya	2005-2009	Exposure is change in HH SES and outcome is transfer from public to private school	A	E	C	D	E	C	(Oketch et al. 2012)
Trends and factors related to adolescent pregnancies: an incidence trend and conditional inference trees	Cuatro Santos, Nicaragua	2001-2014	Adolescent child-bearing and associations with living arrangements	A; G; H	D	G	F	F	A	(Pérez et al. 2021)

Title	HDSS, country	period	Subject	Data manip ulation	Data set	Stati stics	Repe ated data	Migr ation	Miss -ing data	Referenc e
analysis of northern Nicaragua demographic ...			and other household factors							
Age patterns of HIV incidence in eastern and southern Africa: a modelling analysis of observational population-based cohort studies	Multiple (ALPHA)	1989- 2017	HIV incidence	B	A	H	D; E	A; E	B	(Risher et al. 2021)
The dynamics of household dissolution and change in socio-economic position: A survival model in a rural South Africa	Agincourt, South Africa	1993- 2008	Summary household measures used to assess association with household dissolution	G; I	A	H	D	A	C	(Sartorius et al. 2014)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
Dependent or productive? A new approach to understanding the social positioning of older South Africans through living arrangements	Agincourt, South Africa	2000-2010	Household composition in older people	G	E	D	D	A	A	(Schatz et al. 2015)
Effect of maternal death on child survival in rural West Africa: 25 years of prospective surveillance data in The Gambia	Farafenni, Gambia	1989-2014	Maternal vital status and child survival	J	A	B	D	B	A	(Scott et al. 2017)
An intimate epidemic: HIV and marriage in rural Uganda [Phd thesis]	Masaka & Rakai, Uganda		HIV incidence & concordance	B; J	A	B	E	A; F	B; E	(Sully 2015)
Does early linear growth failure influence later school performance? A cohort study in Karonga district, northern Malawi	Karonga, Malawi	2002-2015	Changes in stunting in childhood used as exposure for later school performance (age for grade)	A	D	A	F	D	B	(Sunny et al. 2018)

Title	HDSS, country	period	Subject	Data manipulation	Data set	Statistics	Repeated data	Migration	Missing data	Reference
Lusting, learning and lasting in school: sexual debut, school performance and dropout among adolescents in primary schools in Karonga district, northern Malawi	Karonga, Malawi	2007-2016	Age at sexual debut examined for association with school drop-out	B; D	A	B	A; D	A	A	(Sunny et al. 2019)
Use of antiretroviral therapy in households and risk of HIV acquisition in rural KwaZulu-Natal, South Africa, 2004–12: a prospective cohort study	uMkhanyakude, South Africa	2004-2012	Household ART usage and HIV acquisition	B; J	A	B	D; E	A	A; C	(Vandormael et al. 2014)
Adult mortality and its impact on children in two informal settlements in Nairobi, Kenya [PhD thesis]	Nairobi, Kenya	2003-2007	Child migration following death in household	J	C	A	D	A	C	(Ziraba 2013)

Codes:

Data manipulation: A=Data from ≥ 2 time points used to create summary exposure or outcome measure; B=Data from ≥ 2 time points used to create 'event' with date; C=Data from ≥ 2 time points used to create sequence; D=Data from ≥ 2 time points used to create time-varying exposure; E=Lagged data used as exposure; F=Matched analysis; G=Data from linked household members used to create summary

exposures or outcomes; H=Data from members of other groups used to create summary exposures or outcomes; I=Data from linked household members used to create household event with date; J=Data from linked person(s) used to create exposure for index

Dataset: A=Time to event (continuous); B=Time to event (monthly); C=Time to event (yearly); D=One record per person; E=Multi-record; F=Summaries

Statistical methods: A=Logistic/linear regression; B=Survival analysis; C=Multi-level modelling; D=Descriptive only; E=Sequence analysis; F=Multi-state transition model; G=Conditional inference tree analysis; H=Bayesian modelling; I=Sankey diagrams

Repeated data: A=First event/report used; B=Events treated separately; C=Excluded inconsistencies; D=Accounted for in model; E=Unclear; F=N/A

Migration: A=Contribute time when present; B=In-migrants excluded; C=Part of analysis; D=Restricted to those with all data points; E=Attempt to reduce effect; F=Potential effect of out-migration

Missing data: A=Not mentioned/none; B=Excluded (cases or variables); C=Missing kept as category; D=Statistical imputation; E=Imputed from other data points

References

- Agarwal, D. et al. 2017. Potential of Health and Demographic Surveillance System in Asthma and Chronic Obstructive Pulmonary Disease Microbiome Research. *Frontiers in Public Health* 5: p.262279.
- Allison, P.D. 1982. Discrete-Time Methods for the Analysis of Event Histories. *Sociological Methodology* 13: p.61.
- Alter, G. et al. 2020. Provenance metadata for statistical data: An introduction to Structured Data Transformation Language (SDTL). *IASS/ST Quarterly* 44(4): p.1–26. Available at: <https://doi.org/10.29173/iq983> [Accessed June 2, 2023].
- Anane-Sarpong, E. et al. 2018. “You cannot collect data using your own resources and put it on open access”: Perspectives from Africa about public health data-sharing. *Developing World Bioethics* 18(4): p.394–405. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/dewb.12159> [Accessed January 4, 2024].
- Anekwe, T.D., Newell, M.L., Tanser, F., Pillay, D., & Bärnighausen, T. 2015. The causal effect of childhood measles vaccination on educational attainment: A mother fixed-effects study in rural South Africa. *Vaccine* 33(38): p.5020–5026.
- Ardington, C., Bärnighausen, T., Case, A., & Menendez, A. 2014. The economic consequences of AIDS mortality in South Africa. *Journal of Development Economics* 111: p.48–60. Available at: [/pmc/articles/PMC4233148/](https://pubmed.ncbi.nlm.nih.gov/24233148/) [Accessed July 18, 2023].
- Ardington, C., Menendez, A., & Mutevedzi, T. 2015. Early childbearing, human capital attainment, and mortality risk: Evidence from a longitudinal demographic surveillance area in rural KwaZulu-Natal, South Africa. *Economic Development and Cultural Change* 63(2): p.281–317. Available at: <https://www.journals.uchicago.edu/doi/10.1086/678983> [Accessed December 1, 2022].
- Arthur, S., Bangha, M., & Sankoh, O. 2013. Review of contributions from HDSSs to research in sexual and reproductive health in low- and middle-income countries. *Tropical Medicine & International Health* 18(12): p.1463–1487. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/tmi.12209> [Accessed July 19, 2023].
- Asiki, G., Newton, R., Marions, L., Kamali, A., & Smedman, L. 2019. The effect of childhood stunting and wasting on adolescent cardiovascular diseases risk and educational achievement in rural Uganda: a retrospective cohort study. *Global Health Action* 12(1): p.1626184. Available at: <https://www.tandfonline.com/doi/full/10.1080/16549716.2019.1626184> [Accessed December 2, 2022].
- Bagayoko, M., Akeyo, D., Kadengye, D.T., & Iddi, S. 2020. Understanding wealth transitions

- among households in urban slums of Nairobi: A multi-state transition modelling approach. *Global Epidemiology* 2: p.100037.
- Becher, H. et al. 2016. Decreasing child mortality, spatial clustering and decreasing disparity in North-Western Burkina Faso. *Tropical Medicine & International Health* 21(4): p.546–555. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/tmi.12673> [Accessed December 1, 2022].
- 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].
- Bishai, D., Razzaque, A., Christiansen, S., Mustafa, A.H.M.G., & Hindin, M. 2015. Selection Bias in the Link Between Child Wantedness and Child Survival: Theory and Data From Matlab, Bangladesh. *Demography* 52(1): p.61–82. Available at: <http://read.dukeupress.edu/demography/article-pdf/52/1/61/877341/61bishai.pdf> [Accessed December 19, 2022].
- Bocquier, P. 2016. Migration Analysis Using Demographic Surveys and Surveillance Systems. In *International Handbook of Migration and Population Distribution*, 205–223. Springer, Dordrecht Available at: https://link.springer.com/chapter/10.1007/978-94-017-7282-2_10 [Accessed July 19, 2023].
- Bocquier, P., Ginsburg, C., & Collinson, M.A. 2019. A training manual for event history analysis using longitudinal data. *BMC Research Notes* 12(1): p.506. Available at: <https://bmcresearchnotes.biomedcentral.com/articles/10.1186/s13104-019-4544-1> [Accessed July 19, 2023].
- Bocquier, P., Ginsburg, C., Herbst, K., Sankoh, O., & Collinson, M.A. 2017. A training manual for event history data management using Health and Demographic Surveillance System data. *BMC Research Notes* 10(1): p.224. Available at: <http://bmcresearchnotes.biomedcentral.com/articles/10.1186/s13104-017-2541-9> [Accessed July 19, 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].
- Bocquier, P., Sankoh, O., & Byass, P. 2017. Are health and demographic surveillance system estimates sufficiently generalisable? *Global Health Action* 10(1). Available at: <https://doi.org/10.1080/16549716.2017.1356621> [Accessed January 6, 2023].
- Breznau, N. et al. 2022. Observing many researchers using the same data and hypothesis

- reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences of the United States of America* 119(44): p.e2203150119. Available at: <https://doi.org/10.1073/pnas.2203150119> [Accessed February 23, 2024].
- Byass, P., Worku, A., Emmelin, A. *et al.* DSS and DHS: longitudinal and cross-sectional viewpoints on child and adolescent mortality in Ethiopia. *Popul Health Metrics* 5, 12 (2007). <https://doi.org/10.1186/1478-7954-5-12>
- Cantone, G.G., & Tomaselli, V. 2023. Theory and methods of the multiverse: an application for panel-based models. *Quality and Quantity*: p.1–34. Available at: <https://doi.org/10.1007/s11135-023-01698-5> [Accessed February 23, 2024].
- Chandramohan, D. *et al.* 2008. Should Data from Demographic Surveillance Systems Be Made More Widely Available to Researchers? *PLoS Medicine* 5(2): p.e57. Available at: <https://dx.plos.org/10.1371/journal.pmed.0050057> [Accessed April 17, 2023].
- Chang, A.Y. *et al.* 2018. Improving the Validity of Mathematical Models for HIV Elimination by Incorporating Empirical Estimates of Progression Through the HIV Treatment Cascade. *JAIDS Journal of Acquired Immune Deficiency Syndromes* 79(5): p.596–604. Available at: <https://journals.lww.com/00126334-201812150-00009> [Accessed December 1, 2022].
- Chawinga, W.D., & Zinn, S. 2019. Global perspectives of research data sharing: A systematic literature review. *Library and Information Science Research* 41(2): p.109–122.
- Chu, X., Ilyas, I.F., Krishnan, S., & Wang, J. 2016. Data cleaning: Overview and emerging challenges. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 2201–2206. Association for Computing Machinery Available at: <https://dl.acm.org/doi/10.1145/2882903.2912574> [Accessed June 10, 2023].
- Clark, S.J. *et al.* 2013. Young Children's Probability of Dying Before and After Their Mother's Death: A Rural South African Population-Based Surveillance Study J. K. Tumwine (ed). *PLoS Medicine* 10(3): p.e1001409. Available at: <https://dx.plos.org/10.1371/journal.pmed.1001409> [Accessed December 1, 2022].
- Cleland, J. 2010. Demographic Data Collection in Less Developed Countries 1946–1996. *Population Studies* 50(3): p.433–450. Available at: <https://www.tandfonline.com/doi/abs/10.1080/0032472031000149556> [Accessed July 14, 2023].
- Colombe, S. *et al.* 2019. HIV-seroconversion among HIV-1 serodiscordant married couples in Tanzania: A cohort study. *BMC Infectious Diseases* 19(1): p.518. Available at: <https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-019-4151-8> [Accessed December 1, 2022].
- Corsi, D.J., Neuman, M., Finlay, J.E., & Subramanian, S. 2012. Demographic and health

- surveys: a profile. *International Journal of Epidemiology* 41(6): p.1602–1613. Available at: <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dys184> [Accessed June 2, 2023].
- Crampin, A.C. et al. 2012. Profile: the Karonga Health and Demographic Surveillance System. *International journal of epidemiology* 41(3): p.676–85. Available at: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3396313&tool=pmcentrez&rendertype=abstract> [Accessed February 2, 2015].
- Crowther, M.J., Andersson, T.M.-., Lambert, P.C., Abrams, K.R., & Humphreys, K. 2016. Joint modelling of longitudinal and survival data: incorporating delayed entry and an assessment of model misspecification. *Statistics in Medicine* 35(7): p.1193–1209. Available at: <https://onlinelibrary.wiley.com/doi/10.1002/sim.6779> [Accessed July 22, 2023].
- Dobra, A., Bärnighausen, T., Vandormael, A., & Tanser, F. 2017. Space-time migration patterns and risk of HIV acquisition in rural South Africa. *AIDS* 31(1): p.137–145. Available at: </pmc/articles/PMC5131684/> [Accessed July 18, 2023].
- Dobra, A., Bärnighausen, T., Vandormael, A., & Tanser, F. 2019. A method for statistical analysis of repeated residential movements to link human mobility and HIV acquisition P. Anglewicz (ed). *PLOS ONE* 14(6): p.e0217284. Available at: <https://dx.plos.org/10.1371/journal.pone.0217284> [Accessed July 17, 2023].
- Dzomba, A., Tomita, A., Govender, K., & Tanser, F. 2019. Effects of Migration on Risky Sexual Behavior and HIV Acquisition in South Africa: A Systematic Review and Meta-analysis, 2000–2017. *AIDS and Behavior* 23(6): p.1396–1430. Available at: <https://doi.org/10.1007/s10461-018-2367-z> [Accessed July 20, 2023].
- Edson Utazi, C. et al. 2018. Geographic coverage of demographic surveillance systems for characterising the drivers of childhood mortality in sub-Saharan Africa. *BMJ Global Health* 3(2): p.611. Available at: <http://gh.bmj.com/> [Accessed January 4, 2023].
- 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].
- Eyre, R.W., House, T., Xavier Gómez-Olivé, F., & Griffiths, F.E. 2021. Bayesian belief network modelling of household food security in rural South Africa. *BMC Public Health* 21(1): p.1–16. Available at: <https://doi.org/10.1186/s12889-021-10938-y> [Accessed January 6, 2023].
- Fabir, M.S., Choi, Y., & Bird, S. 2012. Systematic reviews A systematic review of Demographic and Health Surveys: data availability and utilization for research. *Bull World Health Organ* 90: p.604–612.

- Del Fava, E., Piccarreta, R., Gregson, S., & Melegaro, A. 2016. Transition to Parenthood and HIV Infection in Rural Zimbabwe P. Anglewicz (ed). *PLOS ONE* 11(9): p.e0163730. Available at: <https://dx.plos.org/10.1371/journal.pone.0163730> [Accessed December 16, 2022].
- Finlay, J.E. et al. 2015. The Effects of Maternal Mortality on Infant and Child Survival in Rural Tanzania: A Cohort Study. *Maternal and Child Health Journal* 19(11): p.2393–2402. Available at: <https://link.springer.com/article/10.1007/s10995-015-1758-2> [Accessed December 1, 2022].
- Fotso, J.C., Cleland, J., Mberu, B., Mutua, M., & Elungata, P. 2013. Birth spacing and child mortality: An analysis of prospective data from the nairobi urban health and demographic surveillance system. *Journal of Biosocial Science* 45(6): p.779–798. Available at: <https://doi.org/10.1017/S0021932012000570> [Accessed December 22, 2022].
- Gansaonré, R.J., Moore, L., Bleau, L., Kobiané, J., & Haddad, S. 2022. Stunting, age at school entry and academic performance in developing countries: A systematic review and meta-analysis. *Acta Paediatrica* 111(10): p.1853–1861. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/apa.16449> [Accessed July 20, 2023].
- Gaydos, L. 2015. Childhood Risk of Parental Absence in Tanzania. *Demography* 52(4): p.1121–1146. Available at: <http://read.dukeupress.edu/demography/article-pdf/52/4/1121/877699/1121gaydos.pdf> [Accessed December 1, 2022].
- Gaydos, L. 2017. Beyond Orphanhood: Parental Nonresidence and Child Well-being in Tanzania. *Journal of Marriage and Family* 79(5): p.1369–1387. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/jomf.12422> [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>.
- Ginsburg, C. et al. 2016. Healthy or unhealthy migrants? Identifying internal migration effects on mortality in Africa using health and demographic surveillance systems of the INDEPTH network. *Social Science and Medicine* 164: p.59–73.
- Graybill, L.A. et al. 2020. Incident HIV among pregnant and breast-feeding women in sub-Saharan Africa: A systematic review and meta-analysis. *AIDS* 34(5): p.761–776. Available at: </pmc/articles/PMC7275092/> [Accessed July 20, 2023].
- Haber, N. et al. 2017. From HIV infection to therapeutic response: a population-based longitudinal HIV cascade-of-care study in KwaZulu-Natal, South Africa. *The Lancet HIV* 4(5): p.e223–e230.
- 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., Law, M., et al. 2015. Innovations in health and demographic surveillance systems to establish the causal impacts of HIV policies. *Current Opinion in HIV and AIDS* 10(6): p.483–494. Available at: </pmc/articles/PMC4982533/> [Accessed July 19, 2023].
- Herbst, K., Juvekar, S., et al. 2015. The INDEPTH Data Repository: An International Resource for Longitudinal Population and Health Data from Health and Demographic Surveillance Systems. *Journal of Empirical Research on Human Research Ethics* 10(3): p.324–333. Available at: <https://journals.sagepub.com/doi/full/10.1177/1556264615594600> [Accessed July 19, 2023].
- 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].
- Heuveline, P. et al. 2016. *The Mekong Island Population Laboratory (MIPopLab), A Demographic Surveillance System in Rural Cambodia (2000-06)*. Available at: <http://128.97.186.17/index.php/pwp/article/view/PWP-CCPR-2015-012> [Accessed June 6, 2023].
- Heuveline, P., & Hong, S. 2017. Household structure and child education in Cambodia. *International Journal of Population Studies* 3(2): p.1. Available at: </pmc/articles/PMC7668423/> [Accessed December 1, 2022].
- 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].
- Houle, B. et al. 2013. Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status. *International journal of epidemiology* 42(5): p.1444–54. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/23912808> [Accessed July 23, 2018].
- Houle, B. et al. 2021. Linking the timing of a mother's and child's death: Comparative evidence from two rural South African population-based surveillance studies, 2000–2015 K. Navaneetham (ed). *PLOS ONE* 16(2): p.e0246671. Available at: <https://dx.plos.org/10.1371/journal.pone.0246671> [Accessed December 1, 2022].
- Hunter, L.M. et al. 2021. Working toward effective anonymization for surveillance data: innovation at South Africa's Agincourt Health and Socio-Demographic Surveillance Site. *Population and Environment* 42(4): p.445–476. Available at:

- <https://doi.org/10.1007/s11111-020-00372-4> [Accessed July 19, 2023].
- Jia, P., Sankoh, O., & Tatem, A.J. 2015. Mapping the environmental and socioeconomic coverage of the INDEPTH international health and demographic surveillance system network. *Health and Place* 36: p.88–96.
- Kahn, K. et al. 2012. Profile: Agincourt Health and Socio-demographic Surveillance System. *International Journal of Epidemiology* 41(4): p.988–1001. Available at: <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dys115> [Accessed June 8, 2023].
- Källestål, C. et al. 2020. Assessing the Multiple Dimensions of Poverty. Data Mining Approaches to the 2004–14 Health and Demographic Surveillance System in Cuatro Santos, Nicaragua. *Frontiers in Public Health* 7: p.409. Available at: <https://www.frontiersin.org/article/10.3389/fpubh.2019.00409/full> [Accessed January 6, 2023].
- Kim, H.-Y. et al. 2020. HIV seroconcordance among heterosexual couples in rural KwaZulu-Natal, South Africa: a population-based analysis. Available at: <http://onlinelibrary.wiley.com/doi/10.1002/jia2.25432/full> [Accessed July 17, 2023].
- King, C. et al. 2020. Population impact and effectiveness of sequential 13-valent pneumococcal conjugate and monovalent rotavirus vaccine introduction on infant mortality: prospective birth cohort studies from Malawi. *BMJ Global Health* 5(9): p.e002669. Available at: <http://gh.bmj.com/> [Accessed December 1, 2022].
- Korinek, K., & Punpuing, S. 2012. The Effect of Household and Community on School Attrition: An Analysis of Thai Youth. *Comparative Education Review* 56(3): p.474–510. Available at: <https://www.journals.uchicago.edu/doi/10.1086/666911> [Accessed December 23, 2022].
- Larmarange, J., Mossong, J., Bärnighausen, T., & Newell, M.L. 2015. Participation Dynamics in Population-Based Longitudinal HIV Surveillance in Rural South Africa A. G. Pacheco (ed). *PLOS ONE* 10(4): p.e0123345. Available at: <https://dx.plos.org/10.1371/journal.pone.0123345> [Accessed December 16, 2022].
- Lawrence Gould, A. et al. 2015. Joint modeling of survival and longitudinal non-survival data: Current methods and issues. Report of the DIA Bayesian joint modeling working group. *Statistics in Medicine* 34(14): p.2181–2195. Available at: <https://pubmed.ncbi.nlm.nih.gov/2677775/> [Accessed July 22, 2023].
- Lovblom, L.E., Briollais, L., Tomlinson, G., & Perkins, B.A. 2023. 1336-P: Predicting the Progression of Stages of Neuropathy Using Routine Eye and Kidney Tests—An Application of Novel Statistical Methods in the Diabetes Control and Complications Trial (DCCT). *Diabetes* 72(Supplement_1). Available at: <https://dx.doi.org/10.2337/db23-1336-P> [Accessed July 22, 2023].

- Lynch, R. et al. 2022. Religiosity is associated with greater size, kin density, and geographic dispersal of women's social networks in Bangladesh. *Scientific Reports* 12(1): p.18780. Available at: <https://www.nature.com/articles/s41598-022-22972-w> [Accessed January 16, 2023].
- Machiyama, K. et al. 2015. An Assessment of Childbearing Preferences in Northern Malawi. *Studies in family planning* 46(2): p.161–76. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26059988> [Accessed December 11, 2015].
- Madhavan, S., Myroniuk, T.W., Kuhn, R., & Collinson, M.A. 2017. Household structure vs. composition: Understanding gendered effects on educational progress in rural South Africa. *Demographic research* 37: p.1891–1916. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/29270077> [Accessed July 21, 2018].
- Madhavan, S., Schatz, E., Clark, S., & Collinson, M. 2012. Child mobility, maternal status, and household composition in rural South Africa. *Demography* 49(2): p.699–718. Available at: </pmc/articles/PMC3864506/?report=abstract> [Accessed January 18, 2021].
- Madhavan, S., Schatz, E., Gómez-Olivé, F.X., & Collinson, M. 2017. Social positioning of older persons in rural South Africa: Change or stability? *Journal of Southern African Studies* 43(6): p.1293–1307. Available at: <https://www.tandfonline.com/doi/abs/10.1080/03057070.2017.1365522> [Accessed December 1, 2022].
- Mank, I. et al. 2021. The Impact of Rainfall Variability on Diets and Undernutrition of Young Children in Rural Burkina Faso. *Frontiers in Public Health* 9: p.693281. Available at: </pmc/articles/PMC8489680/> [Accessed December 1, 2022].
- Marston, M. et al. 2013. Is the Risk of HIV Acquisition Increased during and Immediately after Pregnancy? A Secondary Analysis of Pooled HIV Community-Based Studies from the ALPHA Network D. Paraskevis (ed). *PLoS ONE* 8(12): p.e82219. Available at: <https://dx.plos.org/10.1371/journal.pone.0082219> [Accessed July 17, 2023].
- McLean, E. et al. 2021. Data-driven versus traditional definitions of household membership and household composition in demographic studies: does latent class analysis produce meaningful groupings? In *International Population Conference*, Hyderabad, India. Available at: <https://ipc2021.popconf.org/abstracts/210330>.
- McLean, E. et al. 2023. Family network and household composition: a longitudinal dataset derived from the Karonga Health and Demographic Surveillance System, in rural Malawi. *Wellcome Open Research* 8: p.573.
- McParland, D. et al. 2014. Clustering South African households based on their asset status using latent variable models. *Annals of Applied Statistics* 8(2): p.747–776. Available at: </pmc/articles/PMC4256055/> [Accessed January 6, 2023].
- Muniina, P.N. 2016. *Household survival and changes in characteristics of households in*

rural South-western Uganda through the period of 1989 to 2008.

- Murtaza, S.S., Kolpak, P., Bener, A., & Jha, P. 2018. Automated verbal autopsy classification: Using one-against-all ensemble method and Naïve Bayes classifier. *Gates Open Research* 2. Available at: [/pmc/articles/PMC6480413/](https://pmc/articles/PMC6480413/) [Accessed January 9, 2023].
- Mutisya, M., Ngware, M.W., Kabiru, C.W., & Kandala, N. bakwin. 2016. The effect of education on household food security in two informal urban settlements in Kenya: a longitudinal analysis. *Food Security* 8(4): p.743–756. Available at: <https://link.springer.com/article/10.1007/s12571-016-0589-3> [Accessed December 1, 2022].
- Nguemdjo, U., & Ventelou, B. 2021. How Does Migration Affect Under-5 Mortality in Rural areas? Evidence from Niakhar, Senegal. *Population* 76(2): p.341–368. Available at: <https://www.cairn-int.info/journal-population-2021-2-page-341.htm>.
- Nguyen, D.T.N. et al. 2019. Risk of childhood mortality associated with death of a mother in low-and-middle-income countries: A systematic review and meta-analysis. *BMC Public Health* 19(1): p.1–21. Available at: <https://doi.org/10.1186/s12889-019-7316-x> [Accessed July 20, 2023].
- Nyirenda, M. 2014. *Ageing with HIV: An investigation of the health and well-being of older people in a rural South African population with a severe HIV epidemic*. University of Southampton. Available at: https://eprints.soton.ac.uk/366476/1/__soton.ac.uk_uode_personalfiles_users_jep1r13_mydesktop_Makandwe__Approved_eThesis_20140618.pdf.
- Oduro, M.S., Iddi, S., Asiedu, L., Asiki, G., & Kadengye, D.T. 2022. A multi-state transition model for child stunting in two urban slum settlements of Nairobi: a longitudinal analysis, 2011-2014 and for the Nairobi Urban Health and Demographic Surveillance System □. *medRxiv*: p.2022.07.26.22278058. Available at: <https://doi.org/10.1101/2022.07.26.22278058> [Accessed December 1, 2022].
- Oketch, M., Mutisya, M., & Sagwe, J. 2012. Do poverty dynamics explain the shift to an informal private schooling system in the wake of free public primary education in Nairobi slums? *London Review of Education* 10(1): p.3–17. Available at: <https://scienceopen.com/document?vid=7f5fd3e5-cb8b-4e25-ab1f-df6f53340825> [Accessed December 2, 2022].
- Pérez, W. et al. 2021. Trends and factors related to adolescent pregnancies: an incidence trend and conditional inference trees analysis of northern Nicaragua demographic surveillance data. *BMC Pregnancy and Childbirth* 21(1): p.749. Available at: <https://bmcpregnancychildbirth.biomedcentral.com/articles/10.1186/s12884-021-04215-4> [Accessed January 5, 2023].

- Pisani, E., & AbouZahr, C. 2010. Sharing health data: good intentions are not enough. *Bulletin of the World Health Organization* 88(6): p.462–466. Available at: <https://pubmed.ncbi.nlm.nih.gov/20539861/> [Accessed April 19, 2023].
- Reniers, G. et al. 2016. Data Resource Profile: Network for Analysing Longitudinal Population-based HIV/AIDS data on Africa (ALPHA Network). *International journal of epidemiology* 45(1): p.83–93. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/26968480> [Accessed June 24, 2016].
- Reynolds, L. 2015. Category and Kin in “Crisis”: Representations of Kinship, Care, and Vulnerability in Demographic and Ethnographic Research in KwaZulu-Natal, South Africa. *Studies in Comparative International Development* 50(4): p.539–560. Available at: <https://link.springer.com/article/10.1007/s12116-015-9200-8> [Accessed June 6, 2023].
- Risher, K.A. et al. 2021. Age patterns of HIV incidence in eastern and southern Africa: a modelling analysis of observational population-based cohort studies. *The Lancet HIV* 8(7): p.e429–e439. Available at: www.thelancet.com/hiv [Accessed July 17, 2023].
- Roberts, D.A. et al. 2022. Predicting the Risk of Human Immunodeficiency Virus Type 1 (HIV-1) Acquisition in Rural South Africa Using Geospatial Data. *Clinical Infectious Diseases* 75(7): p.1224–1231. Available at: <https://academic.oup.com/cid/article/75/7/1224/6518221> [Accessed July 18, 2023].
- Romero Prieto, J. et al. 2022. *Under-Five Mortality during the Covid-19 Outbreak: Evidence from Four Demographic Surveillance Systems in Low-Income Countries*. Available at: <https://hal.science/hal-03907638> [Accessed July 20, 2023].
- Sartorius, K., Sartorius, B.K., Collinson, M.A., & Tollman, S.M. 2014. The dynamics of household dissolution and change in socio-economic position: A survival model in a rural South Africa. *Development Southern Africa* 31(6): p.775–795. Available at: <http://www.tandfonline.com/doi/abs/10.1080/0376835X.2014.951991> [Accessed December 2, 2022].
- Schatz, E., Madhavan, S., Collinson, M., Gómez-Olivé, F.X., & Ralston, M. 2015. Dependent or Productive? A New Approach to Understanding the Social Positioning of Older South Africans Through Living Arrangements. *Research on Aging* 37(6): p.581–605. Available at: <https://pubmed.ncbi.nlm.nih.gov/26968480/> [Accessed January 14, 2023].
- Scott, S. et al. 2017. Effect of maternal death on child survival in rural West Africa: 25 years of prospective surveillance data in The Gambia R. Paul (ed). *PLOS ONE* 12(2): p.e0172286. Available at: <https://dx.plos.org/10.1371/journal.pone.0172286> [Accessed December 2, 2022].
- Siedner, M.J. et al. 2021. Protocol: Leveraging a demographic and health surveillance system for Covid-19 Surveillance in rural KwaZulu-Natal. *Wellcome Open Research* 5:

- p.1–15. Available at: </pmc/articles/PMC7424917/> [Accessed July 20, 2023].
- Silberzahn, R. et al. 2018. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science* 1(3): p.337–356. Available at: <https://journals.sagepub.com/doi/full/10.1177/2515245917747646> [Accessed February 23, 2024].
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. 2016. Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science* 11(5): p.702–712. Available at: <https://journals.sagepub.com/doi/10.1177/1745691616658637> [Accessed February 23, 2024].
- Sully, E.A. 2015. *An Intimate Epidemic: HIV and Marriage in Rural Uganda*. Princeton, NJ : Princeton University. Available at: <https://dataspace.princeton.edu/handle/88435/dsp012v23vw78h> [Accessed July 18, 2023].
- Sunny, B.S. et al. 2018. Does early linear growth failure influence later school performance? A cohort study in Karonga district, northern Malawi C. M. Schooling (ed). *PLOS ONE* 13(11): p.e0200380. Available at: <https://dx.plos.org/10.1371/journal.pone.0200380> [Accessed December 2, 2022].
- Sunny, B.S. et al. 2019. Lusting, learning and lasting in school: sexual debut, school performance and dropout among adolescents in primary schools in Karonga district, northern Malawi. *Journal of Biosocial Science* 51(5): p.720–736. Available at: https://www.cambridge.org/core/product/identifier/S0021932019000051/type/journal_article [Accessed January 3, 2020].
- Tanser, F. et al. 2008. Cohort Profile: Africa Centre Demographic Information System (ACDIS) and population-based HIV survey. *International Journal of Epidemiology* 37(5): p.956–962. Available at: <https://academic.oup.com/ije/article-lookup/doi/10.1093/ije/dym211> [Accessed June 6, 2023].
- Tatem, A.J., Snow, R.W., & Hay, S.I. 2006. Mapping the environmental coverage of the INDEPTH demographic surveillance system network in rural Africa. *Tropical Medicine and International Health* 11(8): p.1318–1326. Available at: <https://onlinelibrary.wiley.com/doi/10.1111/j.1365-3156.2006.01681.x> [Accessed July 19, 2023].
- Templ, M., Kanjala, C., & Siems, I. 2022. Privacy of Study Participants in Open-access Health and Demographic Surveillance System Data: Requirements Analysis for Data Anonymization. *JMIR Public Health and Surveillance* 8(9): p.e34472. Available at: <https://publichealth.jmir.org/2022/9/e34472> [Accessed June 13, 2023].
- Vandormael, A., Newell, M.L., Bärnighausen, T., & Tanser, F. 2014. Use of antiretroviral

- therapy in households and risk of HIV acquisition in rural KwaZulu-Natal, South Africa, 2004-12: A prospective cohort study. *The Lancet Global Health* 2(4): p.2004–2016. Available at: www.thelancet.com/lancetgh [Accessed July 18, 2023].
- Wasko, Z. et al. 2022. Ethnic diversity and mortality in northwest Burkina Faso: An analysis of the Nouna health and demographic surveillance system from 2000 to 2012 C. C. de A. Pereira (ed). *PLOS Global Public Health* 2(5): p.e0000267. Available at: <https://dx.plos.org/10.1371/journal.pgph.0000267> [Accessed December 2, 2022].
- Werner, L.K. et al. 2022. Health and economic benefits of secondary education in the context of poverty: Evidence from Burkina Faso J. Leight (ed). *PLOS ONE* 17(7): p.e0270246. Available at: <https://dx.plos.org/10.1371/journal.pone.0270246> [Accessed December 2, 2022].
- Williams, J. et al. 2010. Improving public health training and research capacity in Africa: a replicable model for linking training to health and socio-demographic surveillance data. *Global Health Action* 3(1): p.5287. Available at: <https://www.tandfonline.com/doi/full/10.3402/gha.v3i0.5287> [Accessed July 21, 2023].
- Ye, Y., Wamukoya, M., Ezech, 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-Saharan 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].
- Ziraba, A.K. 2013. *Adult mortality and its impact on children in two informal settlements in Nairobi, Kenya*. London School of Hygiene & Tropical Medicine. Available at: <https://researchonline.lshtm.ac.uk/id/eprint/4653405/>.