Exploring 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:* 9 *Abstract:* 299 *Paper:* 3669

# Abstract

## Background

The Enhanced Tuberculosis Surveillance (ETS) system is a routine surveillance system that collects data on all notified tuberculosis (TB) cases in England and is routinely used in research. Routine data often has a large amount of missing data which may introduce bias to analyses.

## Aim

Explore potential sources of bias from missing data for key outcomes in the ETS system.

## Methods

We used data on all notifications in the ETS for England (2000-2015). We summarised the proportion of missing data and then used logistic regression to explore the associations between missingness and demographic variables for the following outcomes: drug resistance; BCG status and year; date of symptom onset, diagnosis, notification, starting treatment, finishing treatment and death; and cause of death. For all date variables, we visualised the distribution annually and by month, identifying variables at risk of bias.

## Results

All demographic variables considered were associated with data being missing for multiple outcomes. Missingness was not associated with all demographic variables for all outcomes. Associations that were present did not all have the same direction of effect. We found that the date of symptom onset and the date of ending treatment had a high proportion of cases occurring in January and on the 1st and 14th of each month indicating the presence of potential bias.

## Conclusions

Missingness in outcomes was associated with multiple demographic variables. These associations could not be generalised, preventing a systematic approach to dealing with missingness and indicating that domain knowledge is required. The identified associations may induce bias, meaning that complete case analysis should not be used for this data source - or other similar surveillance data. Our findings should be used to motivate the use of imputation and to inform the specification of imputation models for similar surveillance data.

## Keywords

Tuberculosis, surveillance, missingness, imputation

## Key messages

* Missing data may induce biases into analyses conducted using surveillance data and needs to be properly accounted for.
* Common practice is to use complete case analysis or multiple imputation. We show that common demographic variables are often associated with key outcomes and hence may induce bias if complete case analysis is used.
* Our findings also indicate that variables outside the analysis model should be considered when defining the multiple imputation model - this may require domain specific expertise.
* Finally we found that date variables were potentially at risk of bias due to the presence of a higher than expected proportion of cases in January, and on the 1st and 14th day of each month for multiple variables.

# Introduction

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. Studies rarely independently assess the impact of missing data - instead aiming to account for it during analysis. This study explores missingness in the ETS - focusing on associations between key outcomes and demographic variables with the aim of highlighting potential sources of bias.

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 - or MNAR - may lead to biases when analysing the data. This means that 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[1,2]. Common practice is to include all variables included in the analyses in the imputation model and best practise is to also include additional variables that may be related to missing data[3]. Several studies have used imputation for analyses using the ETS and found that results were impacted compared to using complete case analysis or imputation using only analysis variables[4,5].

This study aims to explore the evidence for associations between missingness in several key outcomes and demographic variables in the ETS. Any such associations may introduce bias if not accounted for. It also aims to more broadly assess missing data in the ETS, evaluate the impact of the introduction of the web based ETS in 2008 on missing data, and highlight potential biases in reporting for date variables.

# Methods

## 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[6]. Data collection began in 2000 and was expanded, with additional variables, with the launch of a web-based system in 2008[7]. 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[6]. 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. The code used for data cleaning is available as an R package (<https://zenodo.org/badge/latestdoi/93072437>).

### Data completeness

As the ETS is aggregated across England, from a variety of sources, 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 data-set the proportion of missing data can be calculated but care must be taken to account for nested variables (such as cause of death being dependent on date of death). To account for this when estimating the proportion of missing data we have assumed that nested variables take the value of their parent variable when missing. This approach may be biased for rare outcomes (such as death in the ETS) - for this reason we have also estimated the proportion of missing data by filtering top level variables required for the nested variable to be defined and then computed the proportion of notifications that were missing data for the outcome of interest.

### Drivers of Variable completeness

**Overview**

Missing data may be MAR or MNAR, which may introduce biases into any analyses based on these data. 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 may not be included in a model used for analysis. Here we describe a method for this and apply it to several key outcomes including: Drug resistance (any treatment), 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.

We 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. This approach does not account for missingness within exploratory variables.

**Statistical details**

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[9]. 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 we adjusted for the same set of demographic variables that were both highly complete, plausibly linked to missingness for all outcomes considered, and likely to be present in other comparable surveillance data-sets. These were: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status, socio-economic status (national quintiles), and PHE centre (region). Complete case analysis was used, with the data-set limited to notifications from 2010 and on-wards as socio-economic status was not collected prior to this. The code for this approach is available as an R package (<https://doi.org/10.5281/zenodo.3492200>).

### Assessing temporal biases in reporting

In addition to data being MAR there may be other biases present. For date variables this is a particular issue with recall bias, reporting bias etc. potentially distorting temporal trends. We explore this by summarising the distribution of all date variables by month and then by day of the month, stratified by the introduction of the web-based ETS system (2009). The date of notification is then used as a baseline for the inherent seasonal or monthly reporting structure. This approach allows potential biases to be identified and compared across the current and pre-web ETS. For each date data was restricted with only data from 2000 until 2015 being used.

## Patient and public involvement

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

# Results

## Data completeness

We found high completeness for common demographic variables such as sex, age, ethnic group and UK birth status (Supplementary Figure S1, Table 1). More problematically, BCG status and year of BCG status had a high percentage missing, even after accounting for the introduction of national collection of these variables in 2008[6]. Socio-economic status (as national quintiles) was not collected until 2010 but after this point is highly complete[6]. Comparing before and after the launch of the web-based ETS in Table 1 (Supplementary Figure S1) we see completeness changes over time[6]. There was some evidence that groups of variables had correlated missing data (Supplementary Figure S1).

Table 1: Percentage of missing data from the ETS for a subset of variables, prior to the web-based system (pre 2009) and post (post 2008) by variable, ordered by the percentage missing for a subset of variables. 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.

|  |  |  |
| --- | --- | --- |
| Variable | 2000-2008 | 2009-2015 |
| Socio-economic status (quintiles) | 100.0 (63175) | 15.7 (8120) |
| Year of BCG vaccination | 98.9 (62479) | 60.8 (31421) |
| BCG status | 98.0 (61916) | 33.2 (17133) |
| Date of diagnosis | 72.1 (45557) | 19.9 (10303) |
| Sputum smear status | 52.1 (32912) | 62.1 (32094) |
| Time since entry | 46.0 (29084) | 36.2 (18670) |
| Drug resistance | 43.5 (27485) | 40.7 (20995) |
| Occupation | 39.4 (24870) | 10.7 (5513) |
| Date of symptom onset | 37.9 (23937) | 24.8 (12829) |
| Treatment end date | 29.6 (18711) | 2.2 (1137) |
| Previous diagnosis | 20.9 (13204) | 6.1 (3148) |
| Date of starting treatment | 14.5 (9151) | 4.1 (2127) |
| Cause of death | 11.9 (7539) | 2.3 (1191) |
| UK birth status | 9.9 (6230) | 3.5 (1825) |
| Overall outcome | 9.6 (6044) | 0.0 (0) |
| Started treatment | 6.7 (4242) | 1.2 (602) |
| Ethnic group | 4.4 (2811) | 2.4 (1229) |
| Date of death | 2.0 (1235) | 0.7 (357) |
| Pulmonary or extra-pulmonary TB | 0.3 (177) | 0.4 (213) |
| Sex | 0.2 (101) | 0.2 (110) |
| Public Health England Centre | 0.1 (32) | 0.0 (0) |
| Age | 0.0 (25) | 0.0 (0) |
| Date of notification | 0.0 (0) | 0.0 (0) |
| Year | 0.0 (0) | 0.0 (0) |
| Culture | 0.0 (0) | 0.0 (0) |

By filtering nested variables - rather than by using replacement - we found the date of starting treatment was 5.9% (6434/108410) missing, which is more complete than previously estimated. For cases that were known to have completed treatment 16.5% (13804/83891) were missing a date for the end of treatment. In notifications that were 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.

## Drivers of Variable completeness

### Drug resistance

There was evidence that drug resistance was missing with a MAR mechanism for all variables considered (Table 2), excepting year of notification. Men were less likely to be missing than women. Children were much more likely to have a missing drug resistance status than any other age group. The white ethnic group were less likely to be missing drug resistance than all other ethnic groups, excepting the Chinese ethnic group. The UK born population was more likely to be missing as were those from higher economic quintiles. Notifications in London were more likely to be missing drug resistance status than for most other PHE centres.

Table 2: Results from a logistic regression model with data completeness (Complete/Missing) for drug resistance (to any 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Category | Missing (N) | Notifications (41659) | Odds Ratio | P value (Wald) | P value (LRT) |
| Year | 2010 | 40.7% (2905) | 7143 | Reference |  | 0.844 |
|  | 2011 | 40.2% (3126) | 7781 | 0.97 (0.91, 1.04) | 0.428 |  |
|  | 2012 | 40.1% (3107) | 7755 | 0.96 (0.90, 1.03) | 0.278 |  |
|  | 2013 | 40.4% (2839) | 7034 | 0.98 (0.92, 1.05) | 0.625 |  |
|  | 2014 | 39.8% (2519) | 6327 | 0.96 (0.90, 1.03) | 0.29 |  |
|  | 2015 | 40.3% (2267) | 5619 | 1.00 (0.93, 1.07) | 0.896 |  |
| Sex | Female | 43.1% (7613) | 17664 | Reference |  | 4.81e-21 |
|  | Male | 38.1% (9150) | 23995 | 0.82 (0.79, 0.86) | 4.67e-21 |  |
| Age | 0-14 | 76.1% (1365) | 1793 | Reference |  | 2.95e-229 |
|  | 15-44 | 36.0% (9096) | 25235 | 0.18 (0.16, 0.21) | 6e-180 |  |
|  | 45-64 | 43.9% (3961) | 9026 | 0.26 (0.23, 0.29) | 2.07e-105 |  |
|  | 65+ | 41.8% (2341) | 5605 | 0.23 (0.20, 0.26) | 1.48e-112 |  |
| Ethnic group | White | 40.2% (3364) | 8359 | Reference |  | 8.17e-29 |
|  | Black-Caribbean | 40.1% (372) | 928 | 0.99 (0.85, 1.14) | 0.854 |  |
|  | Black-African | 38.5% (2775) | 7204 | 1.07 (0.98, 1.16) | 0.133 |  |
|  | Black-Other | 42.5% (157) | 369 | 1.20 (0.96, 1.49) | 0.105 |  |
|  | Indian | 40.7% (4412) | 10848 | 1.24 (1.15, 1.34) | 1.58e-08 |  |
|  | Pakistani | 42.4% (2885) | 6806 | 1.31 (1.21, 1.41) | 2.7e-11 |  |
|  | Bangladeshi | 47.1% (791) | 1680 | 1.67 (1.48, 1.88) | 9.74e-18 |  |
|  | Chinese | 34.6% (171) | 494 | 0.97 (0.80, 1.18) | 0.787 |  |
|  | Mixed / Other | 36.9% (1836) | 4971 | 1.01 (0.92, 1.10) | 0.911 |  |
| UK birth status | Non-UK Born | 38.6% (11913) | 30880 | Reference |  | 3.1e-08 |
|  | UK Born | 45.0% (4850) | 10779 | 1.19 (1.12, 1.26) | 3e-08 |  |
| Socio-economic status | 1 | 40.0% (6454) | 16131 | Reference |  | 0.000369 |
|  | 2 | 39.7% (5005) | 12621 | 1.02 (0.97, 1.07) | 0.487 |  |
|  | 3 | 40.3% (2633) | 6530 | 1.06 (1.00, 1.13) | 0.0563 |  |
|  | 4 | 41.1% (1561) | 3796 | 1.10 (1.02, 1.19) | 0.0125 |  |
|  | 5 | 43.0% (1110) | 2581 | 1.21 (1.10, 1.32) | 4.57e-05 |  |
| Public Health England centre | London | 40.4% (7135) | 17658 | Reference |  | 6.46e-15 |
|  | West Midlands | 43.6% (2274) | 5217 | 1.07 (1.00, 1.15) | 0.0416 |  |
|  | North West | 39.2% (1597) | 4075 | 0.87 (0.81, 0.94) | 0.000464 |  |
|  | South East | 38.2% (1542) | 4037 | 0.87 (0.81, 0.94) | 0.000287 |  |
|  | Yorkshire and the Humber | 40.9% (1257) | 3077 | 0.91 (0.84, 0.99) | 0.027 |  |
|  | East of England | 38.3% (1019) | 2662 | 0.87 (0.80, 0.95) | 0.00171 |  |
|  | East Midlands | 40.2% (1025) | 2548 | 0.95 (0.87, 1.04) | 0.27 |  |
|  | South West | 42.3% (674) | 1595 | 1.06 (0.95, 1.18) | 0.3 |  |
|  | North East | 30.4% (240) | 790 | 0.60 (0.51, 0.70) | 3.43e-10 |  |

### BCG status and year of BCG vaccination

Similarly to drug resistance there was evidence that BCG status was missing with a MAR mechanism for all variables considered (Table 3) with the stronger evidence for an association with year but reduced evidence of an association with socio-economic status. After adjusting for other variables data completeness increased from 2010 until 2012 but has since showed no clear trend. Men appeared 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 increased with age, with those aged 65+ being over 4 times more likely to be missing BCG status than those aged 0-14 years old. The White ethnic group was more likely to have a missing BCG status than any other ethnic group. London was associated with less reduced missingness for BCG status compared to other PHE centres.

Missingness for year of BC vaccination had similar associations as BCG status. However, there was less evidence of an association with sex, the white ethnic group were less likely to have a missing status than other ethnic groups, and there was strong evidence of an association with socio-economic status with those in the lowest quintile being more likely to have a missing year of BCG vaccination. London was much more likely to be missing BCG status than any other PHE centre, a reversal of the relationship observed for BCG status.

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 | Reference |  | 1.6e-08 |
|  | 2011 | 29.8% (2319) | 7781 | 0.94 (0.88, 1.01) | 0.107 |  |
|  | 2012 | 27.9% (2164) | 7755 | 0.85 (0.79, 0.92) | 1.93e-05 |  |
|  | 2013 | 27.1% (1907) | 7034 | 0.79 (0.73, 0.85) | 1.3e-09 |  |
|  | 2014 | 30.1% (1907) | 6327 | 0.90 (0.83, 0.97) | 0.00672 |  |
|  | 2015 | 29.7% (1668) | 5619 | 0.88 (0.81, 0.95) | 0.00104 |  |
| Sex | Female | 27.4% (4847) | 17664 | Reference |  | 5.21e-14 |
|  | Male | 30.6% (7353) | 23995 | 1.19 (1.14, 1.24) | 5.97e-14 |  |
| Age | 0-14 | 13.1% (235) | 1793 | Reference |  | 8.49e-162 |
|  | 15-44 | 26.0% (6557) | 25235 | 2.24 (1.94, 2.60) | 5.72e-27 |  |
|  | 45-64 | 32.8% (2964) | 9026 | 3.05 (2.63, 3.55) | 3.38e-47 |  |
|  | 65+ | 43.6% (2444) | 5605 | 4.82 (4.13, 5.64) | 1.93e-87 |  |
| Ethnic group | White | 35.4% (2959) | 8359 | Reference |  | 1.18e-14 |
|  | Black-Caribbean | 24.6% (228) | 928 | 0.88 (0.74, 1.03) | 0.124 |  |
|  | Black-African | 27.3% (1966) | 7204 | 0.87 (0.79, 0.95) | 0.00235 |  |
|  | Black-Other | 24.1% (89) | 369 | 0.87 (0.67, 1.12) | 0.275 |  |
|  | Indian | 25.9% (2805) | 10848 | 0.71 (0.65, 0.77) | 3.69e-16 |  |
|  | Pakistani | 33.2% (2258) | 6806 | 0.85 (0.78, 0.93) | 0.000209 |  |
|  | Bangladeshi | 27.9% (469) | 1680 | 0.92 (0.81, 1.05) | 0.214 |  |
|  | Chinese | 33.6% (166) | 494 | 0.91 (0.74, 1.12) | 0.395 |  |
|  | Mixed / Other | 25.3% (1260) | 4971 | 0.80 (0.72, 0.88) | 5.15e-06 |  |
| UK birth status | Non-UK Born | 29.5% (9104) | 30880 | Reference |  | 7.78e-28 |
|  | UK Born | 28.7% (3096) | 10779 | 0.68 (0.63, 0.73) | 2.69e-27 |  |
| Socio-economic status | 1 | 30.7% (4948) | 16131 | Reference |  | 0.0647 |
|  | 2 | 26.8% (3383) | 12621 | 1.01 (0.95, 1.07) | 0.825 |  |
|  | 3 | 29.2% (1905) | 6530 | 1.09 (1.01, 1.16) | 0.0187 |  |
|  | 4 | 30.1% (1142) | 3796 | 0.98 (0.90, 1.06) | 0.616 |  |
|  | 5 | 31.8% (822) | 2581 | 0.96 (0.87, 1.06) | 0.415 |  |
| Public Health England centre | London | 21.0% (3716) | 17658 | Reference |  | 0 |
|  | West Midlands | 22.4% (1171) | 5217 | 1.08 (0.99, 1.16) | 0.066 |  |
|  | North West | 51.8% (2112) | 4075 | 4.16 (3.85, 4.49) | 4.44e-286 |  |
|  | South East | 26.6% (1074) | 4037 | 1.33 (1.23, 1.45) | 7.73e-12 |  |
|  | Yorkshire and the Humber | 37.0% (1138) | 3077 | 2.24 (2.05, 2.44) | 1.35e-72 |  |
|  | East of England | 36.4% (969) | 2662 | 2.12 (1.94, 2.32) | 6.4e-61 |  |
|  | East Midlands | 45.3% (1154) | 2548 | 3.20 (2.93, 3.50) | 4.07e-145 |  |
|  | South West | 41.2% (657) | 1595 | 2.55 (2.28, 2.85) | 5.96e-62 |  |
|  | North East | 26.5% (209) | 790 | 1.31 (1.11, 1.55) | 0.0013 |  |

### 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 4). 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. London was again associated with a increased level of missing data compared to other PHE centres.

Table 4: 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 | Reference |  | 0 |
|  | 2011 | 30.1% (2339) | 7781 | 0.84 (0.78, 0.90) | 1.45e-06 |  |
|  | 2012 | 24.2% (1878) | 7755 | 0.61 (0.57, 0.66) | 1.73e-38 |  |
|  | 2013 | 17.5% (1233) | 7034 | 0.41 (0.37, 0.44) | 2.6e-105 |  |
|  | 2014 | 11.8% (744) | 6327 | 0.25 (0.23, 0.27) | 6.1e-187 |  |
|  | 2015 | 6.9% (390) | 5619 | 0.14 (0.12, 0.15) | 1.7e-245 |  |
| Sex | Female | 22.0% (3894) | 17664 | Reference |  | 0.363 |
|  | Male | 21.3% (5116) | 23995 | 0.98 (0.93, 1.03) | 0.363 |  |
| Age | 0-14 | 38.1% (684) | 1793 | Reference |  | 6.9e-78 |
|  | 15-44 | 20.5% (5182) | 25235 | 0.33 (0.30, 0.38) | 4.33e-78 |  |
|  | 45-64 | 20.7% (1870) | 9026 | 0.36 (0.32, 0.41) | 4.15e-58 |  |
|  | 65+ | 22.7% (1274) | 5605 | 0.44 (0.39, 0.51) | 3.41e-34 |  |
| Ethnic group | White | 20.9% (1749) | 8359 | Reference |  | 1.53e-08 |
|  | Black-Caribbean | 23.1% (214) | 928 | 0.76 (0.64, 0.90) | 0.00216 |  |
|  | Black-African | 23.0% (1654) | 7204 | 0.72 (0.65, 0.79) | 7.47e-11 |  |
|  | Black-Other | 18.7% (69) | 369 | 0.61 (0.45, 0.80) | 0.000611 |  |
|  | Indian | 22.2% (2404) | 10848 | 0.76 (0.70, 0.84) | 1.17e-08 |  |
|  | Pakistani | 19.2% (1305) | 6806 | 0.79 (0.72, 0.87) | 3.23e-06 |  |
|  | Bangladeshi | 23.9% (401) | 1680 | 0.80 (0.69, 0.92) | 0.00178 |  |
|  | Chinese | 18.8% (93) | 494 | 0.68 (0.53, 0.87) | 0.0025 |  |
|  | Mixed / Other | 22.6% (1121) | 4971 | 0.79 (0.71, 0.88) | 1.07e-05 |  |
| UK birth status | Non-UK Born | 21.9% (6774) | 30880 | Reference |  | 0.000152 |
|  | UK Born | 20.7% (2236) | 10779 | 0.86 (0.80, 0.93) | 0.00016 |  |
| Socio-economic status | 1 | 19.9% (3218) | 16131 | Reference |  | 1.06e-06 |
|  | 2 | 22.9% (2888) | 12621 | 0.98 (0.92, 1.05) | 0.63 |  |
|  | 3 | 24.2% (1578) | 6530 | 1.17 (1.08, 1.26) | 7.32e-05 |  |
|  | 4 | 22.0% (837) | 3796 | 1.18 (1.07, 1.29) | 0.000845 |  |
|  | 5 | 18.9% (489) | 2581 | 1.17 (1.04, 1.31) | 0.01 |  |
| Public Health England centre | London | 30.0% (5289) | 17658 | Reference |  | 0 |
|  | West Midlands | 12.0% (627) | 5217 | 0.30 (0.27, 0.33) | 8.63e-137 |  |
|  | North West | 20.6% (841) | 4075 | 0.56 (0.51, 0.61) | 5.62e-36 |  |
|  | South East | 9.0% (363) | 4037 | 0.20 (0.18, 0.23) | 7.15e-156 |  |
|  | Yorkshire and the Humber | 13.2% (407) | 3077 | 0.32 (0.28, 0.35) | 4.19e-83 |  |
|  | East of England | 26.5% (705) | 2662 | 0.80 (0.72, 0.88) | 4.54e-06 |  |
|  | East Midlands | 19.2% (488) | 2548 | 0.52 (0.47, 0.58) | 2.21e-32 |  |
|  | South West | 10.9% (174) | 1595 | 0.27 (0.23, 0.32) | 6.79e-53 |  |
|  | North East | 14.7% (116) | 790 | 0.39 (0.31, 0.47) | 1.9e-19 |  |

### Date of diagnosis

For date of diagnosis there was again strong evidence for an MAR mechanism for all variables considered, except for sex (Supplementary Table S1). 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 more likely to be missing data on the data of diagnosis as compared to the majority of other ethnic groups. The poorest socio-economic group was less likely to be missing data 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. As for other variables London had a much higher proportion of missing data than any other PHE centre.

### Date of starting treatment and ending treatment

For date of starting treatment there was evidence that missing data was again associated with all variables considered, excepting UK birth status and socio-economic status (Supplementary Table S2). Missingness for the date of ending treatment was associated with fewer variables with evidence only of associations between year and PHE centre (Supplementary Table S3). For both variables the proportion of missing data reduced with the year of notification. London had a lower proportion of missing data when compared to most other PHE centres. For the date of starting treatment the White ethnic group were more likely to be missing data than other groups. Older age groups were also more likely to be missing data, as were males.

### Date of death and cause of death.

For date of death there was little evidence of any association, except for PHE centre (Supplementary Table S4). This was also the case for cause of death but there was some additional evidence of an association with ethnic group (Supplementary Table S5). There was little evidence of a clear trend across ethnic groups for cause of death. As for other outcomes London was much more likely to be missing date of death than other PHE centres. This relationship was reversed for cause of death. Both date of death and cause of death had a small sample size and this may mean that these analyses were underpowered.

## Assessing temporal biases in reporting

Notifications showed evidence of a strong seasonal trend with a peak in the number of notifications in May-July each year but had a near uniform distribution within each month (Supplementary Table S6). There was little evidence of strong biases in this reporting and there was little evidence to suggest that the introduction of the web-based ETS impacted the distribution of notifications or the levels of bias. The date of symptom onset showed evidence of an inverted seasonal trend - in comparison to notifications (Table 5) . There was evidence that reporting in January may be biased with a much greater proportion of cases reported as having symptoms starting in this month than in any other. There was also evidence that cases were more likely to have symptoms start on the first and the 14th of each month, again indicating bias. Both of these apparent biases were reduced by the introduction of the web-based ETS but were still present. The date of ending treatment also showed some evidence of these biases and had the same inverted seasonal trend as the date of symptom onset (Supplementary Table S7). The date of diagnosis, date of starting treatment and date of death showed a similar reporting structure to notifications although the strength of the seasonal trend was reduced (see the supplementary information). There was little clear evidence of biases in reporting either by month, or by day for these variables.

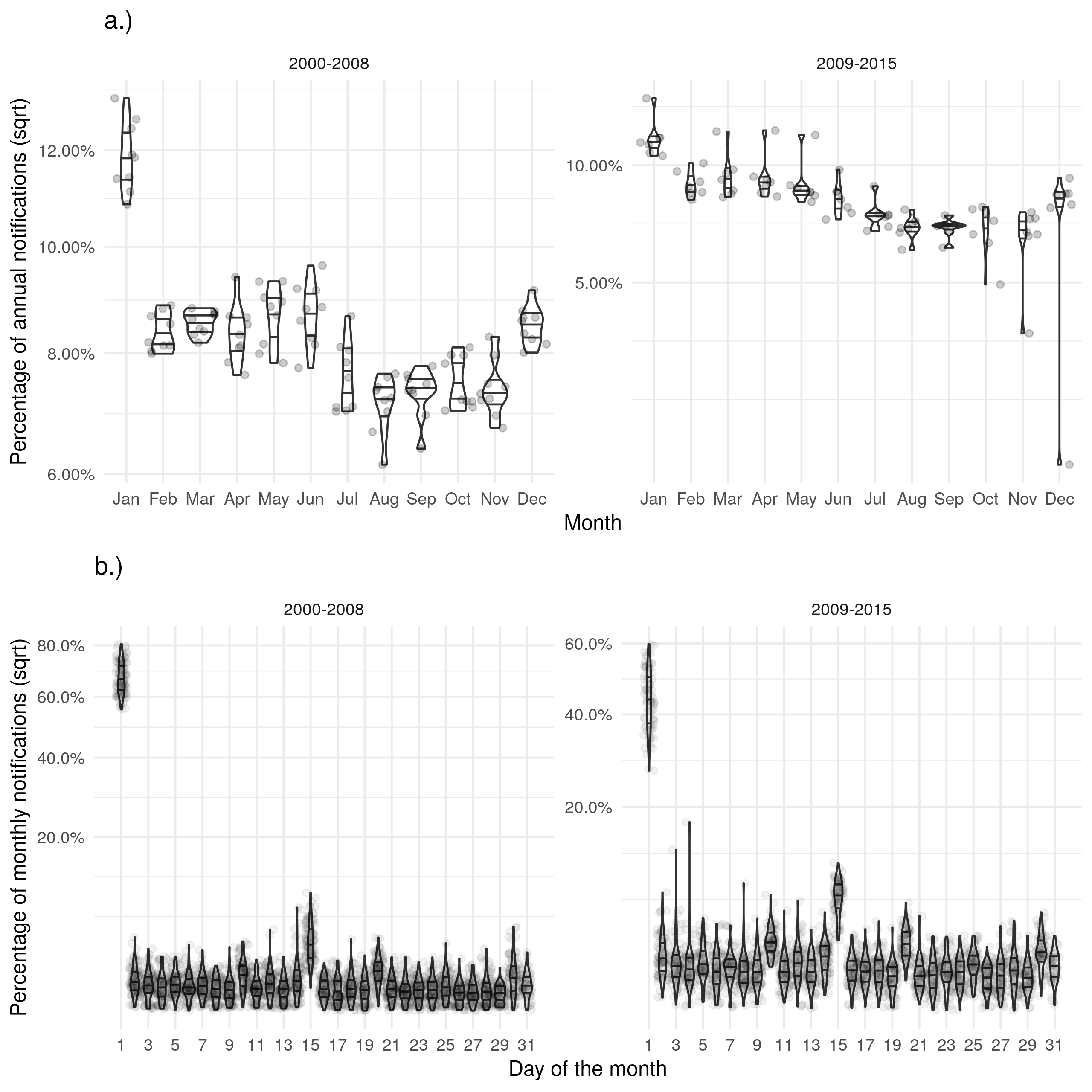


Figure 1: a.) Shows the proportion of cases with symptons starting in a given month for each year with some evidence of bias in January and reduced evidence of a seasonal trend. b.) Shows the proportion of cases with symptons starting on a given day for each month with a strong evidence of biased reporting on the first of the month and the 14th. Stratifying both figures based on the introduction of the web-based ETS indicates that the web-based ETS may have reduced these biases.

# Discussion

We found a high degree of missing data for several variables in the ETS. All demographic variables considered were strongly associated with data being missing for multiple outcomes. However, not all outcome missingness was associated with all demographic variables and for those that were the direction of - or trend in - effect was not consistent. Missingness for drug resistance was associated with all variables - excepting year - with higher levels of missingness being associated with children, non-White ethnic groups, the UK born, higher socio-economic status, and the London PHE centre (though there was between PHE centre variation). In comparison, the associations with BCG status had a similar level of evidence but did not share the same trends with children being much less likely to have missing data than older adults, non-White ethnic groups having a lower levels of missing data than the White ethnic group, the non-UK born being more likely to be missing than the UK born, and the London PHE centre being less likely to be missing data than other PHE centres. Missingness for date and cause of death had less evidence of associations than other outcomes. We also found that date variables in particular suffered from changing data completeness over time. In addition, we found that both the date of symptom onset and the date of ending treatment had a higher than expected proportion of cases occurring in January each year and on the 1st and 14th of each month when compared to notification date. This may indicate the presence of reporting or recall bias. These potential biases were reduced after the launch of the web-based ETS but were still present.

This study has explored missing data in the ETS - which is an example of a well designed surveillance system - in detail and found that missing data can rarely be considered missing completely at random. We highlighted several key demographic variables as being potential sources of bias, but did not find a generalised structure to these biases across key outcomes. Routine observational data-sets are subject to numerous other potential biases not explored in this study. These sources of bias include: selection bias, recall bias, measurement bias, and unmeasured confounding[10]. These sources of bias may be difficult to quantify in a single routine surveillance data-set as they require knowledge of the population in order to identify - except in the instances seen above with spuriously high reporting for some months, or days in a month. This means that studies using routine surveillance data are never likely to be free of bias. However, by using imputation and including a wide range of variables plausibly linked to variable missingness - as we have demonstrated here - this bias can be reduced[11]. Also as demonstrated here, by exploring variable reporting across times additional potential sources of bias can be identified and then potentially mitigated. This study does not detail imputation best practises see [1,11] for further information and [4,5] for case studies in the ETS. 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[12]. These potential sources of bias require additional verification studies in order to identify and account for them. Finally, our study may be used to inform studies in surveillance data-sets other than the ETS but cannot be directly generalised as we only considered a single data-set. For this reason we conducted our analysis within a highly reproducible framework and hence our approach should be readily applicable to other data sources with little alteration.

Missing data in the ETS was highly complex with changing completeness over time and associations with multiple demographic factors. The launch of the web-based ETS system was linked to reduced missing data, followed by improved completeness over subsequent years for multiple outcomes but for most outcomes this improvement then stalled. An updated data collection system may be able to reduce missing data - and MAR missingness - further. The nature of the associations could not be generalised over the outcomes considered with it being clear that individual mechanisms led to each. However, for all outcomes considered - that were well powered - there was evidence of associations with demographic variables. This indicates that missingness is likely to be not MCAR and hence must be accounted for in any analysis, ideally using multiple imputation[2,11]. This means that complete case analysis should not be used on its own for this data source - or other similar surveillance data. Whilst many of the variables we considered may often be considered in analysis models - and hence included in imputation - this may not be the case. Therefore, considerable care should be taken when specifying imputation models with other potentially informative variables also being included. Our findings highlight the need for those involved with data collection to also be involved with downstream analysis. Data issues that could lead to bias could then be readily identified by those with the required domain knowledge. This approach may not be feasible in widely studied data-sets, an alternative would be that data collectors share fully imputed data - using all available variables - rather than raw data with missing values. This would allow those best placed to understand the potential sources of bias to mitigate for them and leave downstream users to conduct analysis without having to account further for missingness.

Whilst this analysis was able to identify multiple potential sources of bias for key outcomes it did not quantify the impact that these may have on analyses using these outcomes. Several studies have made use of the ETS - using imputation to adjust for biases due to missing data - but to our knowledge none have explicitly focused on the impact that biases due to missing data may have had[4,5]. A case study focussing on the impact of missing data on study outcomes could be used to highlight the impact of improperly accounting for missing data in surveillance data-sets. This analysis only used a single surveillance data-set - in order for the findings to be more easily generalised to other data-sets it could be repeated in additional surveillance data. For this reason this analysis has been structured as an R package (<https://doi.org/10.5281/zenodo.3492200>).

**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 conceived and designed the work. SA undertook the analysis with advice from all other authors. All authors contributed to the interpretation of the results. 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**

SA, 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 used to clean the data used in this paper can be found at: <https://doi.org/10.5281/zenodo.2551555>

The code for this analysis, interim results, and final results can be found at: <https://doi.org/10.5281/zenodo.3492200>

# References

1 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.

2 Buuren S van, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles* 2011;**45**:1–67. doi:[10.18637/jss.v045.i03](https://doi.org/10.18637/jss.v045.i03)

3 Rezvan PH, Lee KJ, Simpson JA. The rise of multiple imputation : a review of the reporting and implementation of the method in medical research. 2015;1–14. doi:[10.1186/s12874-015-0022-1](https://doi.org/10.1186/s12874-015-0022-1)

4 Abbott S, Christensen H, Lalor MK *et al.* Exploring the effects of bcg vaccination in patients diagnosed with tuberculosis: Observational study using the enhanced tuberculosis surveillance system. *Vaccine* 2019;**37**:5067–72. doi:[https://doi.org/10.1016/j.vaccine.2019.06.056](https://doi.org/https://doi.org/10.1016/j.vaccine.2019.06.056)

5 Abbott S, Christensen H, Welton N *et al.* Estimating the effect of the 2005 change in bcg policy in england: A retrospective cohort study. *bioRxiv* Published Online First: 2019. doi:[10.1101/567511](https://doi.org/10.1101/567511)

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

7 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.

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

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

10 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.

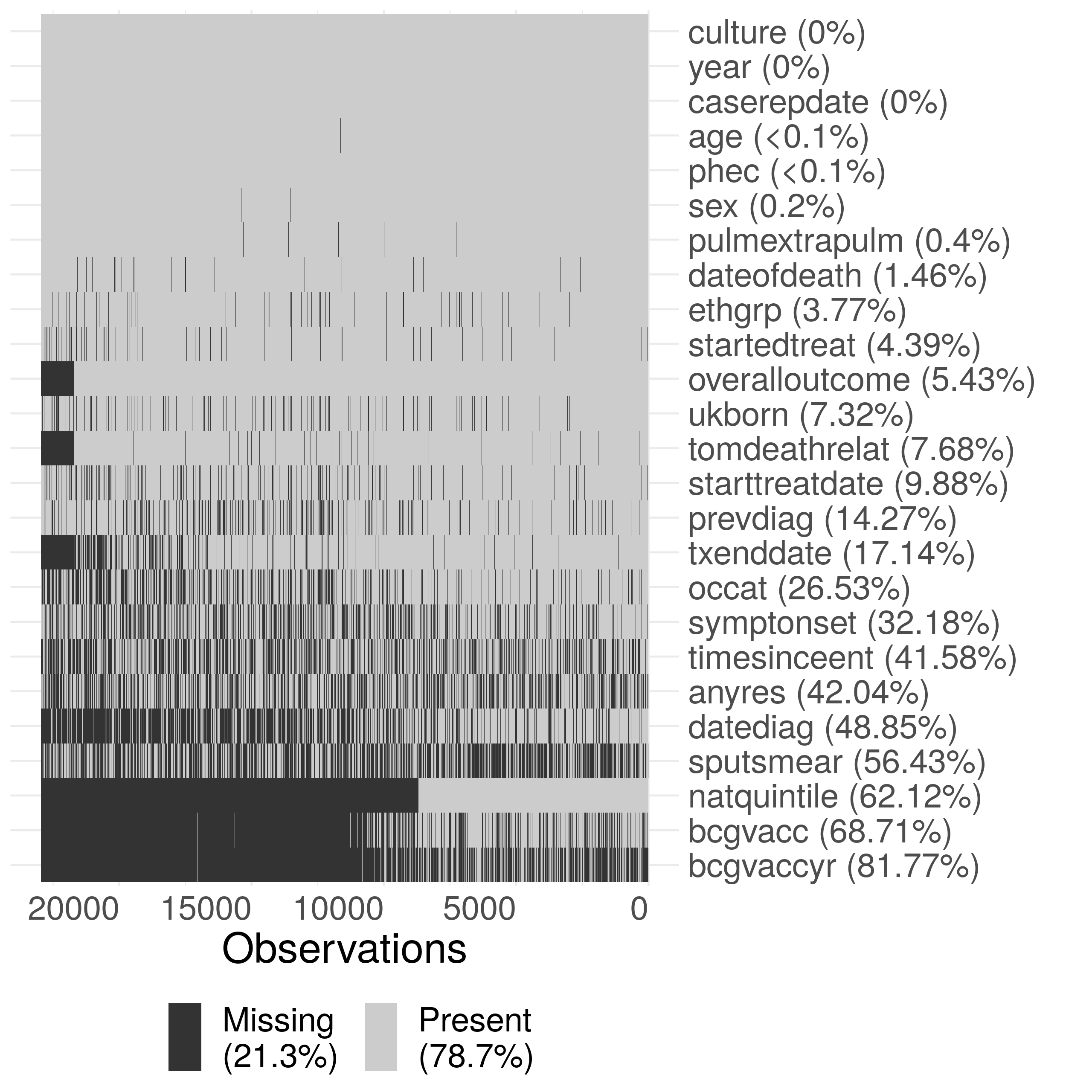
11 Azur MJ, Stuart EA, Frangakis C *et al.* Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011;**20**:40–9. doi:[10.1002/mpr.329](https://doi.org/10.1002/mpr.329)

12 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.

# Supplementary Information: Exploring Missing Data in the Enhanced Tuberculosis Surveillance System

Sam Abbott, Hannah Christensen, Ellen Brooks-Pollock

## Data completeness



Supplementary Figure S1: 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).

## Drivers of data completeness - additional results tables

### Year of BCG vaccination

Supplementary Table S8: Results from a logistic regression model with data completeness (Complete/Missing) for year of 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 year of BCG vaccination is missing at random for the variables considered.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Category | Missing (N) | Notifications (20835) | Odds Ratio | P value (Wald) | P value (LRT) |
| Year | 2010 | 61.0% (2090) | 3424 | Reference |  | 1.59e-09 |
|  | 2011 | 59.6% (2304) | 3869 | 0.90 (0.79, 1.03) | 0.134 |  |
|  | 2012 | 56.2% (2216) | 3945 | 0.73 (0.64, 0.84) | 6.21e-06 |  |
|  | 2013 | 55.7% (2025) | 3638 | 0.75 (0.65, 0.86) | 2.71e-05 |  |
|  | 2014 | 56.6% (1776) | 3138 | 0.83 (0.72, 0.95) | 0.00891 |  |
|  | 2015 | 54.2% (1530) | 2821 | 0.64 (0.55, 0.74) | 1.34e-09 |  |
| Sex | Female | 55.5% (5089) | 9174 | Reference |  | 0.275 |
|  | Male | 58.8% (6852) | 11661 | 1.05 (0.97, 1.13) | 0.275 |  |
| Age | 0-14 | 43.9% (488) | 1111 | Reference |  | 1.21e-20 |
|  | 15-44 | 58.3% (8216) | 14102 | 2.12 (1.77, 2.53) | 1.38e-16 |  |
|  | 45-64 | 57.6% (2526) | 4388 | 2.42 (1.99, 2.94) | 6.72e-19 |  |
|  | 65+ | 57.6% (711) | 1234 | 3.00 (2.36, 3.83) | 5.09e-19 |  |
| Ethnic group | White | 44.2% (1370) | 3102 | Reference |  | 5.86e-12 |
|  | Black-Caribbean | 77.5% (371) | 479 | 1.19 (0.89, 1.61) | 0.242 |  |
|  | Black-African | 65.2% (2524) | 3870 | 0.91 (0.78, 1.07) | 0.261 |  |
|  | Black-Other | 72.0% (154) | 214 | 1.23 (0.80, 1.90) | 0.349 |  |
|  | Indian | 56.1% (3516) | 6267 | 0.75 (0.65, 0.86) | 7.27e-05 |  |
|  | Pakistani | 51.6% (1583) | 3066 | 1.10 (0.95, 1.28) | 0.205 |  |
|  | Bangladeshi | 73.1% (583) | 797 | 1.48 (1.15, 1.90) | 0.00226 |  |
|  | Chinese | 58.2% (142) | 244 | 1.23 (0.83, 1.80) | 0.3 |  |
|  | Mixed / Other | 60.7% (1698) | 2796 | 0.83 (0.70, 0.98) | 0.0318 |  |
| UK birth status | Non-UK Born | 61.1% (9665) | 15808 | Reference |  | 5.14e-08 |
|  | UK Born | 45.3% (2276) | 5027 | 0.74 (0.66, 0.82) | 4.98e-08 |  |
| Socio-economic status | 1 | 55.4% (4221) | 7615 | Reference |  | 4.64e-05 |
|  | 2 | 66.3% (4463) | 6729 | 0.88 (0.79, 0.97) | 0.0118 |  |
|  | 3 | 59.4% (2019) | 3401 | 0.84 (0.74, 0.95) | 0.00684 |  |
|  | 4 | 45.3% (838) | 1848 | 0.70 (0.60, 0.82) | 6.29e-06 |  |
|  | 5 | 32.2% (400) | 1242 | 0.78 (0.65, 0.93) | 0.00583 |  |
| Public Health England centre | London | 91.0% (9421) | 10358 | Reference |  | 0 |
|  | West Midlands | 39.3% (1010) | 2568 | 0.06 (0.05, 0.07) | 0 |  |
|  | North West | 9.2% (116) | 1260 | 0.01 (0.01, 0.01) | 0 |  |
|  | South East | 13.0% (293) | 2261 | 0.01 (0.01, 0.02) | 0 |  |
|  | Yorkshire and the Humber | 45.2% (528) | 1167 | 0.08 (0.07, 0.09) | 2.85e-255 |  |
|  | East of England | 19.9% (260) | 1305 | 0.02 (0.02, 0.03) | 0 |  |
|  | East Midlands | 3.1% (33) | 1066 | 0.00 (0.00, 0.00) | 2.6e-224 |  |
|  | South West | 38.4% (175) | 456 | 0.06 (0.05, 0.08) | 4.24e-153 |  |
|  | North East | 26.6% (105) | 394 | 0.03 (0.03, 0.04) | 2.87e-172 |  |

### Date of diagnosis

Supplementary Table S1: 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 | Reference |  | 7.54e-286 |
|  | 2011 | 22.3% (1736) | 7781 | 0.77 (0.71, 0.83) | 2.11e-10 |  |
|  | 2012 | 18.8% (1458) | 7755 | 0.61 (0.56, 0.66) | 3.93e-31 |  |
|  | 2013 | 12.9% (909) | 7034 | 0.38 (0.35, 0.42) | 6.81e-91 |  |
|  | 2014 | 10.4% (659) | 6327 | 0.30 (0.27, 0.33) | 6.2e-120 |  |
|  | 2015 | 7.4% (415) | 5619 | 0.20 (0.18, 0.22) | 1.56e-158 |  |
| Sex | Female | 16.9% (2984) | 17664 | Reference |  | 0.432 |
|  | Male | 17.1% (4111) | 23995 | 1.02 (0.97, 1.08) | 0.432 |  |
| Age | 0-14 | 19.4% (348) | 1793 | Reference |  | 0.000251 |
|  | 15-44 | 17.8% (4504) | 25235 | 0.74 (0.65, 0.86) | 4.77e-05 |  |
|  | 45-64 | 15.9% (1434) | 9026 | 0.73 (0.62, 0.85) | 3.52e-05 |  |
|  | 65+ | 14.4% (809) | 5605 | 0.79 (0.68, 0.94) | 0.00563 |  |
| Ethnic group | White | 12.5% (1043) | 8359 | Reference |  | 6.85e-08 |
|  | Black-Caribbean | 25.2% (234) | 928 | 1.20 (1.00, 1.43) | 0.0469 |  |
|  | Black-African | 21.9% (1577) | 7204 | 0.99 (0.89, 1.11) | 0.876 |  |
|  | Black-Other | 17.9% (66) | 369 | 0.75 (0.56, 1.01) | 0.0612 |  |
|  | Indian | 18.0% (1957) | 10848 | 0.80 (0.72, 0.89) | 4.94e-05 |  |
|  | Pakistani | 11.8% (805) | 6806 | 0.86 (0.76, 0.97) | 0.0158 |  |
|  | Bangladeshi | 21.5% (361) | 1680 | 0.94 (0.81, 1.10) | 0.469 |  |
|  | Chinese | 13.4% (66) | 494 | 0.66 (0.49, 0.88) | 0.00525 |  |
|  | Mixed / Other | 19.8% (986) | 4971 | 0.91 (0.81, 1.02) | 0.117 |  |
| UK birth status | Non-UK Born | 18.4% (5696) | 30880 | Reference |  | 0.00227 |
|  | UK Born | 13.0% (1399) | 10779 | 0.87 (0.80, 0.95) | 0.00235 |  |
| Socio-economic status | 1 | 14.4% (2317) | 16131 | Reference |  | 6.01e-14 |
|  | 2 | 19.6% (2469) | 12621 | 0.97 (0.90, 1.04) | 0.394 |  |
|  | 3 | 20.3% (1325) | 6530 | 1.22 (1.12, 1.33) | 5.3e-06 |  |
|  | 4 | 17.0% (645) | 3796 | 1.30 (1.17, 1.45) | 1.87e-06 |  |
|  | 5 | 13.1% (339) | 2581 | 1.42 (1.23, 1.62) | 9.74e-07 |  |
| Public Health England centre | London | 31.0% (5471) | 17658 | Reference |  | 0 |
|  | West Midlands | 3.6% (190) | 5217 | 0.08 (0.07, 0.10) | 4.97e-226 |  |
|  | North West | 7.6% (308) | 4075 | 0.18 (0.15, 0.20) | 6.15e-159 |  |
|  | South East | 3.9% (157) | 4037 | 0.08 (0.07, 0.09) | 4e-193 |  |
|  | Yorkshire and the Humber | 3.2% (99) | 3077 | 0.07 (0.06, 0.09) | 1.51e-137 |  |
|  | East of England | 11.3% (302) | 2662 | 0.26 (0.23, 0.30) | 2.32e-93 |  |
|  | East Midlands | 18.9% (482) | 2548 | 0.51 (0.46, 0.57) | 2.4e-33 |  |
|  | South West | 2.8% (45) | 1595 | 0.06 (0.05, 0.08) | 8.96e-73 |  |
|  | North East | 5.2% (41) | 790 | 0.12 (0.09, 0.17) | 5.45e-38 |  |

### Date of starting treatment and ending treatment

Supplementary Table S2: 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 (41659) | Odds Ratio | P value (Wald) | P value (LRT) |
| Year | 2010 | 5.1% (367) | 7143 | Reference |  | 2.48e-37 |
|  | 2011 | 4.7% (368) | 7781 | 0.92 (0.79, 1.07) | 0.281 |  |
|  | 2012 | 4.0% (314) | 7755 | 0.77 (0.66, 0.90) | 0.00121 |  |
|  | 2013 | 3.8% (265) | 7034 | 0.70 (0.59, 0.82) | 1.7e-05 |  |
|  | 2014 | 2.2% (139) | 6327 | 0.39 (0.32, 0.47) | 1.36e-20 |  |
|  | 2015 | 2.0% (115) | 5619 | 0.36 (0.29, 0.45) | 1.65e-20 |  |
| Sex | Female | 3.4% (608) | 17664 | Reference |  | 0.00223 |
|  | Male | 4.0% (960) | 23995 | 1.18 (1.06, 1.31) | 0.00234 |  |
| Age | 0-14 | 3.6% (64) | 1793 | Reference |  | 1.89e-29 |
|  | 15-44 | 3.1% (774) | 25235 | 0.89 (0.68, 1.17) | 0.384 |  |
|  | 45-64 | 3.4% (310) | 9026 | 0.93 (0.70, 1.25) | 0.628 |  |
|  | 65+ | 7.5% (420) | 5605 | 1.96 (1.49, 2.63) | 3.16e-06 |  |
| Ethnic group | White | 5.8% (486) | 8359 | Reference |  | 0.00077 |
|  | Black-Caribbean | 3.4% (32) | 928 | 0.71 (0.48, 1.02) | 0.0765 |  |
|  | Black-African | 2.8% (203) | 7204 | 0.61 (0.49, 0.76) | 7.46e-06 |  |
|  | Black-Other | 3.3% (12) | 369 | 0.79 (0.42, 1.38) | 0.445 |  |
|  | Indian | 3.4% (371) | 10848 | 0.71 (0.59, 0.86) | 0.000401 |  |
|  | Pakistani | 3.6% (243) | 6806 | 0.63 (0.52, 0.77) | 4.66e-06 |  |
|  | Bangladeshi | 3.1% (52) | 1680 | 0.66 (0.48, 0.90) | 0.0108 |  |
|  | Chinese | 3.8% (19) | 494 | 0.78 (0.46, 1.24) | 0.318 |  |
|  | Mixed / Other | 3.0% (150) | 4971 | 0.70 (0.55, 0.87) | 0.00173 |  |
| UK birth status | Non-UK Born | 3.4% (1045) | 30880 | Reference |  | 0.516 |
|  | UK Born | 4.9% (523) | 10779 | 0.95 (0.81, 1.11) | 0.516 |  |
| Socio-economic status | 1 | 3.8% (611) | 16131 | Reference |  | 0.665 |
|  | 2 | 3.7% (462) | 12621 | 1.05 (0.92, 1.20) | 0.481 |  |
|  | 3 | 3.5% (226) | 6530 | 0.92 (0.78, 1.09) | 0.336 |  |
|  | 4 | 4.1% (154) | 3796 | 0.99 (0.82, 1.20) | 0.934 |  |
|  | 5 | 4.5% (115) | 2581 | 1.01 (0.81, 1.25) | 0.925 |  |
| Public Health England centre | London | 3.1% (551) | 17658 | Reference |  | 2.84e-17 |
|  | West Midlands | 3.8% (198) | 5217 | 1.11 (0.93, 1.32) | 0.229 |  |
|  | North West | 4.3% (176) | 4075 | 1.27 (1.05, 1.52) | 0.0112 |  |
|  | South East | 3.0% (121) | 4037 | 0.87 (0.71, 1.07) | 0.194 |  |
|  | Yorkshire and the Humber | 6.6% (202) | 3077 | 2.03 (1.70, 2.43) | 8.5e-15 |  |
|  | East of England | 3.3% (88) | 2662 | 0.97 (0.77, 1.22) | 0.815 |  |
|  | East Midlands | 3.2% (82) | 2548 | 0.93 (0.73, 1.17) | 0.542 |  |
|  | South West | 6.9% (110) | 1595 | 1.94 (1.54, 2.41) | 5.75e-09 |  |
|  | North East | 5.1% (40) | 790 | 1.44 (1.01, 1.99) | 0.0342 |  |

Supplementary Table S3: 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 | Reference |  | 4.89e-15 |
|  | 2011 | 2.6% (177) | 6855 | 0.88 (0.71, 1.08) | 0.228 |  |
|  | 2012 | 2.4% (164) | 6882 | 0.78 (0.63, 0.97) | 0.0274 |  |
|  | 2013 | 1.5% (97) | 6298 | 0.49 (0.38, 0.63) | 3.05e-08 |  |
|  | 2014 | 1.2% (66) | 5341 | 0.38 (0.29, 0.51) | 5.33e-11 |  |
|  | 2015 | 1.4% (28) | 2059 | 0.47 (0.31, 0.69) | 0.000223 |  |
| Sex | Female | 2.1% (311) | 14630 | Reference |  | 0.506 |
|  | Male | 2.1% (403) | 18976 | 1.05 (0.91, 1.23) | 0.507 |  |
| Age | 0-14 | 2.7% (44) | 1617 | Reference |  | 0.52 |
|  | 15-44 | 2.0% (419) | 21027 | 0.81 (0.59, 1.14) | 0.209 |  |
|  | 45-64 | 2.3% (165) | 7272 | 0.83 (0.59, 1.20) | 0.314 |  |
|  | 65+ | 2.3% (86) | 3690 | 0.74 (0.50, 1.11) | 0.141 |  |
| Ethnic group | White | 2.9% (176) | 6076 | Reference |  | 0.0466 |
|  | Black-Caribbean | 2.8% (21) | 753 | 1.51 (0.91, 2.38) | 0.0888 |  |
|  | Black-African | 1.9% (114) | 6071 | 0.90 (0.66, 1.23) | 0.512 |  |
|  | Black-Other | 2.3% (7) | 306 | 1.34 (0.56, 2.75) | 0.464 |  |
|  | Indian | 1.7% (150) | 8842 | 0.72 (0.55, 0.96) | 0.0235 |  |
|  | Pakistani | 2.5% (140) | 5668 | 0.86 (0.65, 1.13) | 0.282 |  |
|  | Bangladeshi | 1.3% (18) | 1409 | 0.65 (0.37, 1.07) | 0.105 |  |
|  | Chinese | 2.8% (11) | 396 | 1.17 (0.58, 2.14) | 0.643 |  |
|  | Mixed / Other | 1.9% (77) | 4085 | 0.98 (0.70, 1.35) | 0.887 |  |
| UK birth status | Non-UK Born | 1.9% (480) | 25174 | Reference |  | 0.959 |
|  | UK Born | 2.8% (234) | 8432 | 1.01 (0.81, 1.25) | 0.959 |  |
| Socio-economic status | 1 | 2.4% (308) | 13080 | Reference |  | 0.257 |
|  | 2 | 1.7% (170) | 10266 | 1.03 (0.84, 1.26) | 0.752 |  |
|  | 3 | 1.9% (100) | 5265 | 1.09 (0.85, 1.38) | 0.498 |  |
|  | 4 | 2.8% (84) | 2994 | 1.36 (1.04, 1.76) | 0.021 |  |
|  | 5 | 2.6% (52) | 2001 | 1.08 (0.78, 1.47) | 0.619 |  |
| Public Health England centre | London | 0.7% (100) | 14747 | Reference |  | 8.46e-59 |
|  | West Midlands | 4.2% (177) | 4240 | 6.68 (5.16, 8.69) | 2e-46 |  |
|  | North West | 2.7% (88) | 3208 | 4.16 (3.07, 5.63) | 2.21e-20 |  |
|  | South East | 2.5% (79) | 3213 | 3.57 (2.62, 4.84) | 3.41e-16 |  |
|  | Yorkshire and the Humber | 2.8% (67) | 2361 | 4.34 (3.12, 6.01) | 1.06e-18 |  |
|  | East of England | 4.0% (83) | 2098 | 5.88 (4.35, 7.94) | 6.86e-31 |  |
|  | East Midlands | 3.1% (63) | 2039 | 4.77 (3.44, 6.58) | 2.87e-21 |  |
|  | South West | 2.9% (32) | 1122 | 4.22 (2.76, 6.29) | 6.37e-12 |  |
|  | North East | 4.3% (25) | 578 | 6.73 (4.19, 10.44) | 2.16e-16 |  |

### Date of death

Supplementary Table S4: 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.

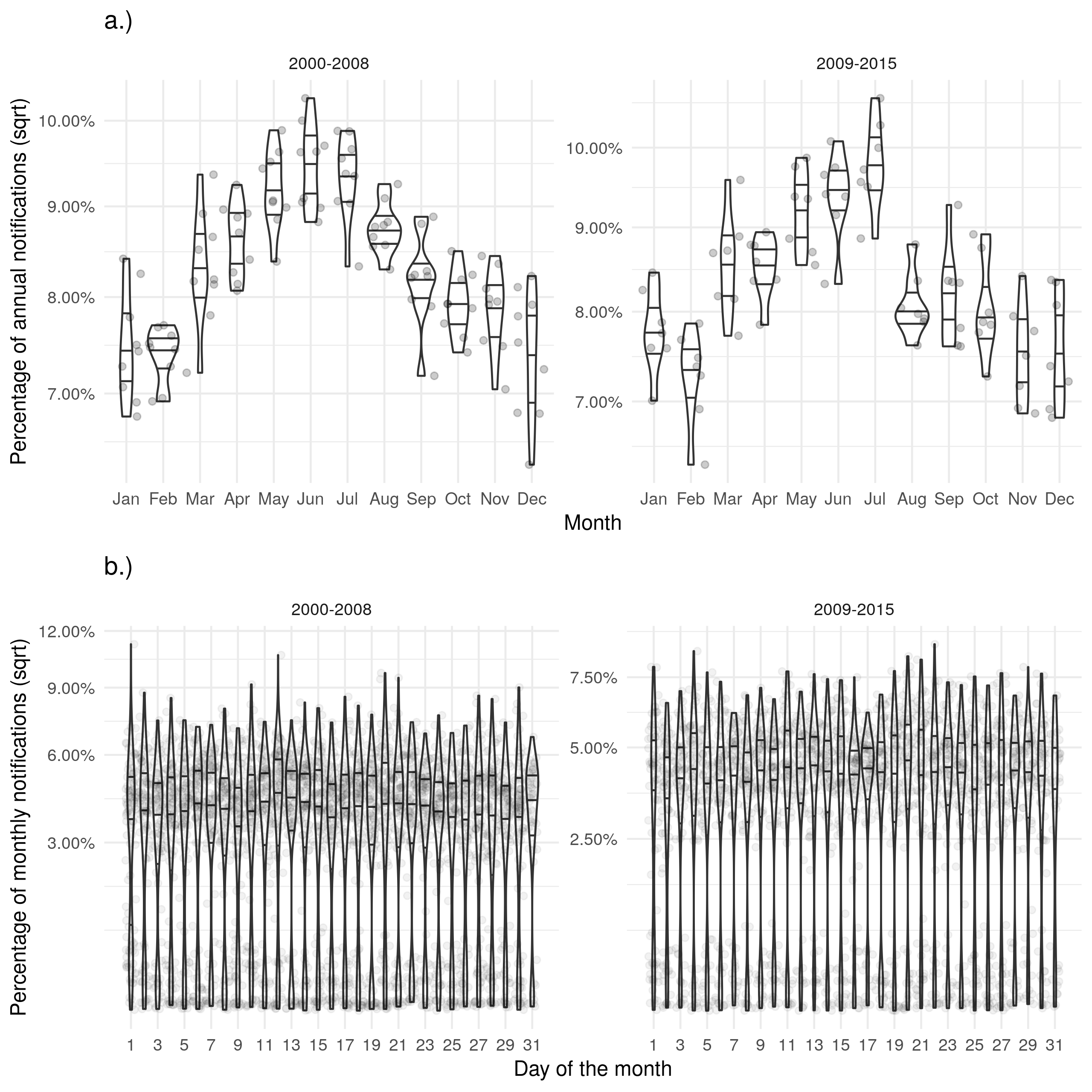
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Category | Missing (N) | Notifications (1883) | Odds Ratio | P value (Wald) | P value (LRT) |
| Year | 2010 | 16.6% (53) | 320 | Reference |  | 0.129 |
|  | 2011 | 15.9% (52) | 327 | 1.02 (0.63, 1.65) | 0.938 |  |
|  | 2012 | 14.5% (51) | 351 | 0.88 (0.54, 1.42) | 0.593 |  |
|  | 2013 | 13.5% (42) | 312 | 0.70 (0.43, 1.16) | 0.169 |  |
|  | 2014 | 9.5% (30) | 317 | 0.55 (0.32, 0.93) | 0.0263 |  |
|  | 2015 | 13.3% (34) | 256 | 0.67 (0.39, 1.14) | 0.14 |  |
| Sex | Female | 14.8% (97) | 657 | Reference |  | 0.569 |
|  | Male | 13.5% (165) | 1226 | 0.91 (0.67, 1.25) | 0.568 |  |
| Age | 0-14 | 10.0% (1) | 10 | Reference |  | 0.799 |
|  | 15-44 | 15.7% (31) | 198 | 1.86 (0.26, 38.77) | 0.596 |  |
|  | 45-64 | 14.6% (68) | 465 | 1.85 (0.26, 38.20) | 0.598 |  |
|  | 65+ | 13.4% (162) | 1210 | 2.11 (0.30, 43.43) | 0.521 |  |
| Ethnic group | White | 11.1% (102) | 920 | Reference |  | 0.9 |
|  | Black-Caribbean | 21.7% (10) | 46 | 0.90 (0.35, 2.18) | 0.817 |  |
|  | Black-African | 20.1% (27) | 134 | 0.92 (0.45, 1.92) | 0.833 |  |
|  | Black-Other | 20.0% (1) | 5 | 0.52 (0.03, 4.31) | 0.586 |  |
|  | Indian | 17.4% (64) | 367 | 0.90 (0.49, 1.70) | 0.747 |  |
|  | Pakistani | 8.0% (20) | 249 | 0.62 (0.30, 1.29) | 0.204 |  |
|  | Bangladeshi | 22.7% (10) | 44 | 0.85 (0.33, 2.12) | 0.731 |  |
|  | Chinese | 14.3% (3) | 21 | 0.80 (0.16, 3.23) | 0.772 |  |
|  | Mixed / Other | 25.8% (25) | 97 | 1.15 (0.55, 2.39) | 0.711 |  |
| UK birth status | Non-UK Born | 16.6% (167) | 1004 | Reference |  | 0.796 |
|  | UK Born | 10.8% (95) | 879 | 1.08 (0.61, 1.92) | 0.796 |  |
| Socio-economic status | 1 | 11.4% (79) | 695 | Reference |  | 0.912 |
|  | 2 | 18.3% (86) | 470 | 0.87 (0.59, 1.29) | 0.499 |  |
|  | 3 | 16.2% (48) | 296 | 1.04 (0.66, 1.64) | 0.87 |  |
|  | 4 | 12.7% (30) | 237 | 1.02 (0.60, 1.71) | 0.937 |  |
|  | 5 | 10.3% (19) | 185 | 0.87 (0.46, 1.59) | 0.651 |  |
| Public Health England centre | London | 37.6% (201) | 534 | Reference |  | 1.92e-57 |
|  | West Midlands | 2.3% (7) | 305 | 0.04 (0.02, 0.07) | 1.61e-16 |  |
|  | North West | 7.0% (16) | 228 | 0.12 (0.07, 0.21) | 5.23e-13 |  |
|  | South East | 4.8% (10) | 208 | 0.08 (0.04, 0.15) | 2.25e-13 |  |
|  | Yorkshire and the Humber | 3.6% (6) | 168 | 0.06 (0.02, 0.12) | 4.81e-11 |  |
|  | East of England | 8.5% (11) | 130 | 0.14 (0.07, 0.26) | 6.32e-09 |  |
|  | East Midlands | 1.9% (3) | 156 | 0.03 (0.01, 0.08) | 2.58e-09 |  |
|  | South West | 6.7% (7) | 105 | 0.11 (0.04, 0.23) | 7.77e-08 |  |
|  | North East | 2.0% (1) | 49 | 0.03 (0.00, 0.15) | 0.000694 |  |

### Cause of death

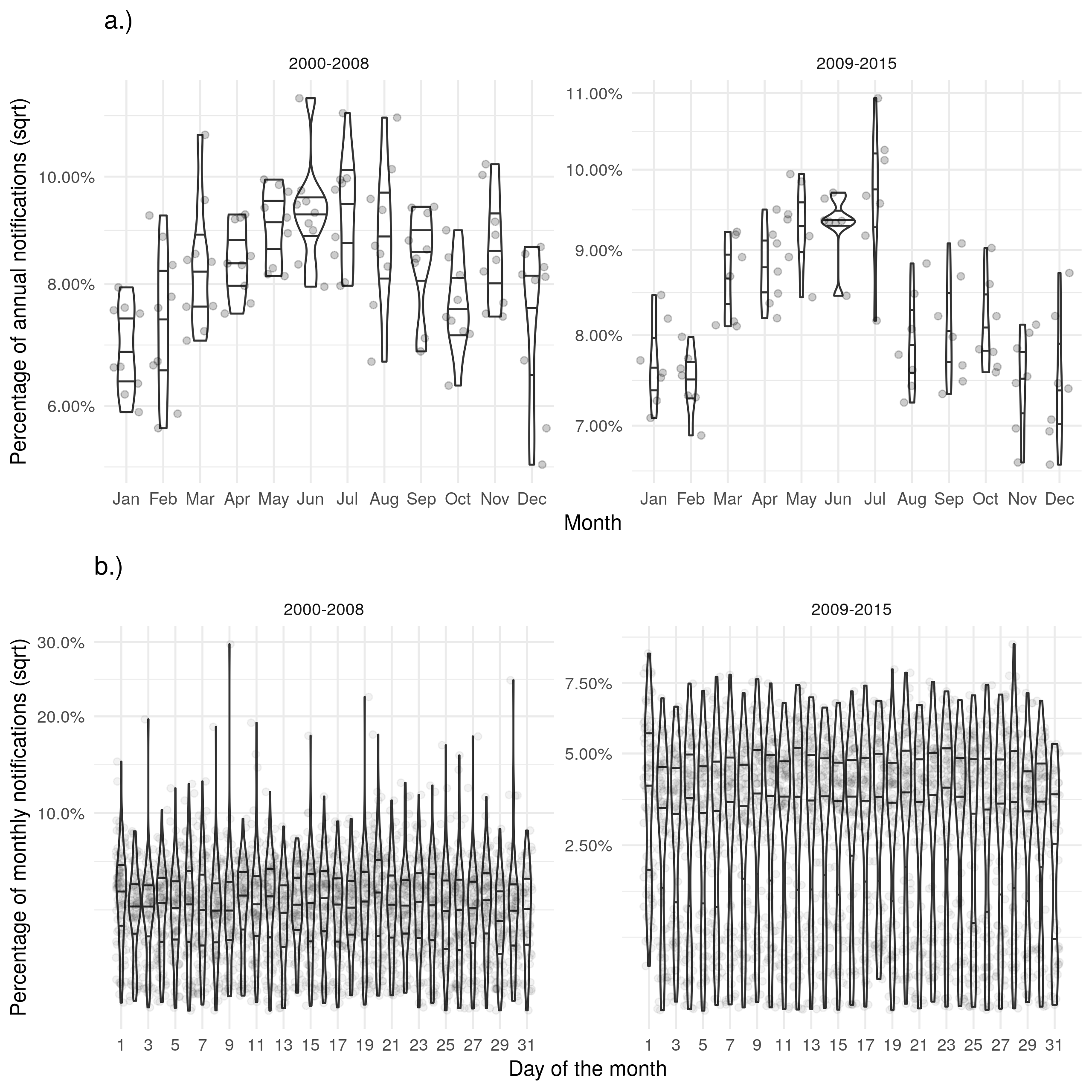
Supplementary Table S5: Results from a logistic regression model with data completeness (Complete/Missing) for cause of death as an outcome, adjusted for: year, sex, age (grouped as 0-14 year olds, 15-65 year olds and 65+), ethnic group, UK birth status and socio-economic status (national quintiles). 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 | Reference |  | 0.576 |
|  | 2011 | 45.6% (149) | 327 | 0.99 (0.71, 1.37) | 0.944 |  |
|  | 2012 | 45.3% (159) | 351 | 0.94 (0.68, 1.29) | 0.694 |  |
|  | 2013 | 43.9% (137) | 312 | 0.94 (0.67, 1.30) | 0.693 |  |
|  | 2014 | 44.8% (142) | 317 | 0.86 (0.62, 1.20) | 0.379 |  |
|  | 2015 | 38.7% (99) | 256 | 0.74 (0.52, 1.05) | 0.0933 |  |
| Sex | Female | 44.7% (294) | 657 | Reference |  | 0.763 |
|  | Male | 43.7% (536) | 1226 | 0.97 (0.79, 1.19) | 0.763 |  |
| Age | 0-14 | 50.0% (5) | 10 | Reference |  | 0.14 |
|  | 15-44 | 35.4% (70) | 198 | 0.69 (0.17, 2.82) | 0.6 |  |
|  | 45-64 | 43.0% (200) | 465 | 1.02 (0.25, 4.11) | 0.977 |  |
|  | 65+ | 45.9% (555) | 1210 | 1.03 (0.25, 4.13) | 0.965 |  |
| Ethnic group | White | 48.2% (443) | 920 | Reference |  | 0.00768 |
|  | Black-Caribbean | 21.7% (10) | 46 | 0.47 (0.20, 0.99) | 0.0565 |  |
|  | Black-African | 45.5% (61) | 134 | 1.78 (1.04, 3.03) | 0.0347 |  |
|  | Black-Other | 20.0% (1) | 5 | 0.70 (0.03, 5.37) | 0.761 |  |
|  | Indian | 35.7% (131) | 367 | 0.87 (0.56, 1.36) | 0.545 |  |
|  | Pakistani | 49.4% (123) | 249 | 1.33 (0.84, 2.11) | 0.224 |  |
|  | Bangladeshi | 27.3% (12) | 44 | 0.82 (0.36, 1.78) | 0.625 |  |
|  | Chinese | 52.4% (11) | 21 | 1.70 (0.64, 4.55) | 0.284 |  |
|  | Mixed / Other | 39.2% (38) | 97 | 1.37 (0.78, 2.41) | 0.275 |  |
| UK birth status | Non-UK Born | 40.1% (403) | 1004 | Reference |  | 0.426 |
|  | UK Born | 48.6% (427) | 879 | 1.17 (0.79, 1.74) | 0.427 |  |
| Socio-economic status | 1 | 43.7% (304) | 695 | Reference |  | 0.168 |
|  | 2 | 40.0% (188) | 470 | 1.26 (0.97, 1.64) | 0.0842 |  |
|  | 3 | 42.9% (127) | 296 | 1.20 (0.89, 1.63) | 0.235 |  |
|  | 4 | 49.8% (118) | 237 | 1.43 (1.03, 1.98) | 0.0322 |  |
|  | 5 | 50.3% (93) | 185 | 1.37 (0.96, 1.97) | 0.0841 |  |
| Public Health England centre | London | 25.3% (135) | 534 | Reference |  | 1.1e-20 |
|  | West Midlands | 48.9% (149) | 305 | 3.01 (2.19, 4.14) | 1.17e-11 |  |
|  | North West | 61.8% (141) | 228 | 4.82 (3.39, 6.91) | 5.11e-18 |  |
|  | South East | 46.6% (97) | 208 | 2.36 (1.65, 3.37) | 2.23e-06 |  |
|  | Yorkshire and the Humber | 44.0% (74) | 168 | 2.23 (1.52, 3.26) | 3.55e-05 |  |
|  | East of England | 46.2% (60) | 130 | 2.36 (1.56, 3.55) | 4e-05 |  |
|  | East Midlands | 60.3% (94) | 156 | 4.56 (3.09, 6.77) | 3.07e-14 |  |
|  | South West | 53.3% (56) | 105 | 3.09 (1.97, 4.88) | 1.06e-06 |  |
|  | North East | 49.0% (24) | 49 | 2.84 (1.54, 5.25) | 0.000831 |  |

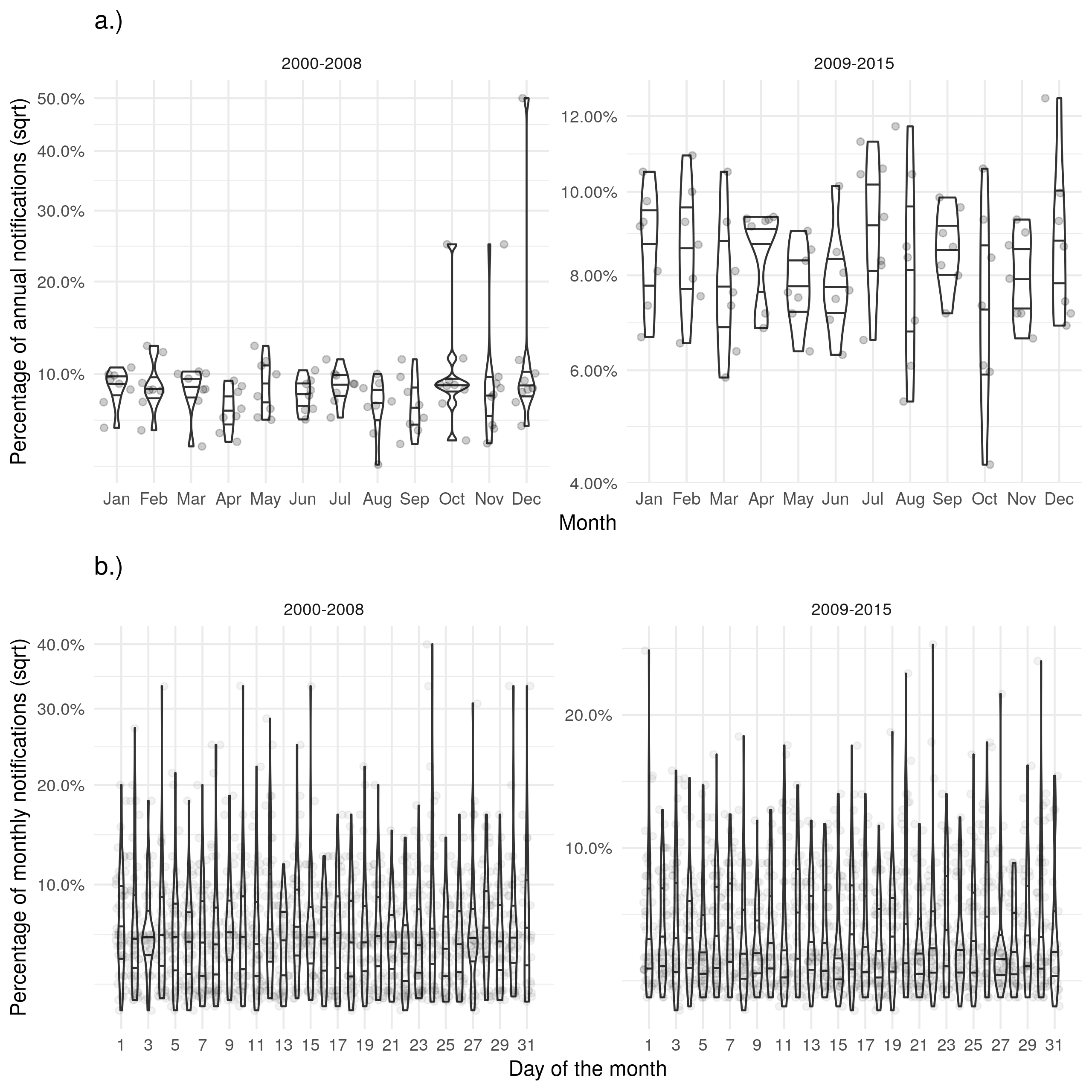
### Assessing temporal biases in reporting



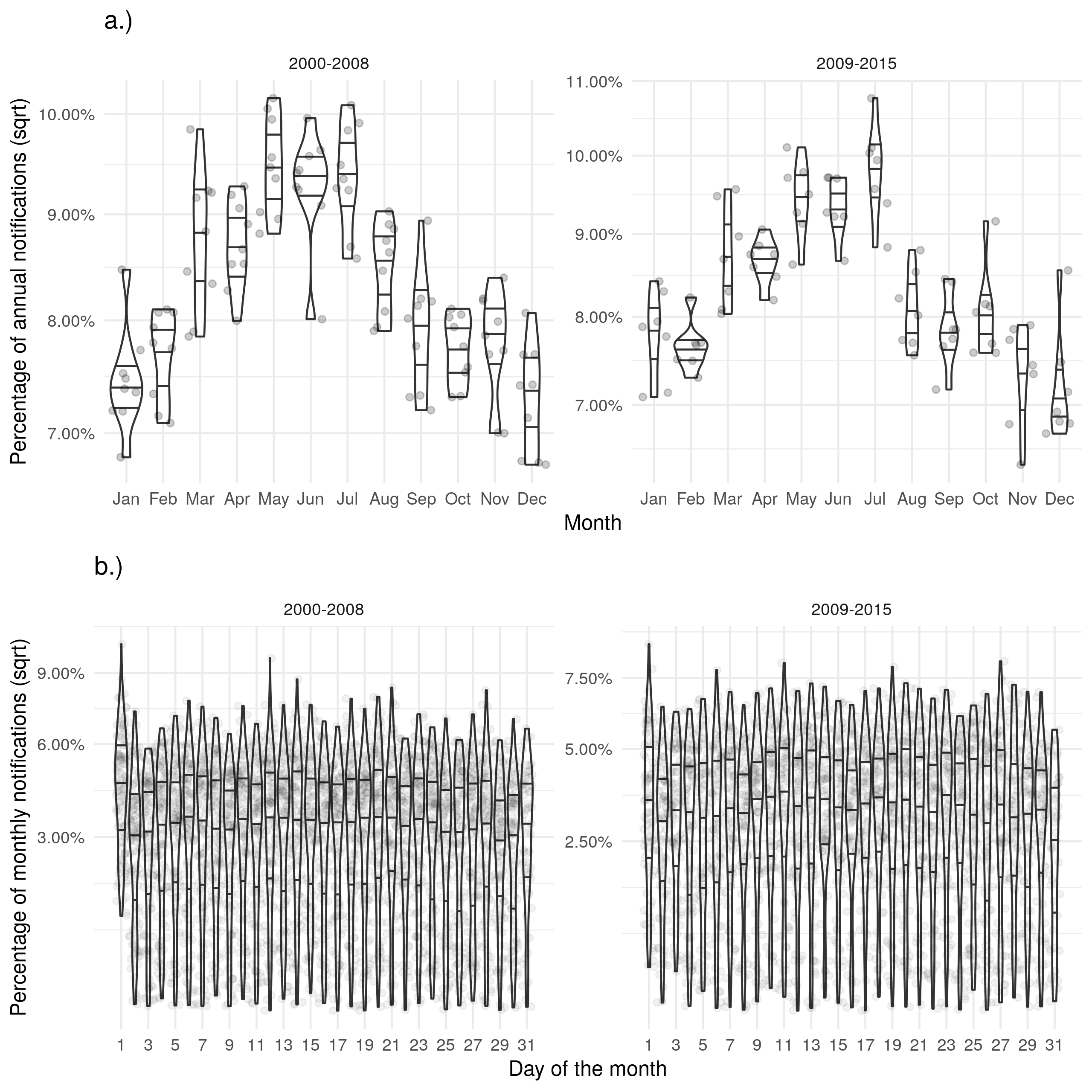
Supplementary Figure S2: a.) Shows the proportion of cases notified in a given month for each year with evidence of a seasonal peak in June. b.) Shows the proportion of cases notified on a given day for each month with a near uniform distribution. Stratifying both figures based on the introduction of the web-based ETS gives little evidence for any change in these trends.



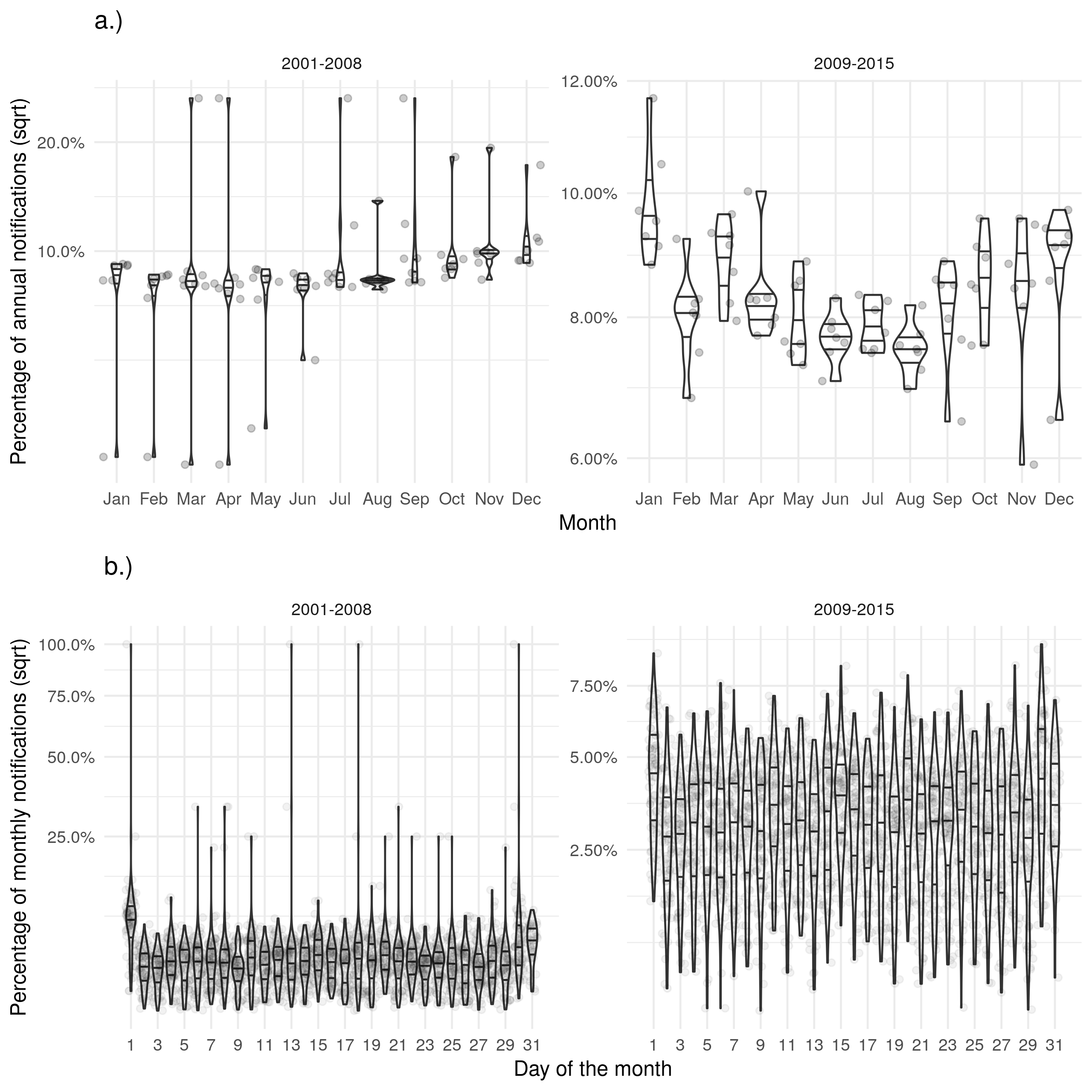
Supplementary Figure S3: a.) Shows the proportion of cases with a diagnosis in a given month for each year. b.) Shows the proportion of cases with a diagnosis on a given day for each month. Trends for the date of diagnosis were similar to those seen for notifications.



Supplementary Figure S4: a.) Shows the proportion of cases that died in a given month for each year. b.) Shows the proportion of cases that died on a given day for each month. Trends for the date of death were similar to those seen for notifications but there was a reduction in the strength of the observed seasonality.



Supplementary Figure S5: a.) Shows the proportion of cases starting treatment in a given month for each year. b.) Shows the proportion of cases starting treatment on a given day for each month. Trends for the date of starting treatment were similar to those seen for notifications.



Supplementary Figure S6: a.) Shows the proportion of cases ending treatment in a given month for each year with a peak in December. b.) Shows the proportion of cases ending treatment on a given day for each month with some evidence of a bias in reporting on the first of the month. Uncertainty reduced after the introduction of the web-based ETS and the level of bias on the first of the month reduced.