Epidemiol. Infect. (2008), **136**, 1606–1616. © 2008 Cambridge University Press
doi:10.1017/S0950268808000496 Printed in the United Kingdom

Record-linkage and capture–recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999–2002

N. A. H. VAN HEST[1,2]*, A. STORY[3], A. D. GRANT[4], D. ANTOINE[3], J. P. CROFTS[3] AND J. M. WATSON[3]

[1] *Tuberculosis Control Section, Division of Infectious Disease Control, Municipal Public Health Service Rotterdam-Rijnmond, Rotterdam, The Netherlands*
[2] *Department of Public Health, Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands*
[3] *Respiratory Diseases Department, Centre for Infections, Health Protection Agency, London, UK*
[4] *Statistics, Modelling and Bioinformatics Department, Centre for Infections, Health Protection Agency, London, UK*

(*Accepted 12 February 2008; first published online 17 March 2008*)

SUMMARY

In 1999 the Enhanced Tuberculosis Surveillance (ETS) system was introduced in the United Kingdom to strengthen surveillance of tuberculosis (TB). The aim of this study was to assess the use of record-linkage and capture–recapture methodology for estimating the completeness of TB reporting in England between 1999 and 2002. Due to the size of the TB data sources sophisticated record-linkage software was required and the proportion of false-positive cases among unlinked hospital-derived TB records was estimated through a population mixture model. This study showed that record-linkage of TB data sources and cross-validation with additional TB-related datasets improved data quality as well as case ascertainment. Since the introduction of ETS observed completeness of notification in England has increased and the results were consistent with expected levels of under-notification. Completeness of notification estimated by a log-linear capture–recapture model was highly inconsistent with prior estimates and the validity of this methodology was further examined.

INTRODUCTION

Since 1987, a rise in notifications of tuberculosis (TB) has been observed in England [1]. This increase is believed to be real, reflecting an increase in diagnoses of TB, rather than an artefact due to improved reporting [2]. Nevertheless, it has been estimated that between 7% and 27% of cases of TB in the United Kingdom are unnotified [3]. In 1999, a revised national routine surveillance system for TB, Enhanced Tuberculosis Surveillance (ETS), was introduced to improve the completeness of reporting as well as the information on reported cases [4]. The aim of this study was to estimate the annual incidence of TB in England and assess the completeness of reporting between 1999 and 2002 using record-linkage and capture–recapture methodology.

The accuracy and completeness of surveillance data can be increased through record-linkage between datasets of cases reported from different sources [5–8]. This is carried out routinely for cases in ETS by linking notifications with reports of *Mycobacterium tuberculosis* isolates from the reference laboratories in the UK Mycobacterial Network (MycobNet). The number of cases missed can then be estimated using the overlap between the two data

* Author for correspondence: Rob van Hest, M.D., Ph.D., Consultant Tuberculosis Control Physician/Epidemiologist, Tuberculosis Control Section, Division of Infectious Disease Control, Municipal Public Health Service Rotterdam-Rijnmond, PO Box 70032, 3000 LP Rotterdam, The Netherlands.
(Email: vanhestr@ggd.rotterdam.nl)

sources through capture–recapture analysis [9]. The preferred capture–recapture method entails log-linear modelling of at least three linked data sources [10–13]. The completeness of the different data sources can be assessed by comparison with the case ascertainment, i.e. the total number of patients observed in at least one data source, or the estimated total number of cases. Capture–recapture analysis has been used to evaluate surveillance systems of various infectious diseases in the United Kingdom [14–16]. The same methodology has been applied to TB surveillance in studies in both the United Kingdom and elsewhere [17–20].

METHODS

Case definition and data sources

For the purpose of estimating the number of unobserved TB cases, i.e. cases not registered ('observed') in at least one of the linked registers studied, we defined as eligible for inclusion those active TB cases first reported to one or more of three data sources in the four years, 1 January 1999 to 31 December 2002. The three data sources were:

(1) Cases notified through ETS (Notification).
(2) Cases with *M. tuberculosis* complex isolates reported to MycobNet (Laboratory).
(3) Cases admitted to National Health Service hospitals with a first or secondary hospital discharge code of TB [International Classification of Disease (ICD-10) code A15-A19] provided from Hospital Episode Statistics (Hospital).

Two other data sources used for cross-validation will be mentioned later. An interval of more than 1 year between entries in each of the data sources was considered as a separate episode of disease. To correct for delays in case reporting and mycobacteriological confirmation, records 3 months before and 3 months after the study period were also examined.

Record-linkage

Duplicate entries within each of the three data sources were excluded. Hospital records were linked to the previously linked Notification and Laboratory records. Record-linkage software developed by the Centre for Infections establishes a likelihood of association between two records based on a core set of identifiers (date of birth, age, full postcode and sex of the patient and proximity of date of notification, initial mycobacterial isolate or hospital admission). It allows for visual inspection of available additional information on geographical location, site of disease, ethnicity and smear, culture or histopathology results (when performed). All cases with incomplete or missing information on both the date of birth and age were labelled as 'insufficient identifiers' and excluded.

The software allocates an *a priori* determined maximum number of points to each core identifier for complete agreement, reflecting the perceived relative importance of that identifier. Record pairs with full agreement of all core identifiers are automatically assigned as true links. Points are deducted proportionally to the presumed loss of information for increasing deviation from perfect linkage of each identifier to generate an aggregate score, reflecting the likelihood of association between two patient records. All categories of candidate links other than automatically assigned links were visually inspected and either accepted or rejected. Linked cases were allocated to the year of first known date of notification, culture-confirmation or hospital admission.

False-positive records and correction

All laboratory-confirmed cases reported through MycobNet were assumed true TB cases, as previously found in a local capture–recapture study in England [17]. Notification and Hospital records not linked with Laboratory could potentially include three groups of false-positive records:

(1) Cases ultimately diagnosed with an infection with Mycobacteria other than tuberculosis (MOTT).
(2) Cases with a final diagnosis other than TB or MOTT infection.
(3) Cases misclassified or miscoded.

The proportion of unlinked Hospital cases attributable to MOTT infection was estimated by linking Hospital data from 2003 with a MOTT database which began in that same year and used to correct the number of unlinked Hospital cases in all years under study using a formula explained below, assuming the annual proportion is similar.

In order to estimate the proportion of cases with a final diagnosis other than TB or MOTT infection Notification cases unknown to Laboratory were linked with Treatment Outcome Monitoring (TOM) data, containing data on Notification cases with a final diagnosis other than TB. At the time of this study

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press

Table 1. *Initial annual number of cases in each of the tuberculosis data sources before record-linkage and the proportion of records excluded from the study because of 'insufficient identifiers' (incomplete or missing date of birth or age)*

Year/data source	Notification (% excluded)	Laboratory (% excluded)	Hospital (% excluded)
1999	5784 (2·2)	3936 (3·9)	4361 (6·4)
2000	6101 (2·1)	3940 (6·7)	4247 (8·0)
2001	6571 (1·6)	4113 (3·7)	4268 (5·1)
2002	6615 (1·2)	4336 (4·3)	4618 (8·3)

TOM data were only available for 2001. The proportion of false-positive Notification cases found was used to correct all years under study assuming the annual proportion is similar.

Previous capture–recapture studies on TB identified a considerable proportion of remaining false-positives among unlinked Hospital cases after examining individual patients' medical files [17, 19, 20]. Examining individual patients' medical files was not feasible due to the scale of this study. We estimated the proportion of these remaining false-positive cases through a population mixture model. Briefly, we used 40 covariates (number of admission days, number of admissions during the TB episode, rank number of TB diagnosis (14 possible positions) and 37 different ICD-10 TB diagnosis codes) and the incidence of Hospital records linked with Notification and/or Laboratory to estimate the number of true TB cases among unlinked records, under the assumption that all linked Hospital cases are true TB cases and unlinked Hospital cases are a mixture of true and false-positive TB cases. The best-fitting logistic regression model calculates for every Hospital case the predicted Bernoulli parameter P (reflecting the probability of being a true TB patient) from the covariates. Linked and unlinked Hospital cases have characteristic frequency distributions of values P as 'signatures'. After standardization we used these signature curves to separate the mixture of unlinked Hospital cases, assuming the subpopulation of true TB cases has a similar signature curve to linked Hospital cases and the false-positive TB cases have a different signature curve (population mixture model available in online Appendix). The corrected annual number of true TB cases known only to Hospital was calculated using the formula:

$$N_{\text{final}} = (Prop_{\text{true}} \times N_{\text{original}}) \times (1 - Prop_{\text{MOTT}}),$$

where N_{original} and N_{final} denote the number of unlinked Hospital cases before and after deducting the projected annual proportion of MOTT infection cases and the estimated annual proportion of remaining false-positive TB cases by logistic regression respectively, $Prop_{\text{true}}$ the estimated annual proportion of true TB cases by logistic regression and $Prop_{\text{MOTT}}$ the projected annual proportion of MOTT infection cases.

Observed source-specific coverage rates were defined as the number of TB cases in each data source divided by the case ascertainment, expressed as a percentage.

Capture–recapture analysis

The annual and total number of unobserved TB cases was estimated on the basis of the final distribution of observed cases over the three data sources. The independence of data sources and other assumptions underlying capture–recapture analysis have been described previously [21]. Interdependencies between the three TB data sources are probable, causing possible bias in two-source capture–recapture estimates. Three-source log-linear capture–recapture analysis was employed to take possible interdependencies into account [17, 19, 20]. Estimated source-specific coverage rates were defined as the number of TB cases in each data source divided by the estimated number of TB cases by capture–recapture analysis, expressed as a percentage.

RESULTS

Table 1 shows the initial annual number of cases in each of the TB data sources before record-linkage and the proportion of records excluded from the study because of 'insufficient identifiers'. The proportion of

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press

excluded records is small for all three TB data sources and consistent over the years examined.

The record-linkage process designated 10 539 of the 16 272 (64·8%) Hospital cases as links while 5733 cases (35·2%) remained unlinked. After visual inspection of the identifiers, 94·9% of all records allocated ⩾3000 points by the record-linkage software (from a maximum of 4000 points) were accepted as true links.

Table 2 shows the number, proportion and distribution of TB cases over the data sources after record-linkage, the corrections for estimated and projected proportions of false-positive cases and the final distribution. Record-linkage between the TOM and Notification data sources for 2001 identified 4·1% of cases known only to Notification and 4·1% of cases known to Notification and Hospital with a final diagnosis of not TB or MOTT infection. Record-linkage between Hospital records and the MOTT database for 2003 identified 3·8% of Hospital cases as having MOTT infection. The population mixture model gave a range of the proportion of true TB cases known only to Hospital of 0–38%, with an upper 95% confidence limit of 50%. The value 28% (uncertainty interval 19–50%) was chosen because of good support by the model and prior expectation based on national and international reports. The total estimated and projected percentage of false-positive cases among all Hospital cases was 26·7% (4352/16 272). Since 2000 the proportion of cases known only to Notification or Laboratory has fallen each year and the number of Notification cases linked to Laboratory or Laboratory and Hospital has increased. Of all 28 678 TB cases included in this study, 2990 (10·4%) were identified in the Laboratory data source with a positive culture for *M. tuberculosis* but unnotified.

Table 3 shows the annual and overall observed number of TB cases after record-linkage and correction for false-positive records. The overall observed source-specific coverage rates of notified, culture-confirmed and hospitalized TB cases were 84·1%, 54·3% and 41·6% respectively. Overall observed under-notification was 15·9%. The annual observed Notification-specific coverage rate increased from 81·8% to 86·7% between 1999 and 2002. The annual observed Laboratory and Hospital source-specific coverage rates were relatively stable over the study period.

Table 4 shows the annual and overall estimated number of unobserved and total TB cases after capture–recapture analysis. For all estimates the saturated log-linear model was preferred based on the Akaike Information Criterion (AIC), as none of the other, more parsimonious, models produced a negative AIC [9, 12]. The overall estimated completeness of case ascertainment was 66·7% (28 678/42 969). The overall estimated source-specific coverage rates of notified, culture-confirmed and hospitalized TB cases were 56·2%, 36·2% and 27·7% respectively. Overall estimated under-notification was 43·8%. The number of unobserved TB cases fell every year. The annual estimated Notification-specific coverage rates between 1999 and 2002 were 48·1%, 51·1%, 59·0% and 66·5% respectively. None of the approximate confidence intervals include expected values of under-notification. We assessed that the interval between the administrative reporting dates used in this study instead of the date of actual disease onset could result in a capture–recapture overestimate of the number of unobserved cases of 1·5% (model available from the authors).

DISCUSSION

Main findings

This study shows that record-linkage of TB data sources and cross-validation with additional TB-related datasets improves data accuracy as well as completeness of case ascertainment. For large TB data sources sophisticated record-linkage software is required and a population mixture model to estimate the proportion of false-positive TB cases among unlinked hospital cases. Since the introduction of ETS the annual observed completeness of notification has increased. However, 10·4% of the observed TB cases in this study were laboratory-confirmed but unnotified. The overall observed under-notification of 15·9% is consistent with previous reports. The 43·8% overall under-notification estimated by a saturated log-linear capture–recapture model is highly inconsistent with previous reports and the validity needs further examination [3, 17].

Under-notification

In this study an interval of more than 1 year between entries in each of the data sources was considered to indicate a separate episode of disease. Although the number of patients with multiple episodes of TB according to this definition was limited, possibly

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press

Table 2. *Number, proportion and distribution of tuberculosis cases between the data sources after record-linkage in England between 1999 and 2002 and correction for estimated and projected proportions of false-positive records*

Year/data source	NOT *N* (%)	LAB *N* (%)	HOSP *N* (%)	NOT + LAB *N* (%)	LAB + HOSP *N* (%)	HOSP + NOT *N* (%)	NOT + LAB + HOSP *N* (%)	*N* total
1999								
Record linkage results*	1764 (21·8)	678 (8·4)	1649 (20·4)	1575 (19·4)	111 (1·4)	903 (11·1)	1417 (17·5)	8097
Correction for TOM†	1692 (21·2)	678 (8·5)	1649 (20·6)	1575 (19·7)	111 (1·4)	866 (10·8)	1417 (17·7)	7988
Correction for MOTT and false-positive hospital records‡	1692 (25·0)	678 (10·0)	444 (6·5)	1575 (23·2)	111 (1·6)	866 (12·8)	1417 (20·9)	6783
2000								
Record linkage results*	2205 (26·8)	795 (9·6)	1313 (16·0)	1324 (16·1)	148 (1·8)	1037 (12·6)	1409 (17·1)	8231
Correction for TOM†	2115 (26·1)	795 (9·8)	1313 (16·2)	1324 (16·3)	148 (1·8)	994 (12·3)	1409 (17·4)	8098
Correction for MOTT and false-positive hospital records‡	2115 (29·6)	795 (11·1)	354 (5·0)	1324 (18·6)	148 (2·1)	994 (13·9)	1409 (19·7)	7139
2001								
Record linkage results*	2148 (25·2)	527 (6·2)	1411 (16·6)	1790 (21·0)	109 (1·3)	996 (11·7)	1534 (18·0)	8515
Correction for TOM†	2060 (24·6)	527 (6·3)	1411 (16·8)	1790 (21·3)	109 (1·3)	955 (11·4)	1534 (18·3)	8386
Correction for MOTT and false-positive hospital records‡	2060 (28·0)	527 (7·2)	380 (5·2)	1790 (24·3)	109 (1·5)	955 (13·0)	1534 (20·9)	7355
2002								
Record linkage results*	1992 (23·4)	478 (5·6)	1360 (16·0)	1814 (21·3)	144 (1·7)	1016 (11·9)	1715 (20·1)	8519
Correction for TOM†	1910 (22·8)	478 (5·7)	1360 (16·2)	1814 (21·6)	144 (1·7)	974 (11·6)	1715 (20·4)	8395
Correction for MOTT and false-positive hospital records‡	1910 (25·8)	478 (6·5)	366 (4·9)	1814 (24·5)	144 (1·9)	974 (13·2)	1715 (23·2)	7401

NOT, Notification data source; LAB, Laboratory data source; HOSP, Hospital data source.

* After correction for multiple links and exclusion of patient records with insufficient identifiers.

† After correction for estimated proportion of cases with diagnosis other than tuberculosis identified in the Treatment Outcome Monitoring (TOM) dataset.

‡ After correction for estimated proportion of unlinked Hospital cases with diagnosis of Mycobacteria other than tuberculosis (MOTT) infection and false-positive hospital records.

Table 3. *Annual and overall observed number of tuberculosis cases after record-linkage and correction for false-positive records and annual and total observed source-specific coverage rates of notified, culture-confirmed and hospitalized tuberculosis cases in England between 1999 and 2002*

	Observed number of tuberculosis cases in at least one data source (case ascertainment)	Notification		Laboratory		Hospital	
Year	Number (UI)*	Number	Percentage (UI)	Number	Percentage (UI)	Number	Percentage (UI)
1999	6783 (6640–7132)	5550	81·8 (77·8–83·6)	3781	55·7 (53·0–56·9)	2838	41·8 (40·6–44·7)
2000	7139 (7025–7417)	5842	81·8 (78·8–83·2)	3676	51·5 (49·6–52·3)	2905	40·7 (39·7–42·9)
2001	7355 (7233–7654)	6339	86·2 (82·8–87·6)	3960	53·8 (51·7–54·7)	2978	40·5 (39·5–42·8)
2002	7401 (7284–7689)	6413	86·7 (83·4–88·0)	4151	56·1 (54·0–57·0)	3199	43·2 (42·3–45·4)
All	28 678 (28 182–29 892)	24 144	84·1 (80·7–85·6)	15 568	54·3 (52·1–55·3)	11 920	41·6 (40·5–43·9)

* UI, Uncertainty interval.

Table 4. *Annual and overall estimated number of unobserved and total tuberculosis (TB) cases by saturated log-linear capture–recapture model in England between 1999 and 2002 (after using a proportion of 28% of true TB cases known only to Hospital in the corrections for false-positive cases)*

Year	Estimated unobserved number of TB cases by the saturated log-linear capture–recapture model (95% ACI)	Estimated total number of TB cases by the saturated log-linear capture–recapture model (95% ACI)
1999	4756 (3717–6087)	11 539 (10 500–12 870)
2000	4294 (3411–5405)	11 433 (10 550–12 544)
2001	3387 (2634–4356)	10 742 (9989–11 711)
2002	2246 (1775–2843)	9647 (9176–10 249)
All	14 291 (12 682–16 105)	42 969 (41 360–44 783)

ACI, Approximate confidence interval.

including a small number of patients whose disease at diagnosis warranted more than 1 year's therapy, an extended definition, e.g. a 2-year interval, would (very) slightly reduce the number of observed cases in the Laboratory and/or Hospital registers and therefore (very) slightly increase the completeness of Notification. Increasing completeness of Notification could be influenced by improved data accuracy and record-linkage over the years.

In comparison with similar studies in Italy and The Netherlands [19, 20], the observed completeness of the Notification register in England is similar. The estimated completeness of notification is low in England due to the high estimated total number of TB patients and highly inconsistent with the results in The Netherlands and Italy, probably due to greater violation of the capture–recapture assumptions. The completeness of the Laboratory register is lower than the completeness of the Notification register due to the proportion of culture-negative TB cases. The observed completeness of the Laboratory register in England is lower compared to The Netherlands but higher compared to Italy, indicating efforts to establish bacteriological confirmation of the diagnosis in England and The Netherlands, whereas in Italy apparently more patients are treated on empirical grounds. In England and The Netherlands, the observed completeness of the Hospital register is low, probably reflecting common policies of preferably treating TB patients as outpatients, including isolation at home for infectious patients. The high proportion of hospitalized TB patients in Italy suggests a system of (initial) clinical analysis, diagnosis, treatment or isolation.

An overall observed under-notification of 15·9% suggests that in England about 1100 TB patients may be unnotified annually of which the majority (2990/4534) is culture-confirmed, representing 10·4% of all TB cases. This reflects the most serious public health aspect of under-notification as culture-confirmed TB cases are assumed true cases and are potentially infectious. Failure to notify laboratory-confirmed cases jeopardizes control measures, including contact tracing. The capture–recapture studies in Italy and The Netherlands show proportions of unnotified culture-confirmed TB cases of 5·5% and 4·9% respectively [19, 20]. The proportion of unnotified culture-confirmed TB cases in England could be an overestimate resulting from possible imperfect record-linkage or, despite our assumption, remaining false-positive records in the Laboratory data source.

Limitations due to imperfect record-linkage and false-positive records

Imperfect record-linkage causes misclassification and results in observed and estimated numbers of TB cases being too low or too high. Our data show that 94·9% of the linked cases have a high likelihood of association score of $\geqslant 3000$ points, and only 5·1% with such a score were unlinked. This indicates that in only a minority of candidate links could an error of classification have occurred. This fulfils our purpose of record-linkage resulting in unbiased numbers in each category, with possibly some balanced misclassification. The relatively stable annual proportional distribution of TB cases and the decreasing annual proportion of unlinked Notification and Laboratory cases give further confidence in the record-linkage software and procedure.

A low positive predictive value of TB data sources results in observed and estimated numbers of TB cases being too high. Lack of specificity of data sources used in capture–recapture studies as a limitation to the validity of this method has previously been described [22, 23]. Not all TB cases are defined by gold-standard laboratory confirmation and diagnosis can be based on a clinical intention to treat. The three data sources used employ different case definitions, with consequent variations in specificity. We demonstrated by cross-validation with additional datasets that failure to de-notify or re-classify patients with a final diagnosis of not TB occurs which will also reduce the positive predictive value.

The population mixture model estimates a proportion of 72% remaining false-positive cases among unlinked Hospital cases, contributing to 26·7% false-positive cases among all Hospital cases, and resulting in a final average proportion of true unlinked Hospital cases of 5·4%. These results are in good agreement with comparable record-linkage studies of TB incidence in the United Kingdom and elsewhere, indicating a plausible logistic regression model but expressing concern about the contribution of unscrutinized Hospital data sources to accurate estimates of TB incidence [8, 17, 19, 20].

Limitations due to violation of the underlying capture–recapture assumptions

The capture–recapture findings have to be placed in the context of the limitations of this study. The assessment of the coverage of the TB data sources was based on three-source log-linear capture–recapture models, only valid in the absence of violation of their underlying assumptions: perfect record-linkage (i.e. no misclassification of records), a closed population (i.e. no immigration or emigration in the time period studied) and a homogeneous population (i.e. no subgroups with markedly different probabilities to be observed and re-observed). In two-source capture–recapture methods one must also assume independence between data sources [i.e. the probability of being observed in one data source is not affected by being (or not being) observed in another] [9]. In the three-source capture–recapture approach dependencies between two data sources (pair-wise interdependencies) can be identified and incorporated in the log-linear model. However, the three-way interaction, i.e. dependency between all three data sources, cannot be incorporated in the model and its absence must be assumed. This and other limitations of capture–recapture analysis are described elsewhere in more detail [12, 22, 24–29].

Violation of the perfect record-linkage assumption and the problem of possible false-positive cases have already been discussed. Violation of the closed population assumption is presumed to be limited for TB as the opportunities for notification, culture confirmation or hospitalization are, also for immigrants, largely determined within a short period of time. However, this violation could result in overestimation of the number of patients.

TB services in England are organized around close collaboration between clinicians, microbiologists and public health professionals such as communicable disease control consultants and TB nurses. The

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press

Table 5. *Annual and overall estimated number of unobserved and total tuberculosis* (*TB*) *cases by structural source model and truncated Poisson mixture model in England between 1999 and 2002* (*after using a proportion of 28% of true TB cases known only to Hospital in the corrections for false-positive cases*)

Year	Estimated unobserved number of TB cases by the structural source model (95% ACI)	Estimated total number of TB cases by the structural source model (95% ACI)	Estimated unobserved number of TB cases by the truncated Poisson mixture model (95% ACI)	Estimated total number of TB cases by the truncated Poisson mixture model (95% ACI)
1999	9151 (3921–12 186)	15 934 (10 704–18 969)	1319 (1137–1509)	8102 (7920–8292)
2000	3737 (2588–4090)	10 876 (9727–11 229)	2019 (1802–2247)	9158 (8941–9386)
2001	2294 (2253–3389)	9649 (9608–10 774)	1256 (1074–1445)	8611 (8429–8800)
2002	1487 (1337–1973)	8888 (8738–9374)	917 (748–1093)	8398 (8229–8574)
All	13 628 (9186–15 563)	42 306 (37 864–44 241)	5417 (5217–5737)	34 149 (33 895–34 415)

ACI, Approximate confidence interval.

log-linear capture–recapture models with the best goodness-of-fit were saturated models, i.e. including all two-way interactions. Violation of the absent (positive) three-way interaction assumption, biasing the estimates of the true population size downwards, cannot be ruled out [12, 26, 27, 30].

Violation of the homogeneity assumption is also likely: age, site of disease and infectiousness, among others, can cause different probabilities of being observed in a TB data source. One way of handling possible heterogeneity is to stratify the population into more homogeneous subpopulations and then to carry out capture–recapture analyses for each of the distinct groups. However, our corrections for the projected and estimated proportion of Notification and especially Hospital records being false-positive, and incomplete availability of relevant identifiers in all data sources prevented meaningful stratification. To investigate possible bias in the log-linear capture–recapture estimates as a result of violation of the homogeneity assumption, we have re-examined the data with alternative models, as described in the capture–recapture literature [12, 30, 31]. These models reportedly perform well when compared to log-linear capture–recapture estimates and are arguably more robust to violation of the homogeneity assumption [30, 32, 33].

(1) We first applied a structural source model [30]. This method models potential heterogeneity of the population, partly based on prior knowledge, and estimates the probabilities of conditions that produce the relationships between the data sources; more specifically in this instance, the proportion of patients with pulmonary or extrapulmonary TB in the population. The annual and overall estimated number of unobserved and total TB cases is shown in Table 5 but the structural source model did not fit well. The number of unobserved TB cases is very high in 1999 but then falls considerably every year to lower estimates compared to the saturated log-linear model, although each year the confidence intervals of both estimates overlap. The estimated annual Notification-specific coverage rate improves every year. The approximate confidence interval of the 2002 estimate includes expected values of under-notification.

The structural source model estimates a large majority of the unobserved TB cases to have extrapulmonary TB. Local under-notification of non-respiratory TB of 47% has been reported in the United Kingdom [8]. This possibly reflects health service organization in the United Kingdom where extrapulmonary cases are less likely to be managed by clinicians familiar with notification of infectious diseases. Apart from underestimating the burden of TB, the implications for public health are limited as extrapulmonary TB patients are rarely infectious.

(2) We tested our data using Zelterman's truncated Poisson mixture model, which is also vulnerable to possible violation of underlying assumptions [34]. This estimator and similar ones have been used in the social sciences to estimate the size of hidden populations such as illicit drug users and homeless persons [33, 35–37]. A recent publication compares three-source capture–recapture model estimates with the estimates of truncated models, including Zelterman's model, for

19 datasets of infectious disease incidence and discusses the conditions where these estimates are similar or dissimilar [38]. The results of this study suggest that for estimating infectious disease incidence and completeness of notification independent (i.e. without pair-wise interdependencies between the data sources) and parsimonious (i.e. incorporating one or two pair-wise interdependencies between the data sources) three-source log-linear capture–recapture models are preferable. However, when saturated models are selected as the best-fit model and the estimates are unexpectedly high and seem implausible the data should be re-examined with truncated models as a heuristic tool, in the absence of a gold standard, to identify possible failure in the saturated log-linear model. When the truncated models produce a lower and more plausible estimated number of infectious disease patients arguments are put forward that the estimates of the truncated models could be preferable. Table 5 shows the annual and overall estimated numbers of unobserved and total TB cases. The estimated numbers of unobserved TB cases were low compared to the structural source model, especially in 1999. From 2000 onwards the estimates fell every year. According to Zelterman's model, estimated completeness of Notification was 70·7% overall and 68·5%, 63·8%, 73·6% and 76·4% for the years 1999–2002 respectively. The confidence intervals do not overlap with the other models but include expected values of under-notification in 2001 and 2002.

In the comparative study mentioned above, the number of TB patients in England was also estimated using a Poisson heterogeneity model and a truncated binomial model [38]. Compared to the Zelterman model, the Poisson heterogeneity model estimated a slightly lower overall completeness of Notification (68·7%) and the truncated binomial model estimated a slightly higher completeness of Notification (73·3%). The latter result could be an overestimate due to some violation of the equiprobability assumption underlying the binomial model [38].

Hook & Regal state that 'In no sense is there any proof or reassurance that application of multiple-source log-linear estimators for any particular observed data on real populations results in a valid estimate, nor even necessarily produce an estimate closer to the true value than some alternative approach' and 'if the saturated log-linear model is selected by any criterion the investigator should be particularly cautious about using the associated outcome' [12]. Confidence in the validity of capture–recapture results may reflect publication bias in favour of successful capture–recapture studies rather than the inherent strength of this methodology [39].

CONCLUSION

Record-linkage, as performed in ETS, improves accuracy of surveillance data as well as completeness of case ascertainment of TB. Hospital-derived data added a limited number of possible true TB patients. Since the introduction of ETS the annual observed completeness of notification has increased. This is probably due to improvements in case reporting combined with improved data collection and record-linkage. This study shows that observed under-notification of TB cases in England might be as high as 10·4% as these cases were laboratory-confirmed but not notified. The overall observed under-notification was 15·9% which is consistent with previous reports. Overall under-notification estimated by a saturated log-linear capture–recapture model was highly inconsistent with previous reports and could be an overestimate due to violation of the underlying assumptions, especially the homogeneity assumption as suggested by the alternative models.

Instead of capture–recapture analysis including hospital episode registers, record-linkage and case ascertainment using the two most relevant sources for infectious disease surveillance, namely notification and laboratory, both with an expected high specificity and hence positive predictive value, as performed in ETS, will often already considerably improve the knowledge of the number of patients and infectious disease incidence rates, as well as the completeness of information on specific demographic, diagnostic or epidemiological variables. All unlinked laboratory cases in addition to the notifications are by definition TB cases. According to Zelterman's truncated model, in England the estimated completeness of the Notification and Laboratory records combined was 78·2%, 74·1%, 81·0% and 83·8% for 1999–2002 respectively, all within the expected range of under-notification and consistent with the results of parsimonious capture–recapture model estimates in some other European countries [19, 20]. Real-time record-linkage of laboratory data and incident case reports in ETS allows for appropriate prospective action to be

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press

taken, such as identifying and approaching the clinicians treating the unlinked culture-positive TB cases by the local consultants in communicable disease control or TB control nurses, considering the unlinked MycobNet reports as 'pre-notifications', and encouraging the clinicians to notify these patients. This would increase the completeness of the notifications register as would campaigns to raise awareness of complying with (compulsory) notifications among clinicians by public health authorities. Appointing a clinician, e.g. one of the consultant chest physicians, as TB coordinator in every hospital, to be consulted for each patient with TB in that hospital, including extrapulmonary cases, could further promote notification.

ACKNOWLEDGEMENTS

We thank David Quinn for designing the record-linkage software and processing the record-linkage procedures, Charlotte Anderson for providing the MOTT data and Valerie Delpech and Filip Smit for their suggestions on alternative models. Permission for this study was obtained from the Hospital Episode Statistics Security and Confidentiality Advisory Group (S&CAG).

NOTE

Supplementary material accompanies this paper on the Journal's website (http://journals.cambridge.org).

DECLARATION OF INTEREST

None.

REFERENCES

1. **Health Protection Agency.** *Tuberculosis cases reported, England and Wales, 1988, 1993, 1998–2005* (http://www.hpa.org.uk/infections/topics_az/tb/epidemiology/table14.htm). Accessed 23 May 2007.
2. **Rose AM, Gatto AJ, Watson JM.** Recent increases in tuberculosis notifications in England and Wales – real or artefact? *Journal of Public Health Medicine* 2002: **24**: 136–137.
3. **Pillaye J, Clarke A.** An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003; **1**: 31.
4. **Van Buynder P.** Enhanced surveillance of tuberculosis in England and Wales: circling the wagons? *Communicable Disease and Public Health* 1998; **1**: 219–220.
5. **Sheldon CD, *et al.*** Notification of tuberculosis: how many cases are never reported. *Thorax* 1992; **47**: 1015–1018.
6. **Roderick PJ, Connelly JB.** The problems of monitoring tuberculosis in an inner-city health district: integrated information is required. *Public Health* 1992; **106**: 193–201.
7. **Devine MJ, Aston R.** Assessing the completeness of tuberculosis notification in a health district. *Communicable Disease Report. CDR Review* 1995; **5**: R137–R140.
8. **Mukerjee AK.** Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data sources. *Communicable Disease and Public Health* 1999; **2**: 143–144.
9. **International Working Group for Disease Monitoring and Forecasting.** Capture-recapture and multiple-record estimation I: History and theoretical development. *American Journal of Epidemiology* 1995; **142**: 1047–1058.
10. **Fienberg SE.** The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika* 1972; **59**: 591–603.
11. **Bishop YMM, Fienberg SE, Holland PW.** *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press, 1975.
12. **Hook EB, Regal RR.** Capture-recapture methods in epidemiology: methods and limitations. *Epidemiologic Reviews* 1995; **17**: 243–263.
13. **International Working Group for Disease Monitoring and Forecasting.** Capture-recapture and multiple-record estimation II: Applications in human diseases. *American Journal of Epidemiology* 1995; **142**: 1059–1068.
14. **Devine MJ, *et al.*** Whooping cough surveillance in the north west of England. *Communicable Disease and Public Health* 1998; **1**: 121–125.
15. **Crowcroft NS, *et al.*** Deaths from pertussis are underestimated in England. *Archives of Disease in Childhood* 2002; **86**: 336–338.
16. **Breen E, *et al.*** How complete and accurate is meningococcal disease notification? *Communicable Disease and Public Health* 2004; **7**: 334–338.
17. **Tocque K, *et al.*** Capture recapture as a method of determining the completeness of tuberculosis notifications. *Communicable Disease and Public Health* 2001; **4**: 141–143.
18. **Cailhol J, *et al.*** Incidence of tuberculous meningitis in France, 2000: a capture-recapture analysis. *International Journal of Tuberculosis and Lung Disease* 2005; **9**: 803–808.
19. **Baussano I, *et al.*** Undetected burden of tuberculosis in a low-prevalence area. *International Journal of Tuberculosis and Lung Disease* 2006; **10**: 415–421.
20. **Van Hest NA, *et al.*** Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiology and Infection* 2007; **135**: 1021–1029.
21. **Van Hest NA, Smit F, Verhave JP.** Underreporting of malaria incidence in The Netherlands: results from a

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press

capture–recapture study. *Epidemiology and Infection* 2002; **129**: 371–377.

22. **Papoz L, Balkau B, Lellouch J.** Case counting in epidemiology: limitations of methods based on multiple data sources. *International Journal of Epidemiology* 1999; **25**: 474–478.
23. **Borgdorff MW, Glynn JR, Vynnycky E.** Using capture-recapture methods to study recent transmission of tuberculosis. *International Journal of Epidemiology* 2004; **33**: 905–906.
24. **Desenclos JC, Hubert B.** Limitations to the universal use of capture-recapture methods. *International Journal of Epidemiology* 1994; **23**: 1322–1323.
25. **Brenner H.** Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995; **6**: 42–48.
26. **Cormack RM.** Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *Journal of Clinical Epidemiology* 1999; **52**: 909–914.
27. **Hook EB, Regal RR.** Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *American Journal of Epidemiology* 2000; **152**: 771–779.
28. **Jarvis SN, *et al.*** Children are not goldfish-mark-recapture techniques and their application to injury data. *Injury Prevention* 2000; **6**: 46–50.
29. **Tilling K.** Capture-recapture methods-useful or misleading? *International Journal of Epidemiology* 2001; **30**: 12–14.
30. **Regal RR, Hook EB.** Marginal versus conditional versus structural source models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Statistics in Medicine* 1998; **17**: 69–74.
31. **Wilson RM, Collins MF.** Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; **79**: 543–553.
32. **Hook EB, Regal RR.** Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *American Journal of Epidemiology* 1982; **116**: 168–176.
33. **Smit F, Reinking D, Reijerse M.** Estimating the number of people eligible for health service use. *Evaluation and Program Planning* 2002; **25**: 101–105.
34. **Zelterman D.** Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *Journal of Statistical Planning and Inference* 1988; **18**: 225–237.
35. **Smit F, Toet J, Van der Heijden PG.** Estimating the number of opiate users in Rotterdam using statistical models for incomplete count data. In: Hay G, McKeganey N, Birks E, eds. *Final report EMCDDA project Methodological Pilot Study of Local Level Prevalence Estimates*. Lisbon: EMCDDA, 1997, pp. 47–66.
36. **Bohning D, *et al.*** Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *European Journal of Epidemiology* 2004; **19**: 1075–1083.
37. **Hay G, Smit F.** Estimating the number of hard drug users from needle-exchange data. *Addiction Research and Theory* 2003; **11**: 235–243.
38. **Van Hest NA, *et al.*** Estimating infectious disease incidence: validity of capture-recapture analysis and truncated models for incomplete count data. *Epidemiology and Infection* 2008; **136**: 14–22.
39. **Hay G.** The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; **46**: 515–520.

https://doi.org/10.1017/S0950268808000496 Published online by Cambridge University Press