Explore and Evaluate 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

## Objectives

## Design

## Setting

## Outcome measures

## Results *Copy from bottom*

## Conclusions

# Strengths and limitations of this study

# Introduction

Although the characteristics of tuberculosis (TB) in England have been reported elsewhere,[1,2] and key risk factors such as non-UK birth status have been identified,[3] little attention has been given to the role of BCG vaccination. In particular, there is little information available regarding the demographics of vaccinated versus unvaccinated cases and the impact of BCG vaccination on TB outcomes in England has not been explored. There has also only been limited reporting of the age distribution, and trends over time, in incidence rates stratified by UK birth status.

In this chapter I explore the epidemiology of TB in England using routine datasets with a particular focus on the impact of missing data, the mechanisms underlying that missing data, seasonal trends, the role of age, UK birth status and BCG status. I also estimate incidence rates, stratified by UK birth status and age, which I use to identify trends in TB incidence over time. Finally I report TB outcomes in England using case rates, again stratified by BCG status and UK birth status.

# Methods

## Data sources

### Enhanced tuberculosis surveillance (ETS) system

The ETS is a database that collects demographic, clinical, and microbiological data on all notified TB cases in England and is maintained by Public Health England (PHE). Notification is required by law, with health service providers having to inform PHE of all confirmed TB cases.[1] Data collection began in 2000 and was expanded, with additional variables, with the launch of a web based system in 2008.[4] It is updated annually with de-notifications, late notifications and other updates. A descriptive analysis of TB epidemiology in England is published each year, which reports on data collection, cleaning, and trends in TB incidence at both a national, and sub-national level.[1]

#### Data extraction and management

Data on all notifications (114,820 notifications) from the ETS system from 2000 to 2015 were obtained from PHE via an application to the TB monitoring team. Data fields included: notification date, age, PHE centre, occupation, ethnic group, UK birth status, years since entry to the UK, date of symptom onset, date of presentation, date of diagnosis, date of treatment start, date of treatment end, date of death, Pulmonary TB status, culture status, sputum smear status, drug resistance, BCG vaccination status, year of vaccination, outcome at 12 months, overall outcome, and cause of death. Notifications were assessed for identifiability and the data release was conditional on the raw data not being shared further. Invalid entries were replaced with missing values unless otherwise noted, with character variables stored as factors using their most common entry as the baseline. Notifications from Scotland, Northern Ireland and Wales were dropped from the dataset. Several variables were created, or modified, for use in further analysis, Table 1summarises these variables. The code used for data cleaning is available as an R package[[1]](#footnote-34).

Table 1: Variables derived or modified from the ETS system for use in the analyses throughout this thesis.

|  |  |
| --- | --- |
| Created/modified variable | Description |
| Years since BCG | Derived using year of vaccination and year of notification. Categorised into and 11+ due to the evidence of waning protection for the BCG vaccine.[5] |
| Age at BCG | Derived using year of vaccination and age at vaccination. Categorised into < 1, 1 to x < 12, 12 to x < 16 and to capture historic vaccination policy.[6] |
| Successful treatment | For cases that had a recorded date of starting treatment, with their outcome recorded at the latest available follow up. Those that completed treatment are defined as successfully treated: treatment failure is defined as those that stopped treatment, were lost to follow up, those that died during follow up from TB, those that died during follow up were TB contributed to their death, and those who were still on treatment. Those that were not evaluated were treated as missing. |
| Mortality | Assessed via follow up at 12 and 24 months: mortality is defined as cases with an overall outcome of death, and survival is defined as those that completed treatment, were still on treatment, and stopped treatment. Those that were lost to follow up, or not evaluated were treated as missing |
| TB mortality | For cases with an overall outcome of died, and whose cause of death was known to be TB or to be related to TB. Those that were known to have not died, or who were known to have died from a cause other than from TB were defined to have not died from TB. |
| Death due to TB | Death due to TB is defined as those that died directly from TB, or where TB had contributed to their death with death not due to TB being cases that died from any other cause. Conditioned on all-cause mortality, for cases with a known cause of death. |

#### Structure of the ETS

The ETS is in a wide format with each notification having a single row, and with each unique variable having a single column. This structure means that the progression of TB in each individual is captured by a series of dates rather than as a series of events. As notifications are not linked to a unique patient I.D it is possible that individuals are duplicated within the ETS, with multiple notifications. These recurrent notifications have been flagged within the data extract by the TB section at PHE. The majority of variables are factors, with a significant minority of numeric and date variables.

#### Data completeness

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).[7] 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.[7]

As the ETS is aggregated across England, from a variety of sources, some level of missing data are inevitable. This takes two forms: under-reporting of notified cases, of which there is some evidence in the literature,[8] and data missing for a notified case. The former is particularly problematic as apart from using comparative studies the characteristics of those that are not notified is unknown. For variables that are missing data within the dataset it is possible to calculate the proportion of missing data (Figure 1, Table 2) but care must be taken to account for nested variables such as date of death and year of BCG vaccination. This can be done by assuming that the nested variables takes the value of the top level variable when it is known that the variable is not truly missing. An example of this is using overall outcome for date of death when notifications are known to have not died. Doing this shows high completeness for common demographic variables such as sex, age, ethnic group and UK birth status. More problematically, BCG status and year of BCG status have a high percentage missing, even after accounting for the introduction of national collection of these variables in 2008. Socio-economic status (as national quintiles) was not collected until 2010 but after this point is highly complete. Comparing pre 2009 and post 2008 in Table 2 (and by inspecting Figure 1) there are also issues of changing completeness over time,[1,9] if this is not accounted for than it may lead to spurious trends. Figure 1 also indicates that there are multiple groups of variables that share a common pattern of missing data.

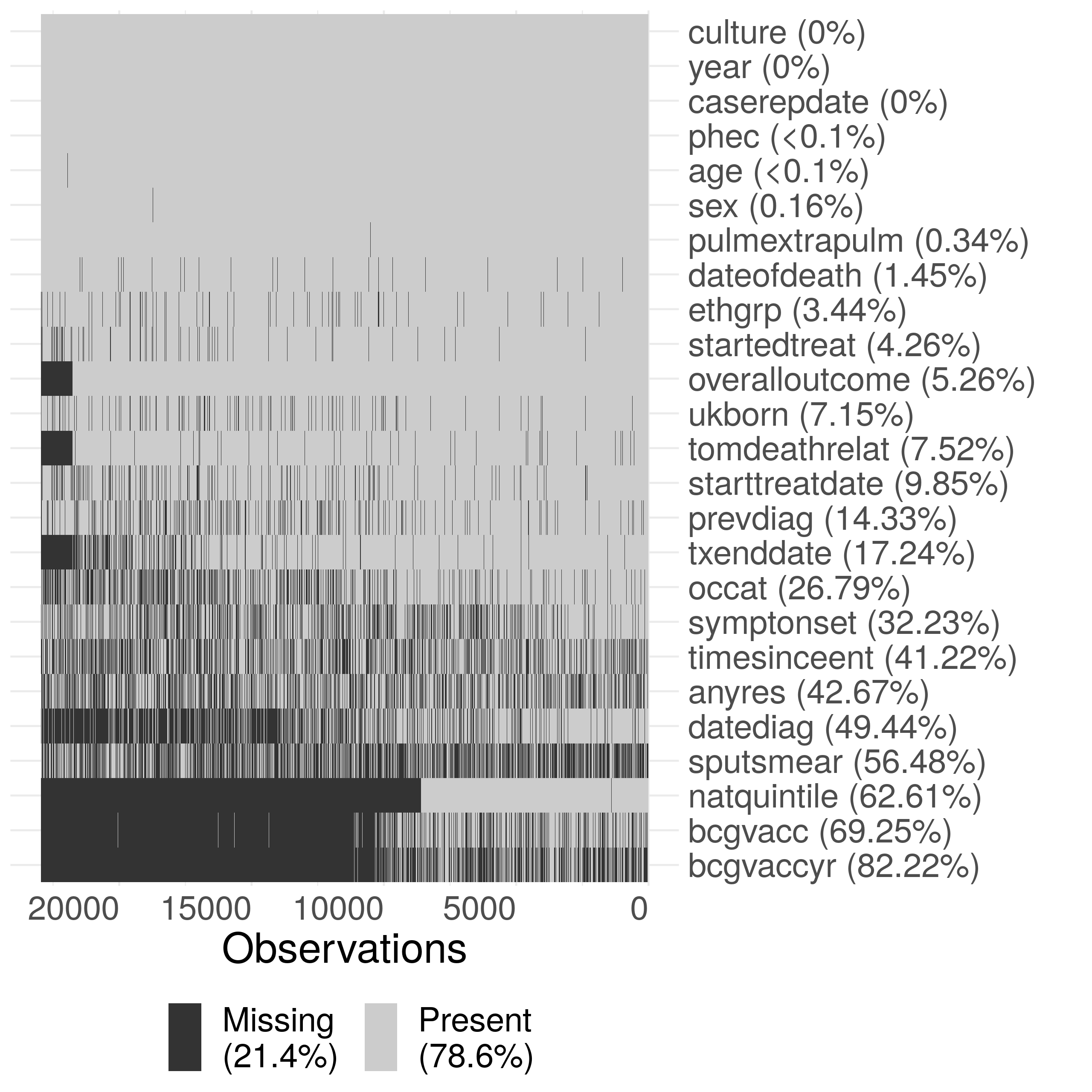


Figure 1: Summary plot of missing data in the extract of the ETS data used in this thesis. Due to the large size of the dataset, the data has been sub-sampled with only 20% of the data shown in this figure. Notifications have been ordered by date of notification from left to right. The following subset of variables are shown: year (year), sex (sex), age (age), PHE Centre (phec), Occupation (occat), Ethnic group (ethgrp), UK birth status (ukborn), Time since entry (timesinceent), date of symptom onset (symptonset), date of diagnosis (datediag), started treatment (startedtreat), date of starting treatment (starttreatdate), treatment end date (txenddate), pulmonary or extra-pulmonary TB (pulmextrapulm), culture (culture), sputum smear status (sputsmear), drug resistance (anyres), previous diagnosis (prevdiag), BCG status(bcgvacc), Year of BCG vaccination (bcgvaccyr), overall outcome (overalloutcome), cause of death (tomdeathrelate), socio-economic status quintiles (natquintile), and date of death (dateofdeath). Nested variables have been accounted for (i.e date of death has had an entry added for cases that are known to have not died), so that true missingness for all variables is estimated.

Table 2: Breakdown of missing data from the ETS prior to the web based system (pre 2009) and post (post 2008) by variable, ordered by the percentage missing for a subset of variables. The following subset of variables are shown year (year), sex (sex), age (age), PHE Centre (phec), Occupation (occat), Ethnic group (ethgrp), UK birth status (ukborn), Time since entry (timesinceent), date of symptom onset (symptonset), date of diagnosis (datediag), started treatment (startedtreat), date of starting treatment (starttreatdate), treatment end date (txenddate), pulmonary or extra-pulmonary TB (pulmextrapulm), culture (culture), sputum smear status (sputsmear), drug resistance (anyres), previous diagnosis (prevdiag), BCG status(bcgvacc), Year of BCG vaccination (bcgvaccyr), overall outcome (overalloutcome), cause of death (tomdeathrelate), socio-economic status quintiles (natquintile), and date of death (dateofdeath). Nested variables have been accounted for (i.e data of death has had an entry added for cases that are known to have not died), so that true missingness for all variables is estimated.

Pre 2009

Post 2008

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

For nested variables with rare outcomes assuming the top level variable value can mask the underlying amount of missing data. An alternative approach is to filter the data for the top level variable required for the nested variable to be defined and to then compute the proportion of these notifications that are missing data for the outcome of interest. For the date of starting treatment this approach leads to an estimate of 5.9% (6434/108410) being missing, which is more complete than previously estimated. For cases that are known to have completed treatment 16.5% (13804/83891) are missing a date for the end of treatment. In notifications that are known to have died, 26.6% (1592/5976) were missing the date of death and 44.9% (2686/5976) were missing the cause of death. In any analysis where these variables are used the missing data for these variables will need to be carefully adjusted for. In particular, if cause of death is used it must be clearly stated that it is highly missing and results based on this variable should be properly caveated.

#### Drivers of Variable completeness

Missing data may be MAR or MNAR, which may introduce biases into any analyses based on these data. This is of particular importance for variables that have high levels of missingness, as any introduced bias is likely to have a greater impact on the overall results, and for variables that are used extensively in analyses later in this thesis. Unfortunately MNAR data cannot be detected, so bias from this source cannot be discounted. However, it is possible to detect potential MAR mechanisms from observed variables that would not necessarily be included in a model used for analysis. Although these associations may themselves be caused by an external factor. In the following section I explore variables associated with data being missing for several key variables including: BCG status, year of BCG vaccination, date of death, cause of death, date of symptom onset, date of diagnosis, date of starting treatment and date of ending treatment. All of these variables were shown to have high levels of missing data in the previous section and they will all be used extensively throughout this thesis.

In order to explore the drivers of missing data I have reformulated the problem as a logistic regression for each variable of interest, with the outcome being data completeness (complete/missing). This allows variables that are hypothesised to be related to missing data to be adjusted for and their independent impact on data completeness to be estimated. Unlike classic approaches to missing data, such as multiple imputation by chained regression (MICE),[10] this is not an imputation. The details of the approach are discussed below.

##### Statistical analysis

In order to reformulate missing data as a logistic regression we took the following steps:

1. For the variable of interest create a new temporary binary variable, called data status, that is “Missing” when the variable of interest is missing and “Complete” when it is not. Specify “Complete” as the baseline.
2. For nested variables exclude notifications that do not have the top level outcome required by the variable of interest. An example of this is excluding cases that did not die, or have a missing overall outcome, when investigating TB mortality.
3. Specify the hypothesised drivers of missingness for the variable of interest. These should be variables with a reasonable hypothesis for how they would drive missingness in the variable of interest. They must also be relatively complete as this approach does not impute missing confounder data.
4. Fit a logistic regression model with the temporary data status variable as the outcome, adjusting for the hypothesised drivers of missingness.
5. Exponentiate the returned coefficients, and confidence intervals so that they represent Odds Ratios (ORs).
6. Refit the model, dropping each variable in turn and then comparing the updated model with the full model using a likelihood ratio test.
7. Interpret the results, using the estimated size of the effect, the width of the confidence intervals and the size of the Wald and likelihood ratio test p values to determine which variables are related to missingness for the variable of interest. Evidence should be interpreted on a spectrum, rather than using arbitrary significance cut-offs.[11] To avoid issues of multiple testing the level of evidence should be weighted based on the number of variables adjusted for and the number of outcomes explored.

For all outcomes considered I adjusted for the same set of demographic variables that were both highly complete and also plausibly linked to missingness for all outcomes considered. These were: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Complete case analysis has been used, with the dataset limited to notifications from 2010 and on-wards as socio-economic status was not collected prior to this.

## Patient and public involvement

We did not involve patients or the public in the design or planning of this study.

# Results

## BCG status

It is clear that BCG status is missing with a MAR mechanism for the variables considered Table 3. BCG data missingness is strongly associated with year of notification, sex age, ethnic group, and socio-economic status. It appears that after adjusting for other variables data completeness increased from 2010 until 2012 but has since showed no clear trend. Men appear to be more likely than women to have a missing BCG status, with the non-UK born also being more likely than the UK born to be missing BCG status. The proportion of those missing BCG status increases with age, with those aged 65+ being over 4 times more likely to be missing BCG status than those aged 0-14 years old. There is also evidence to suggest that notifications in the lowest socio-economic group are more likely to have a missing BCG status but there was no clear evidence of a trend across socio-economic quintiles. The White ethnic group was more likely to have a missing BCG status than any other ethnic group.

Table 3: Results from a logistic regression model with data completeness (Complete/Missing) for BCG vaccination as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that BCG status is missing at random for the variables considered.

Variable

Category

Missing (N)

Notifications (41659)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

31.3% (2235)

7143

1.27e-08

2011

29.8% (2319)

7781

0.93 (0.87, 1.00)

0.0606

2012

27.9% (2164)

7755

0.85 (0.79, 0.91)

6.32e-06

2013

27.1% (1907)

7034

0.79 (0.74, 0.86)

1.01e-09

2014

30.1% (1907)

6327

0.91 (0.85, 0.98)

0.017

2015

29.7% (1668)

5619

0.89 (0.82, 0.96)

0.00348

Sex

Female

27.4% (4847)

17664

8.74e-11

Male

30.6% (7353)

23995

1.16 (1.11, 1.21)

9.49e-11

Age

0-14

13.1% (235)

1793

1.67e-157

15-44

26.0% (6557)

25235

2.10 (1.82, 2.43)

5.83e-24

45-64

32.8% (2964)

9026

2.84 (2.45, 3.30)

3.16e-43

65+

43.6% (2444)

5605

4.42 (3.80, 5.15)

1.43e-81

Ethnic group

White

35.4% (2959)

8359

2.15e-41

Black-Caribbean

24.6% (228)

928

0.62 (0.52, 0.72)

3.96e-09

Black-African

27.3% (1966)

7204

0.73 (0.67, 0.80)

1.34e-12

Black-Other

24.1% (89)

369

0.65 (0.51, 0.83)

0.000717

Indian

25.9% (2805)

10848

0.62 (0.58, 0.68)

5.2e-31

Pakistani

33.2% (2258)

6806

0.89 (0.82, 0.97)

0.00569

Bangladeshi

27.9% (469)

1680

0.71 (0.62, 0.80)

1.03e-07

Chinese

33.6% (166)

494

0.88 (0.72, 1.07)

0.202

Mixed / Other

25.3% (1260)

4971

0.65 (0.59, 0.71)

4.93e-20

UK birth status

Non-UK Born

29.5% (9104)

30880

7.2e-18

UK Born

28.7% (3096)

10779

0.75 (0.70, 0.80)

1.3e-17

Socio-economic status

1

30.7% (4948)

16131

4.88e-08

2

26.8% (3383)

12621

0.84 (0.80, 0.89)

3.58e-10

3

29.2% (1905)

6530

0.92 (0.86, 0.98)

0.0117

4

30.1% (1142)

3796

0.91 (0.84, 0.99)

0.0264

5

31.8% (822)

2581

0.94 (0.85, 1.03)

0.174

## Year of BCG vaccination

As for BCG status, year of BCG vaccination is also clearly missing with MAR mechanisms for the variables considered (Table 4). As for BCG status men were more likely to have a missing year of BCG vaccination as were the non-UK born. Older notifications were again more likely to have missing data, with those aged 65+ being more than 2 times more likely to have a missing year of vaccination. However, unlike BCG vaccination status, year of notification shows a clear trend of increasing data completeness from 2010 until 2015. Additionally, for year of BCG vaccination the White ethnic group is more likely to have complete data than any other ethnic group, with those of Black-Caribbean descent being over 3 times more likely to have a missing year of BCG vaccination. Socio-economic status is highly associated with year of vaccination being missing but there is little clear evidence of a trend. The second, and third, poorest quintiles were more likely to have a missing year of vaccination. Whilst the richest, and second richest quintiles were less likely to have a missing year of vaccination.

## Date of death

For date of death there is some evidence that data is missing with an MAR mechanism for ethnic group and socio-economic status, with little evidence for any other association (Table 5). These associations should be interpreted carefully due to the strength of the evidence when compared to the number of tests conducted, there is a high likelihood of a type 1 error. Whilst the confidence intervals were wide for all ethnic groups there was some weak indication that the White ethnic group were more likely to have a complete date of death than other ethnic groups. Similarly, those in the lowest socio-economic group were somewhat more likely to have a complete date of death than other quintiles. The reduction in the levels of evidence found for case of death may be linked to the reduction in power for this outcome, as mortality is a rare outcome.

Table 5: Results from a logistic regression model with data completeness (Complete/Missing) for date of death as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that there is some evidence that date of death is missing at random for ethnic group, with weaker evidence for all other variables.

Variable

Category

Missing (N)

Notifications (1883)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

16.6% (53)

320

0.0876

2011

15.9% (52)

327

0.95 (0.62, 1.46)

0.818

2012

14.5% (51)

351

0.81 (0.53, 1.25)

0.342

2013

13.5% (42)

312

0.73 (0.46, 1.14)

0.163

2014

9.5% (30)

317

0.52 (0.32, 0.84)

0.0081

2015

13.3% (34)

256

0.69 (0.43, 1.11)

0.133

Sex

Female

14.8% (97)

657

0.609

Male

13.5% (165)

1226

0.93 (0.70, 1.23)

0.608

Age

0-14

10.0% (1)

10

0.929

15-44

15.7% (31)

198

1.90 (0.32, 36.43)

0.556

45-64

14.6% (68)

465

1.92 (0.33, 36.42)

0.549

65+

13.4% (162)

1210

1.95 (0.34, 37.04)

0.536

Ethnic group

White

11.1% (102)

920

0.00373

Black-Caribbean

21.7% (10)

46

1.58 (0.67, 3.51)

0.274

Black-African

20.1% (27)

134

1.49 (0.76, 2.94)

0.251

Black-Other

20.0% (1)

5

1.59 (0.08, 11.72)

0.687

Indian

17.4% (64)

367

1.08 (0.62, 1.92)

0.789

Pakistani

8.0% (20)

249

0.50 (0.25, 0.99)

0.0483

Bangladeshi

22.7% (10)

44

1.65 (0.67, 3.87)

0.261

Chinese

14.3% (3)

21

0.89 (0.19, 3.00)

0.864

Mixed / Other

25.8% (25)

97

1.99 (1.01, 3.92)

0.0462

UK birth status

Non-UK Born

16.6% (167)

1004

0.133

UK Born

10.8% (95)

879

0.67 (0.40, 1.14)

0.128

Socio-economic status

1

11.4% (79)

695

0.0265

2

18.3% (86)

470

1.67 (1.19, 2.35)

0.0033

3

16.2% (48)

296

1.49 (0.99, 2.22)

0.0548

4

12.7% (30)

237

1.21 (0.75, 1.90)

0.429

5

10.3% (19)

185

0.95 (0.54, 1.62)

0.866

## Cause of death

For cause of death there is less evidence of an MAR mechanism, with little evidence of an association for year, sex, age, or socio-economic group (Table 6). There was, however, strong evidence of an association with ethnic group and very weak evidence of an association with UK birth status. The White ethnic group was less likely to have an incomplete cause of death when compared to the majority of other identified ethnic groups but there was evidence to suggest that cause of death was more likely to be missing in those identifying as being of Black-Caribbean, Black-Other, Indian and Bangladeshi descent. The confidence intervals for these estimates were wide, indicating that these estimates may not be reliable. There was again some weak evidence to suggest that the UK born were more likely to be missing a cause of death than the non-UK born, which reverses the trend observed in the other variables explored. The reduction in the levels of evidence found for case of death may be linked to the reduction in power for this outcome, as mortality is a rare outcome.

Table 6: Results from a logistic regression model with data completeness (Complete/Missing) for cause of death as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that cause of death is missing at random for ethnic group and UK birth status, with little evidence for any other variables

Variable

Category

Missing (N)

Notifications (1883)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

45.0% (144)

320

0.724

2011

45.6% (149)

327

1.03 (0.75, 1.41)

0.85

2012

45.3% (159)

351

1.02 (0.75, 1.39)

0.905

2013

43.9% (137)

312

0.99 (0.72, 1.37)

0.954

2014

44.8% (142)

317

0.96 (0.70, 1.32)

0.793

2015

38.7% (99)

256

0.80 (0.57, 1.12)

0.196

Sex

Female

44.7% (294)

657

0.628

Male

43.7% (536)

1226

0.95 (0.78, 1.16)

0.628

Age

0-14

50.0% (5)

10

0.116

15-44

35.4% (70)

198

0.64 (0.17, 2.48)

0.509

45-64

43.0% (200)

465

0.90 (0.24, 3.44)

0.874

65+

45.9% (555)

1210

0.96 (0.25, 3.67)

0.957

Ethnic group

White

48.2% (443)

920

0.000704

Black-Caribbean

21.7% (10)

46

0.40 (0.18, 0.82)

0.0173

Black-African

45.5% (61)

134

1.41 (0.85, 2.36)

0.183

Black-Other

20.0% (1)

5

0.41 (0.02, 2.87)

0.428

Indian

35.7% (131)

367

0.83 (0.55, 1.27)

0.388

Pakistani

49.4% (123)

249

1.47 (0.95, 2.29)

0.0857

Bangladeshi

27.3% (12)

44

0.60 (0.27, 1.26)

0.189

Chinese

52.4% (11)

21

1.64 (0.64, 4.23)

0.302

Mixed / Other

39.2% (38)

97

1.00 (0.58, 1.72)

0.991

UK birth status

Non-UK Born

40.1% (403)

1004

0.072

UK Born

48.6% (427)

879

1.41 (0.97, 2.07)

0.073

Socio-economic status

1

43.7% (304)

695

0.345

2

40.0% (188)

470

0.93 (0.72, 1.18)

0.54

3

42.9% (127)

296

0.98 (0.74, 1.31)

0.916

4

49.8% (118)

237

1.24 (0.91, 1.69)

0.172

5

50.3% (93)

185

1.21 (0.86, 1.71)

0.262

## Date of symptom onset

For date of symptom onset there was strong evidence of an MAR mechanism for all variables considered, except for sex (Table 7). As found previously, the likelihood of date of symptom onset being missing reduced with year of notification. Children (0-14 years old) were more likely to have a missing date of symptom onset than any other age group as were those in any socio-economic quintile when compared to the poorest group. UK born cases were more likely to have a complete date of symptom onset than non-UK born cases, with the White ethnic group being more likely to have a missing date of symptom onset than most other ethnic groups.

Table 7: Results from a logistic regression model with data completeness (Complete/Missing) for date of symptom onset as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that date of symptom onset is missing not at random for the variables for all variables considered, except for sex.

Variable

Category

Missing (N)

Notifications (41659)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

34.0% (2426)

7143

0

2011

30.1% (2339)

7781

0.83 (0.78, 0.89)

2.03e-07

2012

24.2% (1878)

7755

0.61 (0.57, 0.66)

2.24e-40

2013

17.5% (1233)

7034

0.41 (0.38, 0.44)

8.22e-108

2014

11.8% (744)

6327

0.25 (0.23, 0.28)

1.52e-188

2015

6.9% (390)

5619

0.14 (0.13, 0.16)

8.88e-245

Sex

Female

22.0% (3894)

17664

0.93

Male

21.3% (5116)

23995

1.00 (0.95, 1.05)

0.93

Age

0-14

38.1% (684)

1793

3.59e-73

15-44

20.5% (5182)

25235

0.35 (0.31, 0.39)

6.74e-77

45-64

20.7% (1870)

9026

0.37 (0.33, 0.42)

2.18e-58

65+

22.7% (1274)

5605

0.43 (0.38, 0.49)

2.31e-39

Ethnic group

White

20.9% (1749)

8359

3.98e-09

Black-Caribbean

23.1% (214)

928

1.04 (0.88, 1.23)

0.658

Black-African

23.0% (1654)

7204

0.89 (0.80, 0.98)

0.0179

Black-Other

18.7% (69)

369

0.79 (0.60, 1.04)

0.106

Indian

22.2% (2404)

10848

0.86 (0.79, 0.94)

0.00119

Pakistani

19.2% (1305)

6806

0.75 (0.68, 0.83)

5.56e-09

Bangladeshi

23.9% (401)

1680

1.05 (0.91, 1.20)

0.524

Chinese

18.8% (93)

494

0.74 (0.58, 0.94)

0.016

Mixed / Other

22.6% (1121)

4971

0.93 (0.83, 1.03)

0.152

UK birth status

Non-UK Born

21.9% (6774)

30880

5.44e-12

UK Born

20.7% (2236)

10779

0.77 (0.71, 0.83)

7.6e-12

Socio-economic status

1

19.9% (3218)

16131

5e-17

2

22.9% (2888)

12621

1.22 (1.15, 1.29)

9.51e-11

3

24.2% (1578)

6530

1.33 (1.24, 1.43)

5.79e-15

4

22.0% (837)

3796

1.20 (1.09, 1.31)

8.72e-05

5

18.9% (489)

2581

1.00 (0.89, 1.12)

0.991

## Date of diagnosis

For date of diagnosis there was again strong evidence for an MAR mechanism for all variables considered, except for sex for which there was very weak evidence (Table 8). Increasing completeness was found for year of notification as seen previously, as was an increased likelihood of missing data in males and the non-UK born. The White ethnic group was less likely to be missing data on the data of diagnosis as compared to the majority of other ethnic groups, as were the poorest socio-economic group compared to all other socio-economic quintiles. Children (0-14 years old) were again more likely to be missing data than adults in any age group.

Table 8: Results from a logistic regression model with data completeness (Complete/Missing) for date of diagnosis onset as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. The model indicates that date of diagnosis is missing at random for the variables for all variables considered, except for sex.

Variable

Category

Missing (N)

Notifications (41659)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

26.9% (1918)

7143

1.65e-283

2011

22.3% (1736)

7781

0.78 (0.72, 0.84)

6.72e-11

2012

18.8% (1458)

7755

0.63 (0.58, 0.68)

4.31e-31

2013

12.9% (909)

7034

0.41 (0.37, 0.44)

1.46e-89

2014

10.4% (659)

6327

0.32 (0.29, 0.35)

1.05e-118

2015

7.4% (415)

5619

0.22 (0.19, 0.24)

1.63e-154

Sex

Female

16.9% (2984)

17664

0.0296

Male

17.1% (4111)

23995

1.06 (1.01, 1.12)

0.0298

Age

0-14

19.4% (348)

1793

0.000164

15-44

17.8% (4504)

25235

0.76 (0.67, 0.87)

3.85e-05

45-64

15.9% (1434)

9026

0.73 (0.64, 0.84)

1.54e-05

65+

14.4% (809)

5605

0.72 (0.62, 0.84)

1.84e-05

Ethnic group

White

12.5% (1043)

8359

2.91e-67

Black-Caribbean

25.2% (234)

928

2.21 (1.87, 2.61)

2.58e-20

Black-African

21.9% (1577)

7204

1.49 (1.34, 1.66)

5.45e-13

Black-Other

17.9% (66)

369

1.32 (0.98, 1.74)

0.0587

Indian

18.0% (1957)

10848

1.09 (0.99, 1.21)

0.0858

Pakistani

11.8% (805)

6806

0.75 (0.67, 0.84)

8.98e-07

Bangladeshi

21.5% (361)

1680

1.57 (1.35, 1.82)

2.63e-09

Chinese

13.4% (66)

494

0.82 (0.61, 1.07)

0.153

Mixed / Other

19.8% (986)

4971

1.32 (1.18, 1.48)

1.84e-06

UK birth status

Non-UK Born

18.4% (5696)

30880

6.07e-16

UK Born

13.0% (1399)

10779

0.71 (0.65, 0.77)

1.23e-15

Socio-economic status

1

14.4% (2317)

16131

1.05e-45

2

19.6% (2469)

12621

1.48 (1.39, 1.58)

1.7e-32

3

20.3% (1325)

6530

1.62 (1.50, 1.75)

8.07e-34

4

17.0% (645)

3796

1.37 (1.24, 1.52)

4.77e-10

5

13.1% (339)

2581

1.07 (0.94, 1.21)

0.313

## Date of starting treatment

For date of starting treatment there is little evidence that missing data is associated with any variable considered, except for year of notification (Table 9). Variable completeness improved year-on-year, with a 96% drop in missing data in 2015 compared to 2010.

Table 9: Results from a logistic regression model with data completeness (Complete/Missing) for date of starting treatment as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. There is little evidence that the missing data for the date of starting treatment is associated with any variable considered, except for year of notification.

Variable

Category

Missing (N)

Notifications (40977)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

3.5% (244)

7020

2.4e-70

2011

3.2% (242)

7655

0.91 (0.76, 1.08)

0.281

2012

2.5% (187)

7628

0.69 (0.57, 0.84)

0.000211

2013

2.2% (154)

6923

0.63 (0.51, 0.77)

8.29e-06

2014

0.8% (51)

6239

0.23 (0.17, 0.31)

1.29e-21

2015

0.1% (8)

5512

0.04 (0.02, 0.08)

3.62e-19

Sex

Female

2.2% (383)

17439

0.83

Male

2.1% (503)

23538

0.99 (0.86, 1.13)

0.83

Age

0-14

3.0% (54)

1783

0.157

15-44

2.2% (539)

25000

0.72 (0.53, 0.98)

0.0303

45-64

2.0% (180)

8896

0.68 (0.49, 0.95)

0.0209

65+

2.1% (113)

5298

0.69 (0.49, 0.99)

0.042

Ethnic group

White

2.3% (182)

8055

0.423

Black-Caribbean

2.2% (20)

916

0.89 (0.54, 1.39)

0.626

Black-African

1.9% (139)

7140

0.73 (0.55, 0.96)

0.025

Black-Other

3.0% (11)

368

1.33 (0.67, 2.38)

0.379

Indian

2.1% (230)

10707

0.86 (0.67, 1.10)

0.23

Pakistani

2.4% (158)

6721

0.92 (0.72, 1.19)

0.536

Bangladeshi

2.2% (37)

1665

0.88 (0.59, 1.29)

0.526

Chinese

1.7% (8)

483

0.68 (0.30, 1.33)

0.307

Mixed / Other

2.1% (101)

4922

0.86 (0.64, 1.15)

0.308

UK birth status

Non-UK Born

2.1% (646)

30481

0.763

UK Born

2.3% (240)

10496

0.97 (0.79, 1.18)

0.763

Socio-economic status

1

2.3% (364)

15884

0.517

2

2.1% (263)

12422

0.92 (0.78, 1.08)

0.32

3

2.0% (131)

6435

0.89 (0.72, 1.09)

0.282

4

1.9% (70)

3712

0.83 (0.63, 1.07)

0.168

5

2.3% (58)

2524

1.04 (0.77, 1.37)

0.804

## Date of ending treatment

For date of ending treatment there is evidence that missing data is associated with year of notification and weaker evidence of an association with ethnic group and socio-economic status, with little evidence for any other variable (Table 10). As found previously, variable completeness increased over time. There was some evidence that poorest socio-economic group was more likely to be missing the date of ending treatment but the evidence for this was mixed. The White ethnic group was more somewhat likely to be missing date of treatment ending than most other ethnic groups.

Table 10: Results from a logistic regression model with data completeness (Complete/Missing) for date of starting treatment as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). For socio-economic group 1 indicates the most deprived quintile. Notifications from 2010 onwards were included as socio-economic status was not collected before this. Complete case analysis was used. Odds ratios shown are adjusted for all explanatory variables. There is little evidence that the missing data for the date of starting treatment is associated with any variable considered, except for year of notification.

Variable

Category

Missing (N)

Notifications (33606)

Odds Ratio

P value (Wald)

P value (LRT)

Year

2010

2.9% (182)

6171

2.52e-14

2011

2.6% (177)

6855

0.88 (0.71, 1.08)

0.223

2012

2.4% (164)

6882

0.80 (0.64, 0.99)

0.0379

2013

1.5% (97)

6298

0.51 (0.39, 0.65)

8.39e-08

2014

1.2% (66)

5341

0.40 (0.30, 0.53)

2.2e-10

2015

1.4% (28)

2059

0.45 (0.30, 0.66)

0.00011

Sex

Female

2.1% (311)

14630

0.859

Male

2.1% (403)

18976

1.01 (0.87, 1.18)

0.86

Age

0-14

2.7% (44)

1617

0.711

15-44

2.0% (419)

21027

0.83 (0.60, 1.18)

0.282

45-64

2.3% (165)

7272

0.88 (0.62, 1.27)

0.479

65+

2.3% (86)

3690

0.83 (0.56, 1.23)

0.338

Ethnic group

White

2.9% (176)

6076

0.00931

Black-Caribbean

2.8% (21)

753

1.01 (0.62, 1.57)

0.972

Black-African

1.9% (114)

6071

0.69 (0.52, 0.93)

0.0162

Black-Other

2.3% (7)

306

0.88 (0.37, 1.78)

0.751

Indian

1.7% (150)

8842

0.66 (0.51, 0.87)

0.00317

Pakistani

2.5% (140)

5668

0.94 (0.72, 1.22)

0.63

Bangladeshi

1.3% (18)

1409

0.48 (0.28, 0.78)

0.00533

Chinese

2.8% (11)

396

1.09 (0.54, 1.99)

0.787

Mixed / Other

1.9% (77)

4085

0.75 (0.54, 1.02)

0.0724

UK birth status

Non-UK Born

1.9% (480)

25174

0.153

UK Born

2.8% (234)

8432

1.17 (0.94, 1.45)

0.151

Socio-economic status

1

2.4% (308)

13080

0.000621

2

1.7% (170)

10266

0.72 (0.60, 0.87)

0.000888

3

1.9% (100)

5265

0.82 (0.65, 1.03)

0.0917

4

2.8% (84)

2994

1.19 (0.92, 1.52)

0.178

5

2.6% (52)

2001

1.07 (0.78, 1.44)

0.681

#### Biases in the ETS

Routine observational datasets are subject to numerous potential biases, such as selection bias, recall bias, measurement bias, and unmeasured confounding.[12] Additionally, as the data has not been collected with a specific analysis in mind there maybe issues with the specificity of variables. The ETS system is likely to suffer from all of the above biases to some extent, which must be accounted for as far as possbile, and explicitly stated at every level of analysis. The most important consideration is that the ETS system is unlikely to be representative of the general population as it contains only notified TB cases that occurred in England during the study period, research questions must therefore be either limited to active TB patients, or when extended to the general population the differing population demographics must be accounted for. If this is not done then any results may be due to selection bias. Additionally, multiple variables may suffer from misclassification bias, including BCG status which can be assessed via vaccination record, the presence of a scar, or case recall: this may lead to spurious associations.[13] Validation studies would be required to account for this, which is beyond the scope of this thesis.

# Discussion

### Statement of primary findings

In the ETS system I found a high degree of missing data for several important variables. I also found that there is likely to be strong missing at random (MAR) mechanism underlying this missing data for multiple variables. Several factors are strongly associated with data being missing for many variables, including UK birth status, ethnic group, socio-economic status and year. These MAR mechanisms must be adjusted for in future analysis to avoid bias. I found that date variables in particular suffered from changing data completeness over time, which may introduce spurious temporal trends if not fully understood. I also found that for several variables, including the date of symptom onset, there was a large degree of recall bias when aggregating by day or month. Several variables, including date of notification and date of starting treatment, showed a seasonal trend with a maximum in the summer months. The date of ending treatment showed less evidence of a seasonal trend.

### Strengths and limitations of the study

### Strengths and limitations in comparison to the literature

### Meaning of the study

### Unanswered questions and future research

**Acknowledgements**

The authors thank the TB section at Public Health England (PHE) for maintaining the Enhanced Tuberculosis Surveillance (ETS) system; all the healthcare workers involved in data collection for the ETS.

**Contributors**

SA, HC, and EBP conceived and designed the work. SA undertook the analysis with advice from all other authors. All authors contributed to the interpretation of the data. SA wrote the first draft of the paper and all authors contributed to subsequent drafts. All authors approve the work for publication and agree to be accountable for the work.

**Funding**

SEA, HC, and EBP are funded by the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) in Evaluation of Interventions at University of Bristol in partnership with Public Health England (PHE). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England.

**Conflicts of interest**

HC reports receiving honoraria from Sanofi Pasteur, and consultancy fees from AstraZeneca, GSK and IMS Health, all paid to her employer.

**Accessibility of data and programming code**

The code and data for the analysis contained in this paper can be found at:

# References

1 Public Health England. Tuberculosis in England 2017 report ( presenting data to end of 2016 ) About Public Health England. 2017.

2 Public Health England. Tuberculosis in England: 2018. 2019;1–218.

3 French CE, Antoine D, Gelb D *et al.* Tuberculosis in non-UK-born persons, England and Wales, 2001-2003. *Int J Tuberc Lung Dis* 2007;**11**:577–84.

4 Kruijshaar M, French C, Anderson C *et al.* Tuberculosis in the UK, Annual report on tuberculosis surveillance and control in the UK 2007. *Thorax* 2007;**50**:703–3.

5 Abubakar I, Pimpin L, Ariti C *et al.* Systematic review and meta-analysis of the current evidence on the duration of protection by bacillus Calmette-Guérin vaccination against tuberculosis. *Health technology assessment* 2013;**17**:1–372, v–vi.

6 Public Health England. The Green Book. 2013;391–409.

7 Sterne JAC, White IR, Carlin JB *et al.* Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 2009;**338**:b2393–3.

8 Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003;**3**:31.

9 PHE. Tuberculosis in England 2016 Report (presenting data to end of 2015). 2016.

10 Groothuis-oudshoorn K. Journal of Statistical Software MICE : Multivariate Imputation by Chained.;**VV**.

11 Sterne JA, Davey Smith G. Sifting the evidence-what’s wrong with significance tests? *Bmj* 2001;**322**:226–31.

12 Benchimol EI, Smeeth L, Guttmann A *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *The American Statistician* 2016;**115-116**:1–22.

13 Fewell Z, Davey Smith G, Sterne JAC. The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study. *American Journal of Epidemiology* 2007;**166**:646–55.

## Results *Copy to top*

# Supplementary Information: Explore and Evaluate the Mechanisms for Missing Data in the Enhanced Tuberculosis Surveillance System

Sam Abbott, Hannah Christensen, Ellen Brooks-Pollock

1. Data cleaning code: <https://www.samabbott.co.uk/tbinenglanddataclean/reference/clean_munge_ets_2016.html> [↑](#footnote-ref-34)