Evaluating the Mechanisms for Missing Data in the Enhanced Tuberculosis Surveillance System

Sam Abbott, Hannah Christensen, Ellen Brooks-Pollock

**Authors:**

Sam Abbott, Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK

Hannah Christensen, Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK

Ellen Brooks-Pollock, Bristol Medical School: Population Health Sciences, University of Bristol, Bristol, UK

**Correspondence to:** Sam Abbott, Bristol Medical School: Population Health Sciences, University of Bristol, Bristol BS8 2BN, UK; [sam.abbott@bristol.ac.uk](mailto:sam.abbott@bristol.ac.uk); 01173310185

**Words:** *Title:* *Abstract:* *Paper:*

# Abstract

## Background

The Enhanced Tuberculosis Surveillance (ETS) system is a routine surveillance system - with a similar structure to other such systems - that collects data on all notified tuberculosis (TB) cases in England. It is routinely used to study the epidemiology of TB. Routine data often has a large amount of missing data which may not be fully accounted for when used in analyses. This study explores the evidence for associations between missingness in several key outcomes and demographic variables. Any such associations may introduce bias if not accounted for.

## Methods

* Introduce ETS
* Data extraction and management
* Structure of the ETS
* Data completeness
* Drivers of variable completeness (regression)

## Results *Copy from bottom*

* Missing structure
* Drivers of variable completeness

## Conclusions

* Surveillance data is likely to have a high degree of misising data. In the ETS missing for key outcomes is associated with demographic factors such as….
* To avoid biasing analysis studies should make use of imputed data - rather than complete case analysis - and extend their imputation models to other demographic variables that may not be included in the analysis model.
* This analysis should be repeated in other datasets - for this reason the code is available as an R package (<https://doi.org/10.5281/zenodo.3492200>).

# Introduction

*Background*

The Enhanced Tuberculosis Surveillance (ETS) system is a routine surveillance system - with a similar structure to other such systems - that collects data on all notified tuberculosis (TB) cases in England. It is routinely used to study the epidemiology of TB. Routine data often has a large amount of missing data which may not be fully accounted for when used in analyses.

*Detail*

Missing data can take several forms, data that are missing completely at random (MCAR), data that are missing at random (MAR) and data that are missing not at random (MNAR).[1] Data that are MAR are missing with a mechanism that is conditional on observed variables, whilst MNAR are missing with a mechanism that is conditional on variables that are not observed. Data that is MAR, and MNAR may lead to biases when analysing the data, however it is not possible to deduce from the observed data what the mechanism driving missing data is. Therefore, it is necessary to account for these potential biases during the analysis stage. This is possible using a variety of methods such as scenario analysis accounting for the ‘best’ and ‘worst’ case scenarios, and multiple imputation of missing data using additional variables in the dataset to inform the imputation model.[1] Common practise is to include all variables included in the analyses in the imputation model, these variables may or may not be those at most risk of introducing bias due to an MAR mechanism.

*Aim*

This study aims to explore the evidence for associations between missingness in several key outcomes and demographic variables. Any such associations may introduce bias if not accounted for.

# Methods

## Enhanced tuberculosis surveillance (ETS) system

Table 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | Missing (N) | Missing (%) | Missing (N) | Missing (%) |
| natquintile | 63175 | 100.0 | 8120 | 15.7 |
| bcgvaccyr | 62479 | 98.9 | 31421 | 60.8 |
| bcgvacc | 61916 | 98.0 | 17133 | 33.2 |
| datediag | 45557 | 72.1 | 10303 | 19.9 |
| sputsmear | 32912 | 52.1 | 32094 | 62.1 |
| timesinceent | 29084 | 46.0 | 18670 | 36.2 |
| anyres | 27485 | 43.5 | 20995 | 40.7 |
| occat | 24870 | 39.4 | 5513 | 10.7 |
| symptonset | 23937 | 37.9 | 12829 | 24.8 |
| txenddate | 18711 | 29.6 | 1137 | 2.2 |
| prevdiag | 13204 | 20.9 | 3148 | 6.1 |
| starttreatdate | 9151 | 14.5 | 2127 | 4.1 |
| tomdeathrelat | 7539 | 11.9 | 1191 | 2.3 |
| ukborn | 6230 | 9.9 | 1825 | 3.5 |
| overalloutcome | 6044 | 9.6 | 0 | 0.0 |
| startedtreat | 4242 | 6.7 | 602 | 1.2 |
| ethgrp | 2811 | 4.4 | 1229 | 2.4 |
| dateofdeath | 1235 | 2.0 | 357 | 0.7 |
| pulmextrapulm | 177 | 0.3 | 213 | 0.4 |
| sex | 101 | 0.2 | 110 | 0.2 |
| phec | 32 | 0.1 | 0 | 0.0 |
| age | 25 | 0.0 | 0 | 0.0 |
| caserepdate | 0 | 0.0 | 0 | 0.0 |
| year | 0 | 0.0 | 0 | 0.0 |
| culture | 0 | 0.0 | 0 | 0.0 |