Practical session: Missing data techniques

1. Background

The incidence of human immunodeficiency virus (HIV) in Sub-Saharan Africa (SSA) remains high and disproportionate compared to the global average. More specifically, in 2017 a five-fold higher number of new HIV cases was reported in SSA compared to the global average (WHO, 2019). Two-thirds of the new HIV cases among youth occur in adolescent girls and young women (AGYW) aged between 15 and 24 years of age (UNAIDS, 2015, 2016). Four out of five AGYW living with HIV reside in SSA, and they acquire HIV infection on average five to seven years earlier than men (UNAIDS, 2014, 2015). The highest number of people living with HIV reside in South Africa (Shisana, 2014) and AGYW aged between 15- and 24-years old account for about 26% of all new HIV infections, despite that AGYW only make up 10% of the entire population (UNAIDS, 2017).

In general, AGYW are at a higher risk of acquiring HIV compared to young men and older women (Dellar et al., 2015; Harrison, 2014). Social and contextual factor including gender inequality and gender-based violence make them vulnerable for risk-taking behavior such as engagement in unprotected sexual intercourse or substance abuse. Consequently, these factors compound to their vulnerabilities and increase risk of HIV acquisition (Harrison et al., 2016; Arnett, 1992; Tapert et al., 2001). Despite the implementation of intervention programs for AGYW, the effects are suboptimal and the decrease in incidence in AGYW remains slow (UNAIDS, 2015).

An adequate level of risk perception is critical to adopt a protective behavior concerning HIV acquisition. The purpose of this study is to **identify factors associated with HIV risk perception among AGYW**. Here, we will focus on risk perception at baseline and which factors are associated therewith.

2. Data description

The **DREAMS** (Determined, Resilient, Empowered, AIDS-free, Mentored and Safe) initiative, funded by the US President's Emergency Plan for AIDS Relief (PEPFAR) (Saul et al., 2018), aimed at keeping adolescent girls and young women (AGYW) HIV-negative through pre-exposure prophylaxis (PrEP) and secondary distribution of HIV self-test kits to their male sexual partners. The study was conducted between September 18, 2017, and October 31, 2018, at primary healthcare facilities or community outreach programs in Northern Johannesburg, South Africa. Participants were eligible if they were between 16 and 24 years old, tested HIV-negative, had a sexual male partner who was HIV-negative or of unknown HIV status and which were in a heterosexual relationship for at least 3 months.

3. Primary endpoint

The primary endpoint of this analysis was **self-assessed risk of acquiring HIV infection**. At baseline, participants were asked *"What do you think the chance is that you will get HIV in the next six months?",* thereby expressing this on a scale ranging from 0 to 100, with 100 being the maximum value.

Furthermore, next to individual-level characteristics other **potential factors related to HIV risk perception of AGYW** were recorded including socio-economic and demographic variables, overall and sexual health behavior, source of HIV information and relationship characteristics.

1. **How many participants** are included in the study? **How many variables**?
2. What is the **age distribution of individuals** in the DREAMS study?

4. Missing data

For some of the participants, information about the primary endpoint and/or additional covariate information is lacking. Quantify the extent of the missing data in the dataset:

1. For some of the variables, missing observations are coded as 999 instead of NA. First recode missing observations as NA. **NOTE: for the primary endpoint, some observations are equal to 777 which requires recoding to NA as well.**
2. Explore the level of missing data in the dataset by calculating the **frequency of missing observations for each variable**.
3. How many participants have at least one missing observation for the recorded variables?
4. Alternatively, provide a **graphical exploration** of the amount of missing data by variable using the function *aggr* in the R package VIM.

5. Missing data techniques and questions

Let us first focus on the primary endpoint. Study the **marginal distribution of the observed risk perception scores** for the participants in the DREAMS study.

1. Which **missingness mechanisms** do exist? Explain.
2. Formulate an **imputation model for the (pseudo-)continuous endpoint** Z defined as the risk perception score (on a range of 0 – 100). What about the excess number of zero observations?
3. Consider the random variable Y defined as having a risk score equal to zero (no risk of HIV acquisition). Construct an imputation model for Y, under the assumption of missing at random (MAR), and depending on the covariates **age, educational level, residential area,** and **monthly income**. What type of model will you consider?
4. Based on the fitted imputation model (for Y), **predict the missing y-values**, and store the results in a new data frame in R. *Prediction of the probability of a zero risk can be performed using the predict-function in R and prediction of y (being equal to*

zero or one) should be based on an estimated probability of having a zero-risk perception exceeding 0.5.

5. How would you **evaluate the performance of this imputation (or prediction) model**?
6. The procedure in step 5. describes a **single imputation process.** How would you extend the approach to account for the uncertainty related to the imputation step? More specifically, how could you repeatedly and **multiple times** impute the missing observations for the random variable Y? Create 10 different imputation sets based on such an approach.
7. If we impute Y = 1, the resulting imputed value for Z is equal to zero. However, how can we **impute the missing values for our primary endpoint Z**, conditional on Y being equal to 0 (i.e., implying that the risk perception score differs from zero). Formulate an imputation model for Z depending on **age, educational level, residential area,** and **monthly income.** What type of model will you consider?
8. Based on the fitted imputation model (for Z, conditional on Y = 0), **predict the missing z-values**, and store the results in a new data frame in R. *Prediction can be done using the predict-function in R.*
9. The procedure in step 6. (i.e., also referred to as **regression imputation**) describes a **single imputation process.** How would you extend the approach to perform **multiple imputation**?
    a. Consider random generation of error terms (random noise);
    b. Random generation of the asymptotic distribution of the estimators of the model parameters in combination with random noise.
10. Construct **M = 10 imputation** sets and estimate whether there exists a significant difference in average baseline risk perception score among AGYW living in Cosmocity versus those living in Fourways. How would you **combine the results of the different analyses**?
11. Check whether the number of imputations is sufficient. More specifically, create a graph in which you assess the **convergence** of the pooled difference in average baseline risk perception score among AGYW living in Cosmocity versus those living in Fourways depending on the number of imputations.
12. The MICE package in R provides an automated way of performing Multiple Imputation with Chained Equations (Van Buuren, 2011). The MICE approach is also referred to as Full Conditional Specification, meaning that available information on all variables, except for the one to be imputed, will be used in the imputation model to impute missing values. **Use the MICE package to perform multiple imputation for the DREAMS dataset at hand**.
13. Explain the difference between **MICE** and **Multivariate Normal Imputation**? Which method is to be preferred?
14. What is the difference between **predictive mean matching** (pmm) and **regression imputation** considered in the previous steps? What are the advantages and disadvantages of both approaches?

References

J. Arnett, *Reckless behavior in adolescence: A developmental perspective*, Developmental Review, 12(4), 339–373, 1992.

R. Dellar, S. Dlamini and Q. Karim, *Adolescent girls and young women: key populations for HIV epidemic control*, J Int AIDS Soc., 18(1), 2015.

A. Harrison, *HIV Prevention and Research Considerations for Women in Sub-Saharan Africa: Moving toward Biobehavioral Prevention Strategies*, Afr J Reprod Health, 18(3), 17–24, 2014.

A. Harrison, C. Colvin, C. Kuo and et al., *Sustained High HIV Incidence in Young Women in Southern Africa: Social, Behavioral and Structural Factors and Emerging Intervention Approaches*, Curr HIV/AIDS Rep., vol. 12, no. 2, 207–215, 2015.

J. Saul, G. Bachman, S. Allen and et al., *The DREAMS core package of interventions: A comprehensive approach to preventing HIV among adolescent girls and young women*, *PLoS One,* vol. 13, no. 12, 2018.

O. Shisana, T. Rehle, L. Simbayi and et al, *South African National HIV Prevalence, Incidence and Behaviour Survey, 2012*, HSRC Press, Cape Town, 2014.

S. Tapert, G. Aarons, G. Sedlar and S. Brown, *Adolescent substance use and sexual risk-taking behavior*, J Adolesc Health, 28(3), 181-189, 2001.

UNAIDS, *The Gap Report*, Geneva, Switzerland, 2014.

UNAIDS, *Empower Young Women and Adolescent Girls: Fast-tracking the end of the AIDS Epidemic in Africa*, Geneva, Switzerland, 2015.

UNAIDS, *Prevention Gap Report*, Geneva, Switzerland, 2016.

UNAIDS, *UNAIDS Data 2017*, Geneva, Switzerland, 2017.

S. van Buuren, and K., Groothuis-Oudshoorn (2011). *mice: Multivariate Imputation by Chained Equations in R*. Journal of Statistical Software, 45(3), 1–67. https://doi.org/10.18637/jss.v045.i03

World Health Organization, *Number of new HIV infections by WHO region, 2017,* Available at: http://apps.who.int/gho/data/view.main.HIVINCIDENCEREGIONv?lang=en.