COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH EPIDEMIOLOGY

RUTGERS New Jersey Medical School

Development and Standardization of SAS Code to Generate Aggregate Numbers for Biorepository Dashboarding

Pradeeti Mainali, MPH Candidate[1]; Priyal Matreja, MS[2]

[1]Columbia University, Mailman School of Public Health, Department of Epidemiology, [2] Rutgers University, New Jersey Medical School, Department of Medicine-Infectious Disease.

BACKGROUND

Tuberculosis remains a critical and active global health issue, affecting 10.6 million people and causing 1.3 million deaths in just 2022.[1] Tuberculosis remains a difficult infectious disease to tackle due to the occurrence of multi-drug-resistant strains (MDR-TB), co-infection with HIV, diagnostic difficulties, treatment barriers, and more. Continued efforts focus on understanding patient demographics, experience, and outcomes to inform treatment strategies. Additionally, observational studies are needed for TB to understand its complex transmission dynamic, drug resistance patterns, co-infections like HIV, and the effectiveness of interventions across diverse populations. The Department of Medicine–Infectious Diseases of New Jersey Medical School conducts research across international sites (e.g., Brazil, India, Vietnam, South Africa, Uganda, and the Philippines) to analyze TB in populations with high prevalence of disease.

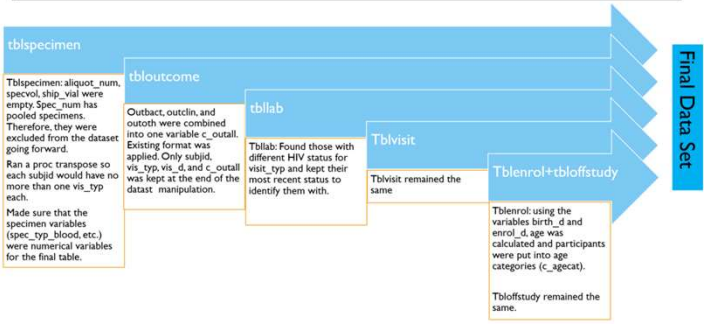
PROJECT DESCRIPTION

As part of a global data harmonization effort, this subproject focused on handling data from various case report forms, merging different datasets to produce interpretable aggregate numbers for key variables. Data harmonization allows for establishment of comparability, greater statistical power and generalizability through a larger sample size, exploring trends, and can be useful to inform/analyze public health interventions.[2] The goal was to create a standardized code, starting with data from Brazil, for consistent analysis of key variables across multiple countries.

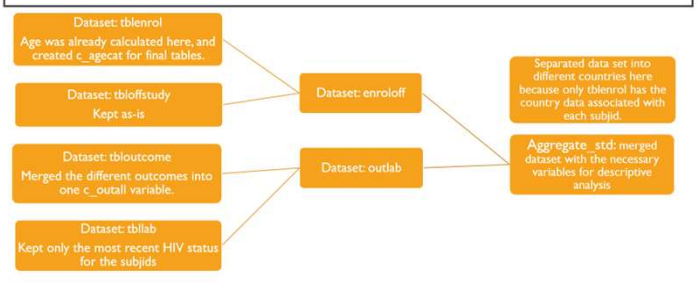
METHODS

The Brazil datasets included 1,188 patients who were enrolled in the study. Initially, there were six datasets provided (each representing a case report form) which had information on enrollment, visit dates/notes, lab tests, TB outcome, specimen collected, and more. Each dataset was coded to include only the PID, visit date, visit type, and descriptive variable of interest (sex, age, HIV status, mortality, treatment outcomes, and specimen types). After gaining a thorough understanding of the variables and their interactions, a complex merge was performed while preventing duplicate data points. PROC TABULATE was used to create a descriptive table. The Brazil code was adapted to create a standardized version for use with data from other countries with the use of SAS macros.

REPORT SITE SPECIFIC DATA: METHODS



STANDARDIZED CODE FOR AGGREGATE TABLES: METHODS



RESULTS AND NEXT STEPS

Both the Brazil and standardized code resulted in a large table that organizes the different combinations of sex, HIV status, age category, mortality, and treatment outcome as rows and specimen type as columns. Both codes also produce descriptive Table 1 statistics through PROC FREQs. The final table, exported as a CSV, will be used as a launchpad for a dashboarding project to visualize the statistics.

REPORT SITE SPECIFIC DATA: DESCRIPTIVE RESULTS TEMPLATE

```
proc freq data=report_br_1; table sex hiv_r c_agecat death_y c_outall; run;
```

Category:	Count:
Total (N):	N
Sex:	
Male	X
Female	X
HIV Status:	
Positive	X
Negative	X
Age Categories:	
<18	X
18-<50	X
>=50	X
Death Status	
Yes	X
No	X

TB Outcome Status:	Count:
Total (N):	N
Bacteriologic cure	X
Bacteriologic status indeterminate	X
Bacteriologic failure	X
Clinical response	X
Clinical relapse	X
Not Tuberculosis	X
Death	X
Treatment incomplete	X
Lost to follow-up/unknown	X
Unknown	X

REPORT SITE SPECIFIC DATA: TABULATE RESULTS TEMPLATE

STUDENT CONTRIBUTION

As the primary programmer, I developed the SAS codes for Brazil Biorepository datasets, merging strategically to ensure data integrity. I coded with SAS using formats, PROC SQL, PROC TABULATE. Additionally, I learned large dataset handling, manual data checks, and SAS macro standardization when creating the standardized code.

EXAMPLES OF HOW I APPLIED COMPETENCIES

Competency	I applied it by. . .
Analyze public health problems in terms of magnitude; person, time, and place; and the distribution and determinants of both chronic and infectious diseases; and principles of disease prevention in different populations	Assisting NJMS DOMC team with data analysis and coding for Tableau dashboards that allowed for exploration of time, place, demographics, samples, and more for different sites internationally. I also analyzed data here and was able to notice a specific pattern previously overlooked in terms of HIV diagnosis with the patients.
Apply appropriate epidemiologic and statistical measures to generate, calculate, and draw valid inferences from public health data.	I helped generate one big cleaned and merged dataset for this project with the Brazil data, which were used for descriptive statistics. This will be used for a bigger global biorepository project by PIs.

REFERENCES

1) Global Tuberculosis Report 2023. World Health Organization. November 7, 2023. Accessed September 27, 2024. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023.
2) Fortier I, Raina P, Van den Heuvel ER, et al. Maelstrom Research guidelines for rigorous retrospective data harmonization. *Int J Epidemiol*. 2017;46(1):103-105. doi:10.1093/ije/dyw075