README File

Replication Package for "Adverse Selection in the Marriage Market: HIV Testing and Marriage in Rural Malawi"

Manuela Angelucci and Daniel Bennett
November 2020

Data Availability Statement

This paper uses data from the Tsogolo la Thanzi (TLT) Panel Study, which was collected by Jenny Trinitapoli and Sarah Yeatman (Trinapoli and Yeatman 2018a, Trinitapoli and Yeatman 2018b). The TLT Study website can be accessed here: https://tsogololathanzi.uchicago.edu. We wrote this paper using an earlier version of the dataset that we obtained from the researchers. Since then, the researchers have placed the data in the ICPSR repository at the University of Michigan, where it is available at the following link: https://www.icpsr.umich.edu/web/DSDR/series/767. To obtain the data, applicants must apply to ICPSR for restricted data access. Applicants must have IRB approval or an exemption to proceed. The application instructions are available on the ICPSR website.

Because Drs. Trinitapoli and Yeatman cleaned the data further before submitting it to ICPSR, the files that are available through the ICPSR are not identical to the files we used for this analysis.

Our analysis uses data that can be found in the following ICPSR data files:

- Baseline Wave (Rounds 1-8): women dataset, male partner dataset
- Biomarker Data: HIV dataset, pregnancy dataset
- Couples Data: file linking women to male partners.

We also use data from the 1992, 2000, 2004, and 2010 waves of the Malawi Demographic and Health Survey (DHS) (Malawi National Statistical Office 1992, 2000, 2004, 2010). These data are available through a brief application though the DHS website: https://www.dhsprogram.com/Where-We-Work/Country-Main.cfm?ctry_id=24&c=Malawi

Description of Datasets in the Analysis

TLT Baseline Wave (Rounds 1-8): these files provide the foundation for our analysis. Beginning in 2009, surveyors administered eight longitudinal survey waves to 1500 sample

women and male partners. Waves were spaced four months apart. The ICPSR site has separate files by round (1-8) and for women and male partners. Another file with randomly chosen men is not used our analysis. These files contain all survey responses. Our code appends these files together.

Biomarker Data: the HIV dataset provides all HIV test results for respondents who were tested as part of the study. Eligible respondents were offered testing at the TLT clinic immediately after completing the survey interview. The pregnancy dataset provides pregnancy test results. Female respondents were asked to complete urine-based pregnancy tests during every survey interview.

Couples Data: this uses respondent identifiers to provide a crosswalk between the identities of female respondents and male partners. Observations are by couple identifier and wave, allowing participants to have partnerships with multiple people over the study period.

We use these files to construct two primary estimation datasets:

Main Estimation Dataset: this dataset contains observations for all female respondents and male partners over Waves 1-8 and includes HIV and pregnancy test results. We subsequently limit the sample in order to estimate effects for subgroups, such as our main estimates, which focus on baseline-unmarried female respondents.

Single-Test Estimation Dataset: For the analysis in Table 6 of the paper, we rearrange the data to create a comparison between respondents who were offered testing in Wave 4 and Wave 8, and respondents who were only offered testing in Wave 8. Since HIV tests occurred after the survey interviews, this comparison allows us to assess the impact of being offered an HIV test only once (in Wave 4). To construct this dataset, we realign the waves for the treatment arms to make the comparison between the single-test arm in Waves 4-8 and the multiple-test arm in Waves 1-5. In both cases, we compare the treatment groups to the control group over comparable time periods.

Variables Used in the Analysis

The main variables used in this analysis are:

- A marriage indicator that equals 1 if the respondent is married in Wave t and 0 otherwise. A person is defined as being married in she identifies a person as her husband.
- A pregnancy indicator that equals 1 if the respondent tested positive in a pregnancy test in Wave t and zero otherwise.
- A treatment assignment variable that indicates which intervention arm the respondent was assigned to. There were three intervention arms. The first arm was offered HIV

testing after every survey wave, the second arm was offered testing after Waves 4 and 8, and the third arm was only offered HIV testing after Wave 8. Our analysis primarily relies on a comparison of the multiple-test arm and the control arm over Waves 2-8. However Table 6 relies on a comparison of the single-test arm and the control arm over Waves 5-8, coupled with a comparison of the multiple-test arm and the control arm over Waves 1-5 (with the sample limited in this way for comparability with the single-test comparison).

- Age of the respondent is specified in years and ranges from 15-25 in Wave 1. Because the multiple-test arm and the control arm are not balanced by age (as we explain in the paper), most estimates use entropy weights to balance across intervention arms by age.
- Sexual Safety Index: we define a sexual risk score based on several baseline variables. These include (1) whether the respondent reports more than two lifetime partners, (2) whether the respondent has been with more than one partner in the past 12 months, (3) whether the respondent has had sexual contact for money, (4) whether the respondent reports having sex at least four times per week, and (5) whether the respondent has ever taken antiretroviral medication. A person is defined as safe if she has zero of these traits and is defined as unsafe if she has any of these traits at baseline. This definition is primarily used in Panel A of Tables 4 and 5.
- Perceived HIV Infection Risk: respondents report the probability that they are currently HIV positive on a probability scale in 10 percentage point increments. Respondents are defined as safe if they report a 0% or 10% chance of having HIV. This safety definition is primarily used in Panel B of Tables 4 and 5.
- Attractiveness is measured on a five-point Likert scale. Attractiveness was assessed by female surveyors during interviews. We use baseline values of this variable and create a binary version in which people with the top three values of this variable are coded as attractive and the bottom two values are coded as unattractive. We primarily use this variable for the interactions in Table 5.

Analysis Code

This replication package contains all of the programs needed to replicate our results using the data that is available by application from the ICPSR.

- *restud_build.do* assembles the main dataset and creates the Table 6 dataset.
- *restud_build_dhs.do* assembles the datasets used for the DHS analysis (Figures 1 and 5). These build files may be run in any order.
- *restud_analysis.do* uses the output data from *restud_build.do* and *restud_build_dhs.do* to produce all tables and figures in the paper and the appendix, as well as the statistics cited in the manuscript.

These files were written using Stata 14. Each of these do-files includes several programs that are called within the file. The programs needed to run each file are defined within that file, so it should not be necessary to run all of the do files within the same Stata session.

Our do files require the following non-standard packages:

- ebalance
- estout

The user can install these packages by typing "ssc install ebalance" and "ssc install estout" in the Stata command line.

Output Files

Since we cannot provide the raw data files used for this analysis, this replication package includes log files that document the output from running these files. These files can be found in the output folder.

- *restud_build.log*
- *restud_build_dhs.log*
- *restud_analysis.log*: generates all tables, figures, and cited numbers in the paper.
- *marriage_restud_accepted.pdf*: the final version of the manuscript and appendix, which includes all tables, figures, and numbers reported in the text.
- *Figures_v10_Restud_Final.xlsx:* contains all figures in the manuscript and appendix as well as underlying data for these figures.
- .tex files that save the regression output associated with tables in the paper and appendix.
- .csv files that save the regression output associated with figures in the paper and appendix.

References

National Statistical Office [Malawi] and Macro International. 1992. Malawi Demographic and Health Survey 1992 [Dataset]. MWIR22FL.dta.

National Statistical Office [Malawi] and ORC Macro. 2000. Malawi Demographic and Health Survey 2000 [Dataset]. MWIR41FL.dta.

National Statistical Office [Malawi] and ORC Macro. 2004. Malawi Demographic and Health Survey 2004 [Dataset]. MWIR4DFL.dta.

National Statistical Office [Malawi] and ICF Macro. 2010. Malawi Demographic and Health Survey 2010 [Dataset]. MWIR61FL.dta.

Trinitapoli, Jenny Ann, and Yeatman, Sara. Tsogolo La Thanzi (TLT): Baseline Wave, Malawi, 2009-2012 [Healthy Futures]. Inter-university Consortium for Political and Social Research [distributor], 2018-10-22. https://doi.org/10.3886/ICPSR36863.v3

Trinitapoli, Jenny Ann, and Yeatman, Sara. Tsogolo La Thanzi (TLT): Biomarker Data, Malawi, 2009-2012, 2015 [Healthy Futures]. Inter-university Consortium for Political and Social Research [distributor], 2018-11-29. https://doi.org/10.3886/ICPSR37200.v2