A replication of a simulation study on measuring and correcting bias of under-5 mortality estimation in populations affected by HIV/AIDS

Sihao Miao, Ruian Yang, Xiaolu Qian, Jiyu Wang, David Wei

24 February, 2020

# Introduction

Generalized linear model (GLM) is a widely used method in research. It is well-established with high flexibility and relatively straightforward interpretation. However, GLMs usually require making assumptions and having the underlying assumptions not hold could invalidate the whole model.

The application of GLM can be observed in many fields including public health. Quattrochi et al. (2019) introduced an GLM via penalized maximum likelihood to correct the under-5 mortality (U5M) in populations affected by HIV/AIDS. This method was developed since the U5M is commonly estimated using survey-based approaches in less developed regions and it is based on a fundamental assumption that the mother’s survival is not related to her children’s, which is often violated in populations affected by HIV/AIDS. Therefore, the author sought a method to correct this bias using GLM.

In our work, we present our endeavor to replicate the model building and regression analysis introduced in Quattrochi et al. (2019). In the meantime, we attempted using different parameters in both constructing the model and applying it to test how robust the original method was. We will in this paper report our method and results from the replication attempt.

# Methods

## Overview

The original model was implemented in R (v3.5.2) [team2013r] with web application (<https://johnquattrochi.com/bias/>) to show that indirect method such as surveying can underestimate U5M by 0-41% in populations where 0-40% of the people are affected by HIV/AIDS. The model was written using the R language [team2013r] and some of the main packages used were tidyverse (v1.2.1) and foreach (v1.4.4). The original implementation utilized the doReids package (v1.1.1) that launches a back-end server in Docker for parallel processing that was not adopted in our replication. The main limitation is the computing resource we had for this project within a short period of time. Based on experiments performed on a m4.xlarge EC2 instance on AWS, the original simulation settings in the target paper consumed over 63% CPU and 4GB of memory and kept running for over 3 days using Docker doRedis package. The operator estimated of finishing the entire simulation step would require an 8-core CPU machine more than 1 week. In order to complete the simulation for constructing the model in a reasonable time frame without parallel computing, we had to stop the simulation and reduced the number of populations used in the simulation input from 4480 to 1000. The code provided by the author on Github (<https://github.com/jquattro/hiv-childmort-bias>) was mostly well-organized and documented. However, we discovered some mistakes regarding output inconsistency and wrong variable names. Despite the debugging process, we were able to replicate the implementation.

## Model Description

The goal of the large-scale simulation used in the original paper was to build a demographic history that accounted for the variability among actual populations to measure if biases such as the independence assumption of mother and children survival rate in a population influenced by HIV/AIDS existed. Quattrochi et al. (2019) used stochastic simulation methods to generate a set of populations with input parameters including demographic features, mortality rate for U5M, maternal mortality rate, HIV prevalence, etc. The input simulation parameters vary over a large interval for different age groups across different populations. In each time step of one year, each woman in a population has some probabilities to give birth (for women between 15 to 49 years old), being infected with HIV, initiating ART (if HIV-positive), and dying given the simulation parameters. Through the aging process in the simulation population, children born with infected HIV mothers will have a probability of getting infected while all children will have a probability of mortality.

For calculating relative and absolute bias, we used the formulas provided in the original paper.

According to (Quattrochi et al. 2019), was the estimattion of U5M based on living women in 2010, and was the estimation of U5M based on living women in 2010 and women died from HIV/AIDS before 2010 given the condition that they would have been 15-49 in 2010 if they had survived.

## Simulation Construction

### Range of Inputs

Since the primary goal in Quattrochi et al. (2019) was to measure bias in indirect estimates across a set of populations that have experienced different rates of fertility, mortality, HIV infection, and ART (antiretroviral therapy) initiation. We adopted the same method to generate a set of populations using the following inputs: fertility, adult mortality, U5M, percent of 15–19 year olds who are sexually active, maternal mortality in 1990, percent annual decline in the maternal mortality rate, HIV incidence, duration of breastfeeding, and ART coverage. The original paper simulated one population for each combination of inputs resulting in a total of 4480 populations. Since the original paper did not provide information on the amount of time required to complete the simulation, we started our process with the same 4480 populations but without parallel computing, we could not complete the simulation within the timeframe of our study and stopped the simulation when it reached 1000 populations. For the purpose of our replication study, we set the seed for simulation to be 1, which was different from the seed value 1000 in the original paper.

### Simulation Parameters

Initially, we attempted to some parameters in the simulation construction process. We tried to reduce the year interval of our study from [1906, 2010] to [1906, 1966] as well as the population age interval from [0,120] to [0,75]. This turned out to be unsuccessful since the the above parameters were not only used in the simulation process but also in the following model selection and regression application. For instance, to estimate HIV prevalence, the model divides the population into 7 age groups, which cannot be achieved if the year or age parameter was reduced. As a result, we used the same size and date for the initial population in the simulation. We initiated with 22,500 women who were aged 15 years in 1906, and ran the simulation through 2010.

For other parameters in simulation, we followed the original paper and used the same data sources. Annual probability of birth was defined as a function of calendar year and mother’s age and set to zero for women younger than 15 years and older than 49 years. Estimates of age-specific fertility rates (ASFR) from the United Nations Population Division’s World Fertility Data (“World Fertility Data 2012” 2012) was used and for years when ASFR were not available, the adjusted the nearest available ASFR using the interpolated estimates of the total fertility rate (TFR) from the United Nations Population Division’s World Population Prospects (“World Population Prospects - Population Division” 2013) was used.

As mentioned in the original paper, according to (Chen and Walker 2010), women between 15–19 years old who were HIV-positive experienced higher ASFRs compared to HIV-negative women, depending if they were sexually active. Therefore the ASFR parameter was adjusted using the ratios estimated by (Chen and Walker 2010). ASFR was also adjusted based on the studies that have found an increase in pregnancy following ART (Myer et al. 2010)(Makumbi et al. 2011)(Homsy et al. 2009). The effect of ART on fertility likely depends on a variety of factors including age, cluster of differentiation 4 (CD4) count at initiation according to the original paper. For the simulation model, the assumption that the ASFR for 15–19 year olds is not affected by ART was made to further simplify the simulation. Inputs relating to maternal mortality was calculated using MMR (maternal deaths per 100,000 births) in 1990 and the annual decline in MMR since 1990. The initial value of the MMR according to (M. C. Hogan et al. 2010) was either 0.0012 or 0.012 and the annual rate of decline was set to 0 or 7.3%.

For other inputs including the annual probability of HIV infection(D. R. Hogan and Salomon 2012)(Heuveline 2003), the parameters governing CD4(Hallett et al. 2008), ART coverage(“World Development Indicators” 2012), the annual probability of death(“Levels and Trends in Child Mortality 2012” 2013) and the probability of mother-to-child transmission of HIV(Stover et al. 2008), we also used the same data sources and derivation methods.

## Model Fitting and Model Selection

Newly-generated simulation data were then imported and fitted to various models to select for a best-performing model and the selection process used least errors to represent model performances.

The simulated data were first divided into training set (80%) and test set (20%). The training set was fitted using eight different models, which were further evaluated by calculating several error metrics (root mean square error, root median square error, mean relative error and median relative error) on both training set (in-sample) and test set (out-of-sample) shown in Table 1(Fig. 1).

Compared to the estimation errors calculated in the original paper, we had very similar results of the root mean square errors and root median square errors for all models. However, our in-sample mean relative errors were consistently greater compared to the original paper and we suspect this could be due to the reduction in population size during the simulation. Consistent with the original paper, no single model in the table showed clear advantages for all the metrics over others we selected the generalized linear regression with alpha equal to 1 as our predictive model since it showed the best performance for most of the metrics.

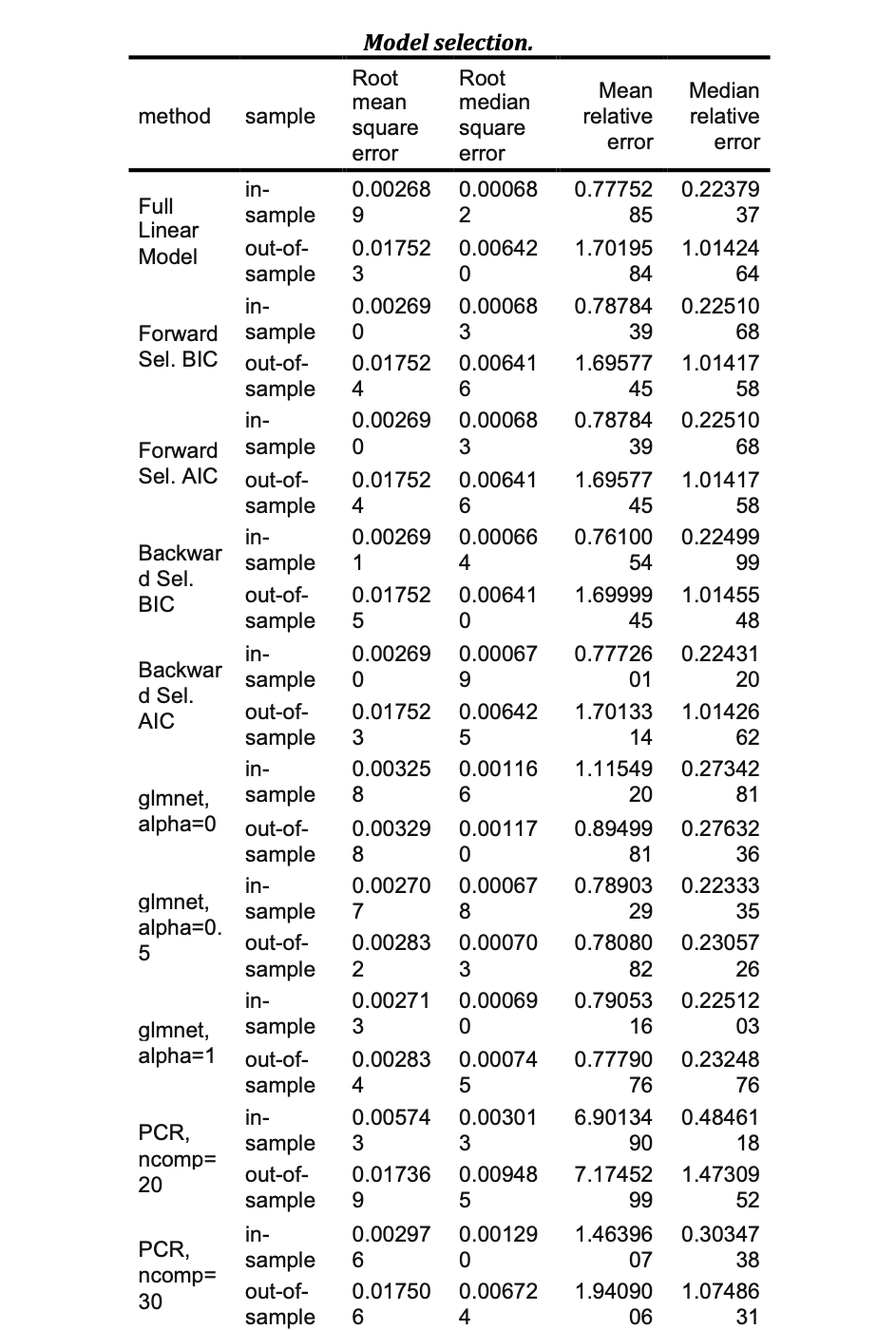


Figure 1: Table 1

## Replication of Figures

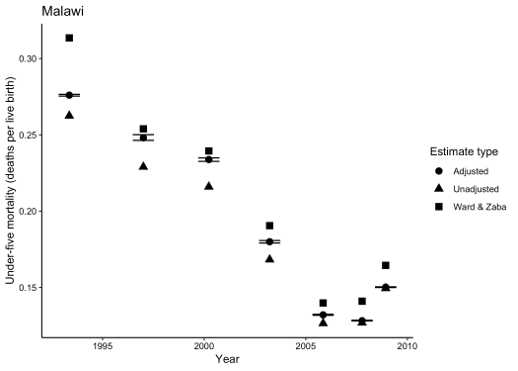


Figure 2: Original figure 4

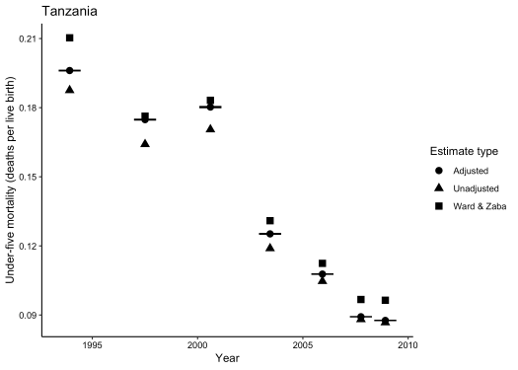


Figure 3: Original figure 5

Figure 4 (Fig. 2) and 5 (Fig. 3) in the original article were created by applying the best-performing model as well as an independent model (Ward and Zaba 2008) to the empirical data from 2010 DHS for Malawi and Tanzania (Statistics - NBS/Tanzania and Macro 2011)(Office and Macro 2011) respectively to estimate U5M of different age groups of population. By comparing differences in U5M estimates of unadjusted data and adjusted data by two models, the authors suggested that the new model was able to correct the bias from indirect estimates.

With the new simulation data and best-fitting model, we re-generated both figures by using R scripts provided (Fig. 4 and Fig. 5).

To prepare data for plotting, we created three functions with R code provided by the authors to estimate U5M using the indirect method, the Ward-Zaba model and our best-performing model. These three functions were then further integrated into a function called plot\_fig to create our replicated figures. We also included a file with necessary data for calculating the confidence interval by bootstrapping for prediction with our best-fitting model, since this step usually takes more than 40 minutes to finish.

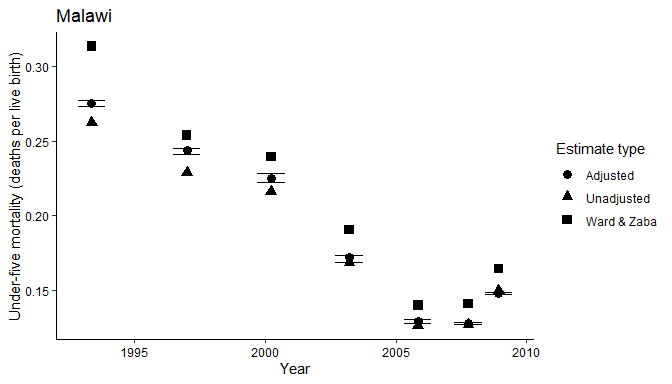


Figure 4: Replication of figure 4 with new simulation data

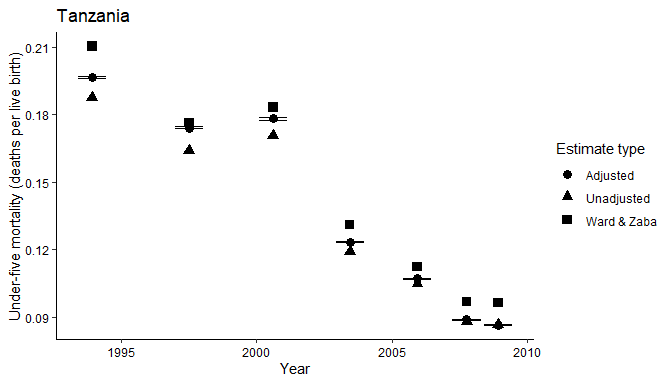


Figure 5: Replication of figure 5 with new simulation data

We revised some parts of the regression model and corrected inconsistent variable names in the original code during the debugging process. For instance, caught corrected the variable name “lambda.min” where it should have been “lambda”.

boot\_models <- list(length=5)  
 t0 <- predict(best.fitting.model,newx = x,s="lambda.min") %>% as.vector  
 tt <- boot\_models[1:5] %>%purrr::map(~predict(best.fitting.model,newx = x,s="lambda.min") %>% as.vector) %>% unlist()  
 #lambda <- best.fitting.model$lambda.min  
 lambda <- best.fitting.model$lambda  
 theta <- as.matrix(coef(best.fitting.model, s = "lambda.min"))  
 pi\_hat <- predict(best.fitting.model, newx = x, s = "lambda.min", type = "response")  
 n <- length(pi\_hat)  
 y\_star <- sapply(seq\_len(length(tt)), function(i) ifelse(runif(n) <= pi\_hat, 1, 0))  
 suff\_stat <- t(y\_star) %\*% x

# Result

Our implementation of the model to estimate the bias in U5M in HIV/AIDS affected populations from Quattrochi et al. (2019) was able to replicate the process and output target tables and figures. Consistently, indirect methods underestimate U5M in older populations, while little difference was shown in younger age groups between unadjusted and adjusted estimates. Applying our model to 2010 survey data from Malawi and Tanzania, we show that indirect methods would underestimate U5M by up to 7%.

Table 2 (Fig. 6) shows our estimation by applying the best performing model on 1,000 simulated populations and consistent with the original implementation, the mean HIV prevalence we calculated among women aged 15–49 across populations was 7% in 1990, 15% in 2000, and 11% in 2010. The highest HIV prevalence in any simulation was 40% in 2000. Our results related to ART showed the same trend as in the original paper that both prevalence and coverage increase with time, although the exact numbers varied. The mean ART coverage across our simulations was 0% in 2004 and 8% in 2010. The mean ART prevalence was < 0.1% in 2005 and 0.19% in 2009. The highest ART prevalence in any simulation was 2.3% in 2009 and this was 1.7% lower compared to the original model. The mean TFR across simulations was 4.88 in 2000 and 4.28 in 2010, also consistent with the original model.

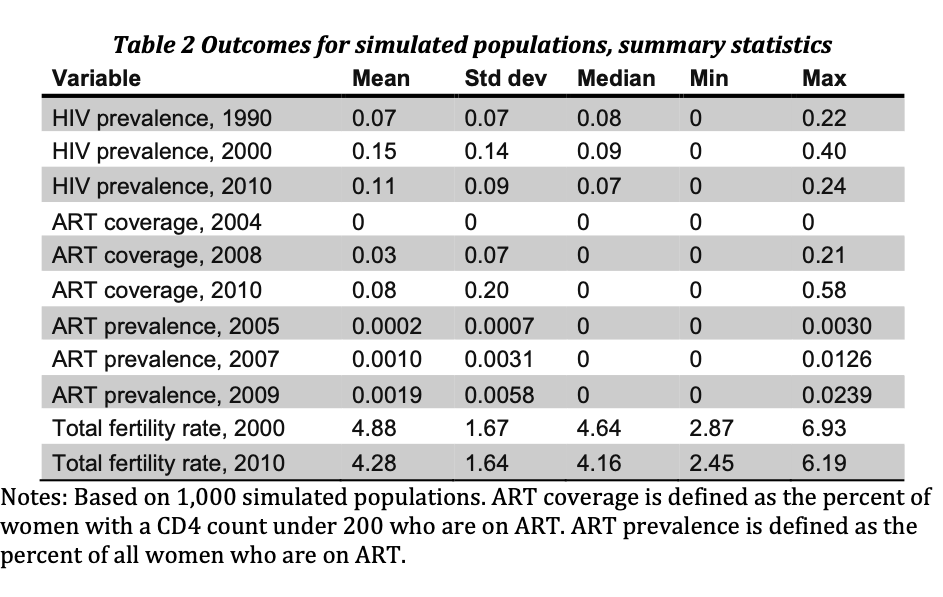


Figure 6: Table 2

Table 3 (Fig. 7) shows the bias in indirect estimates across age groups. According to the original paper, negative numbers indicate that unadjusted estimates were lower than adjusted estimates. The mean absolute bias was largest for estimates from women aged 35–39 and 40–44 (−0.021) and smallest for estimates from women aged 15–19 (0.000). Our results showed the same age group with a small variance in the exact numbers. The original paper also emphasized on the largest absolute bias recorded, which was −0.071 for estimates from women 35–39 in our calculation and -0.069 for them. The paper pointed out that this means that the estimated U5M was 71 deaths per 1000 live births lower when using only reports from surviving women compared to reports from surviving women and women who died from HIV/AIDS.

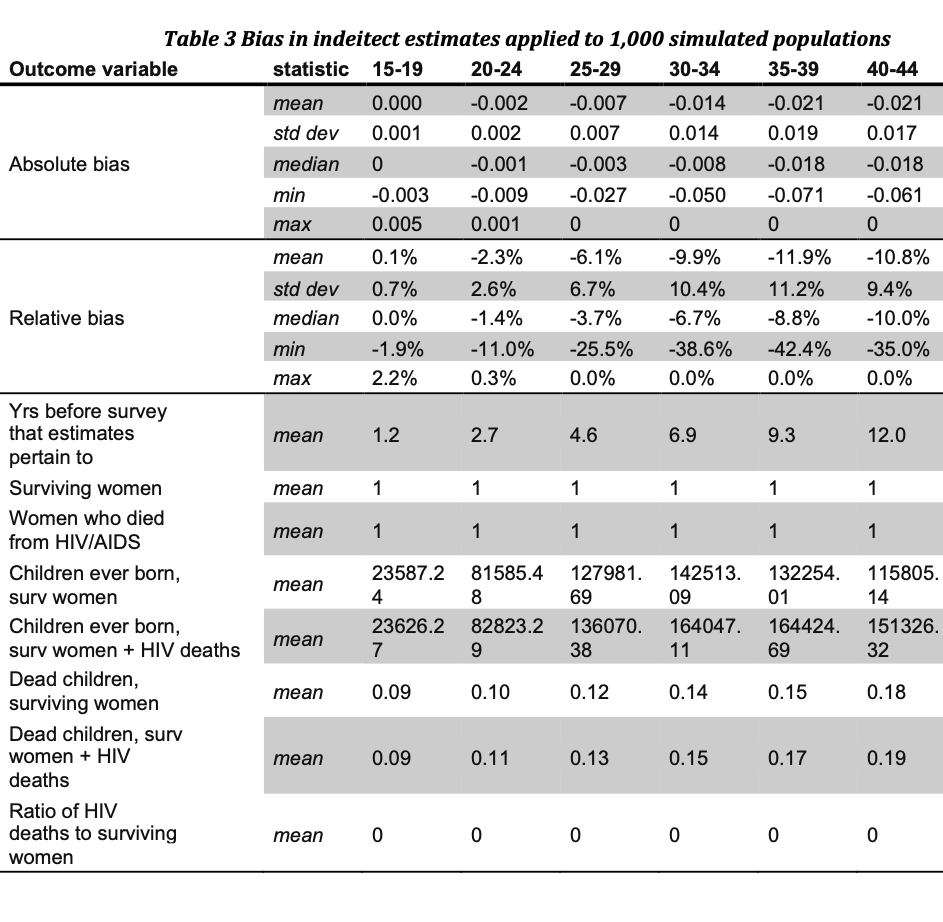


Figure 7: Table 3

Based on our results, we believe that our effort in replicating Quattrochi et al. (2019) was successful despite some inconsistency. We argue that the differences could be due to the smaller population size we used in our simulation though further investigation would be needed to discover the exact causes.

# Conclusion

Our goal of this project was to replicate a published research paper and reproduce figures in it. We were able to accomplish this goal by replicating the key figures in the original paper using R and publishing well-structured source code along with data source in a licensed Github repository. This paper.Rmd R Markdown file is fully executable therefore can be used as the main entry point to our project. However, we intentionally set a couple code chunks with evaluation as False since simulating those data sources does not fit within the timeframe for any group trying to reproduce them. Instead, we created .RData file as a battery-included solution for future reproducers.

We leveraged some useful reproducibility techniques introduced in the class such as Tidyverse libraries and R Markdown syntax. The in-file executable R chunks come extremely handy during our reproduction process.

There are, however, challenges throughout our explorations. The first major issue we encountered was the lengthy source data generation step which required extensive research and deep dive into the source code. Given the nature of the HIV/AIDS estimation research, the data samples required had to be comprehensive enough to cover multiple geological locations and a wide range of populations. Therefore, the original authors designed the raw data to be generated on-the-go rather than archived in open source. Our team initially underestimated the challenge of reproducing all sources and this was also due to the lack of information on completion time in the original paper. We had to reduce population size during simulation to ensure the quality of our research and a timely deliverable. Looking back, we can identify this as an area that we could improve (focusing on verification of data source before replication). Another challenge was using R in an unfamiliar field of research. Given none of us had previous domain knowledge regarding HIV/AIDS research, we spent a lot of time to understand the terminologies while exploring the rationales behind the R code used in the original paper. We had issues trying to understnad specific failures in R givern that multiple team members had different R environment. For example, we still couldn’t fully understand some of the “knit” failures which had been resolved by cleaning up cache in R Studio. We strongly believe that using Docker or a shared development environment could drastically improve our developer experience where most errors could be reproduced easily no matter where the Docker image is running.

In general, we gained hands-on experience from reproducing a research project. After rounds of reviews and merging conflicts, we developed our own Github usage convention and code styles. We would like to carry these invaluable skills to our future Data Science journey.

# References cited

Chen, Wei-Ju, and Neff Walker. 2010. “Fertility of Hiv-Infected Women: Insights from Demographic and Health Surveys.” *Sexually Transmitted Infections* 86 (Suppl 2). The Medical Society for the Study of Venereal Disease: ii22–ii27.

Hallett, Timothy B, Basia Zaba, Jim Todd, Ben Lopman, Wambura Mwita, Sam Biraro, Simon Gregson, J Ties Boerma, Alpha Network, and others. 2008. “Estimating Incidence from Prevalence in Generalised Hiv Epidemics: Methods and Validation.” *PLoS Medicine* 5 (4). Public Library of Science.

Heuveline, Patrick. 2003. “HIV and Population Dynamics: A General Model and Maximum-Likelihood Standards for East Africa.” *Demography* 40 (2). Springer: 217–45.

Hogan, Daniel R, and Joshua A Salomon. 2012. “Spline-Based Modelling of Trends in the Force of Hiv Infection, with Application to the Unaids Estimation and Projection Package.” *Sexually Transmitted Infections* 88 (Suppl 2). BMJ Publishing Group Ltd: i52–i57.

Hogan, Margaret C, Kyle J Foreman, Mohsen Naghavi, Stephanie Y Ahn, Mengru Wang, Susanna M Makela, Alan D Lopez, Rafael Lozano, and Christopher JL Murray. 2010. “Maternal Mortality for 181 Countries, 1980–2008: A Systematic Analysis of Progress Towards Millennium Development Goal 5.” *The Lancet* 375 (9726). Elsevier: 1609–23.

Homsy, Jaco, Rebecca Bunnell, David Moore, Rachel King, Samuel Malamba, Rose Nakityo, David Glidden, Jordan Tappero, and Jonathan Mermin. 2009. “Reproductive Intentions and Outcomes Among Women on Antiretroviral Therapy in Rural Uganda: A Prospective Cohort Study.” *PloS One* 4 (1). Public Library of Science.

“Levels and Trends in Child Mortality 2012.” 2013. *World Health Organization*. World Health Organization. <https://www.who.int/maternal_child_adolescent/documents/levels_trends_child_mortality_2012/en/>.

Makumbi, Fredrick, Gertrude Nakigozi, Steven Reynolds, Anthony Ndyanabo, Tom Lutalo, David Serwada, Fred Nalugoda, Maria Wawer, Ron Gray, and others. 2011. “Associations Between Hiv Antiretroviral Therapy and the Prevalence and Incidence of Pregnancy in Rakai, Uganda.” *AIDS Research and Treatment* 2011. Hindawi.

Myer, Landon, Rosalind J Carter, Monica Katyal, Patricia Toro, Wafaa M El-Sadr, and Elaine J Abrams. 2010. “Impact of Antiretroviral Therapy on Incidence of Pregnancy Among Hiv-Infected Women in Sub-Saharan Africa: A Cohort Study.” *PLoS Medicine* 7 (2). Public Library of Science.

Office, National Statistical, and ICF Macro. 2011. “Malawi Dhs, 2010 - Final Report (English).” *Malawi: DHS, 2010 - Final Report (English)*. <https://dhsprogram.com/publications/publication-fr247-dhs-final-reports.cfm>.

Quattrochi, John, Joshua A Salomon, Kenneth Hill, and Marcia C Castro. 2019. “Measuring and Correcting Bias in Indirect Estimates of Under-5 Mortality in Populations Affected by Hiv/Aids: A Simulation Study.” *BMC Public Health* 19 (1). Springer: 1516.

Statistics - NBS/Tanzania, National Bureau of, and ICF Macro. 2011. “Tanzania Dhs, 2010 - Final Report (English).” *Tanzania: DHS, 2010 - Final Report (English)*. <https://dhsprogram.com/publications/publication-fr243-dhs-final-reports.cfm>.

Stover, J, P Johnson, B Zaba, M Zwahlen, F Dabis, and RE Ekpini. 2008. “The Spectrum Projection Package: Improvements in Estimating Mortality, Art Needs, Pmtct Impact and Uncertainty Bounds.” *Sexually Transmitted Infections* 84 (Suppl 1). The Medical Society for the Study of Venereal Disease: i24–i30.

Ward, Patrick, and Basia Zaba. 2008. “The Effect of Hiv on the Estimation of Child Mortality Using the Children Surviving/Children Ever Born Technique.” *Southern African Journal of Demography*. JSTOR, 39–73.

“World Development Indicators.” 2012. *DataBank*. <http://databank.worldbank.org/reports.aspx?source=world-development-indicators>.

“World Fertility Data 2012.” 2012. *United Nations*. United Nations. <https://www.un.org/en/development/desa/population/publications/dataset/fertility/wfd2012/MainFrame.html>.

“World Population Prospects - Population Division.” 2013. *United Nations*. United Nations. <https://population.un.org/wpp/Publications/>.