TB Fuzzy-Matches Made Easy: Designing shareable SAS code to accurately and efficiently identify exact and close genetic matching tuberculosis isolates

Author: Evan Timme, MPH – Arizona Department of Health Services
Wednesday, April 10th, 2019

The Tuberculosis Genotyping Information Management System (TB-GIMS) is a genetic data source available to public health programs. We attempted to design universally sharable SAS code to efficiently analyze TB GIMS data for the purpose of detecting exact and close matching isolates to an identified active TB case of interest (IATCI).

The Arizona TB-GIMS extract contains 2,800+ isolates. Coding was designed in Base SAS-9.4 and created for easy adaptability by another health department. Modifiable macro thresholds set inclusion criteria for matches. Conventional assessments require an exact match on one 15-digit number and two 12-character alphanumeric variables. Genetic changes overtime are common and fuzzy-matches would accommodate such changes. Using a simple do loop and substr function counter, we were able to count the total number of place-value matches in the 15-digit and two 12-character variables for each IATCI to all other isolates. Only fuzzy-matches meeting or exceeding the macro set thresholds are retained. Fuzzy-matches were then assessed for distance [zipcitydistance function] and the time [years between isolate collections]. Output reports include a dot-plot (time x-axis, distance y-axis), a single .XML with unique sheets for each IATCI, and a single .PDF with individual reports for each IATCI. When combined with epidemiologic data, the outputs were helpful for understanding potential transmission clustering.

The utility of this code is helpful and complementary to epidemiologic-linking data. User defied thresholds allow for simple and easy code adaption, while being an efficient and effective use of staff time and resources. These findings are encouraging and warrant further exploration.