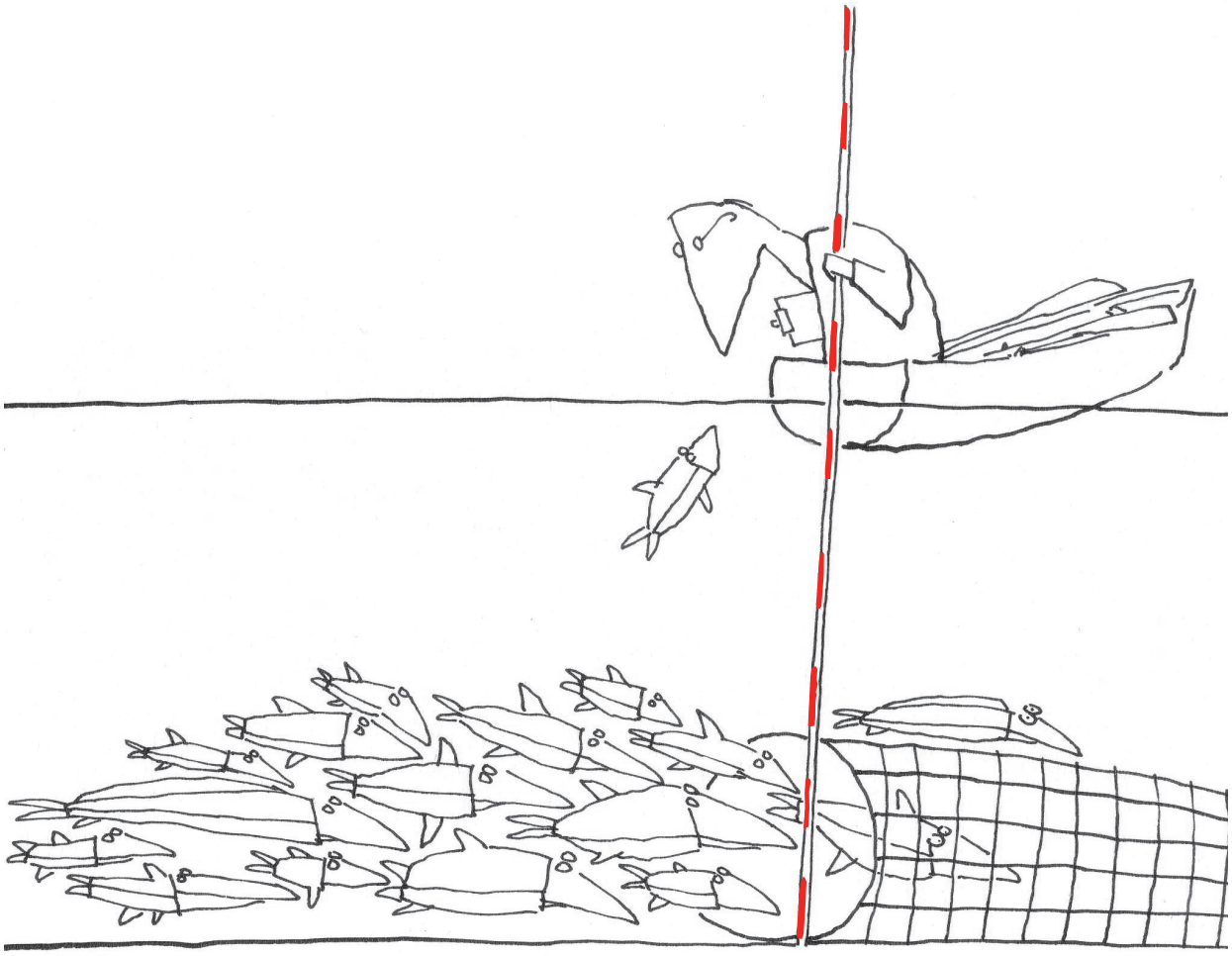
Capture-recapture Methods in Surveillance of Tuberculosis and Other Infectious Diseases



Rob van Hest

Capture-recapture Methods in Surveillance of Tuberculosis and Other Infectious Diseases

Rob van Hest

Colofon
Capture-recapture methods in surveillance of tuberculosis and other infectious diseases / Van Hest, Rob
Thesis Erasmus MC, University Medical Center Rotterdam – With summary in English and Dutch
ISBN: 978-90-9021974-5

Lay-out: Grafisch Bureau DUS bv; Rob van Hest
Cover design: Len Munnik

Printed by Print Partners Ipskamp, Enschede

The printing and distribution of this thesis was financially supported by the Municipal Public Health Service Rotterdam-Rijnmond, the Department of Public Health, Erasmus MC (University Medical Center Rotterdam), the Erasmus University Rotterdam, the Dutch Association of Tuberculosis Control Physicians, the Dr. C. de Langen Foundation for Global Tuberculosis Control.

Capture-recapture Methods in Surveillance of Tuberculosis and Other Infectious Diseases

Vangst-hervangst methoden voor surveillance van tuberculose en andere infectieziekten

Proefschrift

ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam

op gezag van de

rector magnificus

Prof.dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

vrijdag 14 september 2007 om 11.00 uur

door

Norbertus Alphonsus Henricus van Hest

geboren te Tilburg

Promotiecommissie

Promotor:	Prof.dr. J.D.F. Habbema
Overige leden:	Prof.dr. G.J. Bonsel
	Prof.dr. M.W. Borgdorff
	Prof.dr. A. Hofman
Copromotor:	Dr. J.H. Richardus

Quod potes, tenta

Contents **Page**

1

Introduction

1.1 Assessing completeness of ascertainment in epidemiology

Epidemiology is the study of how often diseases occur in different groups of people and why.[1] This includes knowledge about classification errors, i.e. the absence of true cases and the presence of false-positive cases in registrations. Observing and monitoring health and behaviour trends requires a surveillance system that captures useful data on those persons correctly identified with the characteristic under study. This information can be used to identify priorities and evaluate interventions. To determine the usefulness of a surveillance system there is the need to assess the quality of the data and completeness of ascertainment.[2] The number of individuals with a certain condition (i.e. cases) or events in a population can be ascertained directly, by counting every single person or event as attempted in a census, or indirectly, by obtaining sufficient information to estimate prevalence (i.e. the number of cases at a specific point in time) or incidence (i.e. the number of new cases during a specific period of time), as attempted in a survey (active case-finding) or by notification (passive case-finding). Other examples of indirect ascertainment of the number of cases in a population are pharmaco-epidemiolocal studies and record-linkage, i.e. comparing patient data across multiple registers. It is difficult to establish whether these counts are complete or biased to under-ascertainment and only on a few occasions it is attempted to estimate or adjust for missing cases.[3,4] An indirect technique that estimates completeness of ascertainment of surveys and registers used in epidemiological studies is capture-recapture analysis.[5,6]

1.2 Brief introduction to capture-recapture analysis

A more extensive overview of the historical development of capture-recapture analysis is given elsewhere.[5] Briefly, the first use of capture-recapture analysis can be traced back to Graunt who used a similar method for estimating the population of England as early as 1662[7] and Laplace, who attempted to estimate the population size in France in 1782,[8] but usually it is mentioned that capture-recapture analysis was first applied by Petersen in 1894 for the study of fish populations. He used the so-called two-sample method, the simplest capture-recapture model, to estimate the unknown size of a population of plaice in the Limfjord in Denmark.[9] The first sample provides the animals for marking or tagging and is returned to the population, while the second sample provides the recaptures, i.e. the numbers of animals caught in both samples. Using the number of recaptures and the number of animals caught in the first and the second sample, it is possible, under certain assumptions, to estimate the number not caught in either sample, thus providing an estimate of the total population size. Two-sample capture-recapture analysis was extended to multiple-sample capture-recapture analysis by Schnabel in 1938.[10] Unmarked animals in each sample are given individual (i.e. numbered) marks before being returned to the population, resulting in a known capture history of each marked animal. The theory of capture-recapture models was developed more fully in the 1950's, for example by Chapman,[11] who suggested an adjustment of the capture-recapture estimate to reduce small sample bias, known as the "Nearly Unbiased Estimator", and Darroch,[12] who founded the mathematical framework. To tackle the problem of violation of the underlying assumptions, within animal population biology a range of different

models was introduced in the 1970's, associated with the names of Anderson, Burnham, Otis, White and others.[13]

For human conditions two-source capture-recapture analysis was first applied to census data. By taking another data source in addition to the census the undercount can be estimated.[14] The first use of capture-recapture analysis to human health was Sekar and Deming's estimation of completeness of birth and death registers in 1949,[15] translating being captured in wildlife samples into being observed in two incomplete data sources or registers. Personal identifiers such as identification numbers and/or names are used as marks or tags. The use of capture-recapture analysis within epidemiology came relatively late and was introduced by Wittes and colleagues in 1968 and generalised to three registers,[16] four registers,[17] and five registers,[18] benefiting from advanced capture-recapture knowledge and improved (computerised) statistical methods. In 1972 Fienberg approached the problem of violation of some underlying assumptions through the use of a log-linear model, as it had emerged from the analysis of multidimensional contingency tables,[19,20] and since then other models have been added.[21,22] A detailed description of the methodology and mathematical framework of capture-recapture analysis and its application in epidemiological studies is given in chapter 2.

In various disciplines, for example ecology, demography or epidemiology, capture-recapture analysis is known under different names, such as the Petersen estimator, mark-recapture and multiple-recapture method, dual-record system and multiple-record system method, or Bernoulli census estimates and ascertainment corrected rates.

1.3 Application of capture-recapture methods in tuberculosis surveillance

Tuberculosis under-notification

Already in 1952 Rene Dubos wrote, based on medical and social arguments, that "for all these reasons, it is impossible to obtain accurate data concerning the prevalence of tuberculous infection even during recent times".[23] In 1981 Styblo and Rouillon said that "for Africa, Asia and Latin America the reported tuberculosis incidence figures were, with a few exceptions, totally unreliable and incomplete, and should not be extrapolated to areas with no notification of tuberculosis and, most important, should not be taken into consideration to assess the trend of the incidence of tuberculosis in the world".[24] A conventional surveillance system for tuberculosis is (mandatory) notification but tuberculosis under-notification is suspected universally, not only in high tuberculosis-burden countries, although the level can differ considerably.[25,26] In various low tuberculosis-burden countries under-notification of tuberculosis has been reported, estimated through other methods than capture-recapture analysis. In the USA under-notification of tuberculosis was estimated at 37% in 1977[27] but more recent studies in several states found under-notification in the range of 7% to 0.5%.[28-30] In Europe, failure to notify tuberculosis patients has best been demonstrated in the United Kingdom, varying in different settings from 7% to 70% of the patients,[31-39] and as high as 68% to

94% among AIDS patients with tuberculosis.[40,41] In France, local studies suggested under-notification of tuberculosis in the range of 30% to 63%.[42-44] The under-notification of tuberculosis in Italy has been estimated by the World Health Organization at 12%, but may reach between 30% and 54% in some areas of the country.[45-49] In Spain under-notification of tuberculosis has been estimated at 50%.[50,51] In the Netherlands tuberculosis under-notification was estimated at 8% between 1994 and 1998 whilst according to the World Health Organization 100% of the cases were notified in 2002.[26,52]

Under-notification obscures the true burden of tuberculosis, it frustrates proper planning of the human and financial resources needed for adequate tuberculosis control, it hinders meaningful interpretation of figures and trends for surveillance and identifying priorities, it will compromise early signs of location and magnitude of outbreaks and it will also fail to reliably evaluate the effect of interventions. Compared to non-communicable diseases, for tuberculosis under-notification has an important additional consequence. The most serious public health aspect of tuberculosis under-notification, especially for culture-confirmed pulmonary tuberculosis patients, is that it prevents possibly indicated contact investigations around potentially infectious patients.

Methods of estimating tuberculosis incidence or prevalence

Apart from (mandatory) notification different methods can be used to estimate the burden of disease of tuberculosis. As for other diseases a whole population can be examined, but often this is not feasible, expensive and only representative for the area explored,[53] or the ascertainment of the number of tuberculosis patients can be achieved through exhaustive surveys.[54-58] A specific surrogate marker for the incidence of pulmonary tuberculosis is the annual rate of tuberculin skin test conversion. This method assumes that a one percent annual risk of infection with *Mycobacterium tuberculosis* corresponds to an incidence of approximately 50 smear-positive cases of pulmonary tuberculosis per 100 000 population.[59,60] However, these risks were originally drawn from developed nations and, because of variations in the quality of intervention and varying risks of progression from latent tuberculosis infection to the active disease, it is unclear whether these figures can be reliably projected elsewhere.[61] Specific pharmaco-epidemiological studies on anti-tuberculous drug use, especially daily defined doses of pyrazinamide, have been used to estimate tuberculosis incidence.[51,52,62] One of the limitations of using drug prescriptions as a marker for tuberculosis incidence is the difficulty to distinguish between chemoprophylaxis and chemotherapy.[35] The general indirect estimation technique of record-linkage has also been applied to tuberculosis. Through record-linkage it is possible to assess the case-ascertainment and come closer to the true number of cases than by using one source only.[37,45,46,48,63] For tuberculosis and other infectious diseases most often microbiology laboratory records or hospital episode registers are used to supplement notification data. However, proportions of miscoded (false-positive) tuberculosis patients in hospital episode registers of up to 62% and 27% have been reported in the USA and the United Kingdom respectively.[64-66] Pathology records compatible with tuberculosis have been mentioned as an alternative source.[32,33,36,67] Other options are pharmacy data, such as prescriptions,[30,35,38,67-69] death certificates,[28,30,33,34] AIDS registries[30,70] and billing records.[64]

Capture-recapture studies estimating tuberculosis incidence or prevalence

In addition to record-linkage of two or more tuberculosis registers, capture-recapture studies have been performed in the field of tuberculosis surveillance.[65,71-83] An overview of these studies is presented in Table1.1. Nine of the first 11 reports, published prior to the work in this thesis, used simple two-source capture-recapture models and only two studies (Liverpool, United Kingdom and Cayenne, French Guyana) applied three-source capture-recapture models. Seven studies originated from Spain. Five studies estimated the number of patients with pulmonary or respiratory tuberculosis, four studies estimated the number of patients with all forms of tuberculosis and one study estimated the number of patients with tuberculous meningitis. One study did not aim to estimate the total number of tuberculosis patients but the number of tuberculosis patients attributable to recent transmission.[78] Ten studies were performed at the local or regional level and one study was done at the national level. Nearly all studies used mandatory notification data, microbiology laboratory records or hospital episode statistics as data sources. Interviews with local residents and a national reference centre mycobacterial drug resistance survey were used as alternative data sources, apart from the study estimating the number of patients attributable to recent transmission which used an epidemiological recent transmission database and a microbiology DNA fingerprinting database. Two studies included less than 100 patients after record-linkage, eight studies between 100 and 500 patients and one study involved 1248 patients. Three studies describe a sequential analysis over multiple years. Estimated under-notification varied between 7.4% and 65%.

Limitations of capture-recapture studies estimating tuberculosis incidence or prevalence

The limitations to capture-recapture studies estimating tuberculosis incidence or prevalence depend, like in any capture-recapture study, on the violation of the underlying assumptions. These assumptions and methodological aspects of their violation will be discussed in detail in section 2.1.2. Briefly, for tuberculosis, as well as other diseases, violation of the perfect record-linkage assumption (i.e. no misclassification of records) is depending on the availability of a unique identifier in all registers, or sufficient proxy-identifiers. When, as most often, notification, laboratory and hospital registers are used, violation of the closed population assumption (i.e. no immigration or emigration in the time period studied) is presumably limited in countries with a well-organised tuberculosis control system, as the opportunities for notification, culture-confirmation or hospitalisation are largely determined within a short period of time. More likely is the violation of the assumption of independence between the different tuberculosis registers (i.e. the probability of being observed in one register is not affected by being –positive dependence– [or not being –negative dependence–] observed in another) as tuberculosis services are often organised around close collaboration (e.g. laboratory pre-notification, clinical isolation, contact-investigations and referrals) between clinicians, microbiologists and public health professionals such as tuberculosis physicians and tuberculosis nurses. Another more likely violation is that of the homogeneity assumption (i.e. the absence of subgroups in the population with markedly different probabilities of being observed and re-observed), e.g. age, location of disease and infectiousness can cause different probabilities of being observed in a tuberculosis-related register. A problem more specific

Table 1.1 Objectives, methods, data-sources, number of patients included and outcomes of published capture-recapture studies of tuberculosis

Researchers	Objective	Method	Data-source	*N*obs[a]	Outcome
Ferrer Evangelista *et al*[71]	To determine the incidence of pulmonary TB[b] within Health Area 15 of Valencia, Spain, 1990-1993	Two-source CRC[c] model	1. Statutory notification system 2. Hospital microbiology register	150	The annual incidence of TB was estimated at 34.8/100 000 population (95%CI[d] 31.8-39.9%), resulting in a completeness of the statutory notification system of 45% (95%CI 40-50%)
Iváñez Gimeno *et al*[72]	To evaluate the respiratory TB surveillance system in the province of Seville, Spain, in 1996.	Two-source CRC model	1. Notifiable disease surveillance system 2. Microbiological records	348	The completeness of the notifiable disease surveillance system and microbiological register were estimated at 51.3% and 80.7% respectively. Both registers combined captured 90.6% of the TB cases
Sanghavi *et al*[73]	To estimate the incidence of pulmonary TB in the shantytown Las Pampas de San Juan de Miraflores, Peru, 1989-1993	Two-source CRC model	1. Interviews with local residents 2. Local laboratory sputum smear records	121	The average annual incidence of pulmonary TB per 100 000 inhabitants was estimated at 364 (95%CI 293-528) and completeness of Ministry of Health reports at 37% (95%CI 25-46%)
Pérez Ciorda *et al*[74]	To estimate the number of TB cases in the province of Huesca, Spain, 1995-1997	Two-source CRC model	1. Notifiable disease surveillance system 2. Microbiological records	244	The number of TB patients was estimated at 272 and underreporting by the notifiable disease surveillance system at 22.4%
Tocque *et al*[65]	To determine the true number of TB cases in the catchment area of Liverpool Health Authority, United Kingdom, 1989-1996	Three-source log-linear CRC model	1. Notifications 2. Microbiological records 3. Hospital in-patient discharge coding data	473	The number of TB patients was estimated at 485 resulting in a case-ascertainment of 97.5%. Notifications, microbiological records and hospital in-patient discharge coding data identified 92.6%, 56.3% and 36.3% of the TB patients respectively

Researchers	Objective	Method	Data-source	*N*obs[a]	Outcome
Mayoral Cortes *et al*[5]	To estimate the incidence of pulmonary TB and HIV[e] co-infection and to assess the sensitivity and positive predictive value of the notifiable disease surveillance system in the province of Seville, Spain in 1998	Two-source CRC model	1. National notifiable disease surveillance system (NNDSS) 2. Hospital discharge database	308	The annual incidence of pulmonary TB per 100 000 inhabitants was estimated at 25.6 (95%CI 21.5-28.8), and case-ascertainment at 72%. The sensitivity of NNDSS was estimated at 65.3% and the positive predictive value at 89.3%
Iglesias Gozalo *et al*[6]	To estimate the incidence of TB and the epidemiological characteristics of the disease in the Zaragoza province, Spain, 1993-1995	Two-source CRC model	1. Notifiable disease surveillance system 2. Microbiological records	1248	The average annual incidence of TB per 100 000 inhabitants was estimated at 48.5 (95%CI 46.5-50.6), and underreporting by the notifiable disease surveillance system at 38% (95%CI 35-40%)
Tejero Encinas *et al*[7]	To estimate the incidence of pulmonary TB and to assess the completeness of the two sources used in Valladolid, Spain 1996-2000	Two-source CRC model	1. Compulsory notifiable disease system 2. Hospital admission database	255	The average annual incidence of pulmonary TB per 100 000 inhabitants was estimated at 24.4 (95% CI 23.5-25.3), resulting in a case-ascertainment of 95% (95%CI 91-98%)
Iñigo *et al*[8]	To estimate the rate of recent TB transmission in 3 urban districts of Madrid, Spain, 1997-199	Two-source CRC model	1. Epidemiological recent transmission database 2. DNA fingerprinting database	87	According to molecular analysis alone 31.6% of the TB cases were due to recent transmission. This proportion increased to 44.8% after CRC analysis
Cailhol *et al*[9]	To estimate the incidence of culture-positive and culture-negative tuberculous meningitis in France in 2000.	Two-source CRC model	1. Mandatory TB notification system 2. National reference centre TB drug resistance survey	38	After record-linkage a total of 38 cases of culture-positive tuberculous meningitis were identified while CRC analysis estimated 41 cases. Underreporting was estimated at 32%

Researchers	Objective	Method	Data-source	*N*obs[a]	Outcome
Guernier *et al*[80]	To estimate the level of under-reporting and to improve the estimates of the incidence of TB in the vicinity of Cayenne, French Guyana, 1996-2003	Three-source log-linear CRC model	1. Mycobacterial laboratory database 2. Hospital Information database 3. TB Control Service database	381	Log-linear CRC estimated the number of TB patients at 462 (95%CI 423-536) and under-reporting at 49.1%, 38.7% and 41.3% for data sources 1, 2 and 3 respectively
Baussano *et al (this thesis)*[81]	To assess the completeness of TB registration, incidence and underreporting in the Piedmont Region of Italy in 2001	Three-source log-linear CRC model	1. Physician notification system 2. TB laboratory register 3. Hospital records register	657	CRC estimated 47 (95% CI: 31-71) unrecorded cases. Underreporting of the "physician notification system" was estimated at 21% (95%CI 20%-23%)
Van Hest *et al (this thesis)*[82]	To describe a systematic process of record-linkage and case-validation and to assess the completeness of TB notification in the Netherlands in 1998	Three-source log-linear CRC model	1. Mandatory TB notification system 2. National mycobacteriology reference laboratory records 3. Hospital admission database	1499	Observed completeness of TB notification was 86.6%, increasing to 92.7% after adjustment for possible imperfect record-linkage and false-positive hospital cases. Log-linear CRC estimated completeness of notification at 63.2%, increasing to 86.6% after adjustment.
Van Hest *et al (this thesis)*[83]	To observe and estimate the annual incidence of TB and to assess the completeness of TB registers in England, 1999-2002	Three-source log-linear CRC model	1. Mandatory TB notification system 2. National mycobacteriology reference laboratory records 3. Hospital admission database	28768	Observed and estimated annual TB incidence was 6783 and 11 539, 7139 and 11 433, 7355 and 10 742, and 7401 and 9647 patients between 1999 and 2002. Annual estimated completeness of notification between 1999 and 2002 was 48.1%, 51.1%, 59.0% and 66.5% respectively

a: *N*obs = observed number of tuberculosis patients after record-linkage; b: TB = tuberculosis; c: CRC = capture-recapture; d: CI = confidence interval; e: HIV = human immunodeficiency virus; f: PPV = positive predictive value

for tuberculosis than for some other (infectious) diseases is the uniform case-definition in all registers. For laboratory registers the gold standard is a positive culture of a species belonging to the *M. tuberculosis* complex. For the other registers there is a risk of less than perfect positive predictive value, i.e. the registers can include false-positive records, for example due to laboratory-contamination results, infection with Mycobacteria Other Than Tuberculosis or a final diagnosis other than tuberculosis, which are not officially de-notified or adjusted in the hospital discharge codes. As outlined in section 1.3, in case of tuberculosis, studies have shown that the positive predictive value of hospital registers should be judged critically.

1.4 Aim of the thesis and research questions

The aim of this thesis is to investigate the feasibility and validity of capture-recapture methods in surveillance of tuberculosis and other infectious diseases. The specific research questions are:

1. How do the characteristics of various infectious diseases and their registers in the Netherlands influence the feasibility and validity of capture-recapture analysis?
2. How do the characteristics of tuberculosis surveillance systems in different countries influence the feasibility and validity of capture-recapture analysis?
3. What is the feasibility and validity of truncated population estimation models in infectious disease surveillance?

1.5 Outline of the thesis

This thesis begins with three chapters that introduce different aspects of capture-recapture analysis. After a brief general introduction to capture-recapture analysis **Chapter 1** addresses under-notification of tuberculosis and the application of capture-recapture methods in tuberculosis surveillance, with an overview of previous capture-recapture studies performed in this field. In **Chapter 2** we discuss the methodology and mathematical framework of capture-recapture analysis, especially for epidemiological studies. Further to an earlier overview of capture-recapture studies on infectious diseases until 1997, in **Chapter 3** we present a synopsis of capture-recapture studies on infectious diseases published between 1997 and 2006.

To address the first research question of this thesis in **Chapter 4** we estimate malaria incidence and completeness of notification by clinicians and reporting by laboratories in the Netherlands, to assess the effect of the change from clinician-based notification to laboratory-based reporting of malaria in the new Dutch Infectious Diseases Act. We describe an uncomplicated conventional log-linear capture-recapture analysis of three incomplete, partially overlapping registers of malaria cases, resulting in a parsimonious log-linear model, reducing bias due to interdependence between registers. In **Chapter 5** we estimate the incidence and completeness of notification of Legionnaires' disease in the Netherlands. We describe a less conventional and more complicated three-

source capture-recapture analysis, resulting in a covariate log-linear model in order to reduce bias due to expected and observed geographical heterogeneity among the Legionnaires' disease patients. In **Chapter 6** we estimate the completeness of notification of incident tuberculosis cases in the Netherlands. We describe a systematic process of record-linkage of three tuberculosis registers, cross-validation with four other tuberculosis-related registers, case-ascertainment and conventional capture-recapture analysis, initially resulting in a saturated log-linear model, and demonstrate the effect of possible violation of the perfect record-linkage and perfect positive predictive value of registers assumptions.

To address the second research question of this thesis, we show in **Chapter 6** how a well-organised system of tuberculosis control in the Netherlands allows us to correct an implausible high number of tuberculosis patients estimated by a saturated log-linear model in a capture-recapture study at the national level for one year. We introduce the application of a truncated Poisson model, related to capture-recapture analysis, to cross-validate the conventional capture-recapture estimates. In **Chapter 7** we describe a relatively uncomplicated conventional three-source capture-recapture study to estimate tuberculosis incidence and completeness of the tuberculosis registration systems at the regional level in the Piedmont Region in Italy for one year, resulting in a parsimonious log-linear model. We show how the regional scale of the study, a limited number of patients and favourable privacy regulations made inspection of all clinical files possible and allowed for identification of a considerable number of false-positive cases in the hospital register. In **Chapter 8** we show the advantage of two routinely linked tuberculosis registers, as part of Enhanced Tuberculosis Surveillance in England, in a three-source capture-recapture study at the national level for four years. Due to the scale of this study as a disadvantage sophisticated record-linkage software was needed to link the hospital records as a third data source and a population mixture model had to be specified to estimate the proportion of false-positive cases among the unlinked hospital-derived tuberculosis records. The final tuberculosis incidence estimates of the saturated capture-recapture models are cross-validated with a structural source model, a truncated Poisson model and a truncated Poisson mixture model.

To address the third research question of this thesis in **Chapter 9** we estimate the coverage of a mobile targeted digital X-ray tuberculosis screening programme for illicit drug users and homeless persons in Rotterdam with Chao's truncated heterogeneity model and Zelterman's truncated Poisson mixture model. We show how truncated population estimation models can be used relatively easily when only one data source is available. In **Chapter 10** we re-examine 19 published and current international three-source log-linear capture-recapture datasets on estimating tuberculosis and other infectious disease incidence and completeness of registration, with various truncated population estimation models and discuss the performance of these alternative models.

The General Discussion **(Chapter 11)** provides answers to the research questions of this thesis and discusses aspects of the feasibility and validity of three-source log-linear capture-recapture analysis and related truncated population estimators for estimating the incidence of tuberculosis and other infectious diseases, and lists the conclusions and recommendations.

1.6 References

1. Rothman K, Greenland S. *Modern epidemiology*. Philadelphia,: Lippincott, Williams and Wilkins, 1998.
2. Nanan DJ, White F. Capture-recapture: reconnaissance of a demographic technique in epidemiology. *Chronic Dis Can* 1997; 18: 144-8.
3. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol* 1992; 135: 1060-7.
4. Desenclos JC, Bijkerk H, Huisman J. Variations in national infectious diseases surveillance in Europe. *Lancet* 1993; 341: 1003-6.
5. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
6. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-64.
7. Cochran, WG. Laplace's Ratio Estimator. In: David HA, ed. *Contributions to survey sampling and applied statistics*. New York: Academic Press, 1978: pp 3-10.
8. LaPlace SP. Sur les naissances, les mariages, et les morts. In: *Histoire de l' Academie Royale des Sciences Année 1783*. Paris : 1786 : p 693.
9. Petersen CG. The yearly immigration of young place into the Limfjord from the German sea. *Rep Dan Biol Stat* 1896; 6: 1-48.
10. Schnabel ZE. The estimation of the total fish population of a lake. *Am Math Mon* 1938; 45: 348-52.
11. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses. *U California Public Stat* 1951; 1: 131-60.
12. Darroch JN. The multiple-recapture census I. Estimation of a closed population. *Biometrika* 1958; 45: 343-59.
13. Seber GA. A review of estimating animal abundance. *Biometrics* 1986; 42: 267-92.
14. Fienberg SE. Bibliography on capture-recapture modelling with application to census undercount adjustment. *Surv Methodol* 1992; 18: 143-54.
15. Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent of registration. *J Am Stat Assoc* 1949; 44: 101-15.
16. Wittes J, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. *J Chronic Dis* 1968; 21: 287-301.
17. Wittes JT. Applications of a multinominal capture-recapture method to epidemiological data. *J Am Stat Assoc* 1974; 69: 93-7.
18. Wittes JT, Colton T, Sidel VW. Capture-recapture models for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis* 1974; 27: 25-36.
19. Fienberg SE. The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
20. Bishop YM, Fienberg SE, Holland PW. *Discrete multivariate analysis*. Cambridge: MIT-Press, 1975.
21. Pollock KH. Modelling capture, recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. *J Amer Stat Assoc* 1991; 86: 225-38.
22. Seber GA. A review of estimating animal abundance II. *Int Stat Rev* 1992; 60: 129-66.
23. Dubos R, Dubos J. *Tuberculosis, Man and Society*. Boston: Little, Brown and Company, 1952: p 6.
24. Styblo K, Rouillon A. Estimated global incidence of smear-positive tuberculosis. Unreliability of officially reported figures on tuberculosis. *Bull Int Union Tuberc* 1981; 56: 118-26.
25. Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. Global burden of tuberculosis. *JAMA* 1999; 282: 677-86.
26. World Health Organization. *Global tuberculosis control: surveillance, planning, financing. WHO Report 2004*. Geneva: World Health Organization, 2004.
27. Marier R. The reporting of communicable diseases. *Am J Epidemiol* 1977; 105: 587-90.
28. Washko RM, Frieden TR. Tuberculosis surveillance using death certificate data, New York City, 1992. *Public Health Rep* 1996, 111: 251-5.
29. Weinbaum C, Ruggiero D, Schneider E, McCray E, Onorato IM, Phillips L, Donnel HD. TB reporting. *Public Health Rep* 1998; 113: 288.
30. Curtis AB, McCray E, McKenna M, Onorato IM. Completeness and timeliness of tuberculosis case reporting. A multistate study. *Am J Prev Med* 2001; 20: 108-12.
31. Davies PD, Darbyshire J, Nunn P, Byfield SP, Fox W, Citron KM, Raynes RH. Ambiguities and inaccuracies in the notification system for tuberculosis in England and Wales. *Community Med* 1981; 3: 108-18.

32. Bradley BL, Kerr KM, Leitch AG, Lamb D. Notification of tuberculosis: can the pathologist help? *BMJ* 1988; 297:595-96.
33. Sheldon CD, King K, Cock H, Wilkinson P, Barnes NC. Notification of tuberculosis: how many cases are never reported. *Thorax* 1992; 47: 1015-8.
34. 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.
35. Devine MJ, Aston R. Assessing the completeness of tuberculosis notification in a health district. *Commun Dis Re p CDR Rev* 1995; 5: R137-140.
36. Brown JS, Wells F, Duckworth G, Paul EA, Barnes NC. Improving notification rates for tuberculosis. *BMJ* 1995; 310: 974.
37. Mukerjee AK: Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data source *Commun Dis Public Health* 1999, 2: 143-4.
38. Grove A, Valentine J, MacDonald T, Roworth M, Winter J. The ascertainment and management of tuberculosis in Tayside, Scotland during 1993-94. *Health Bull (Edinb)* 2001; 59: 233-7.
39. Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003; 1: 31.
40. Balogun MA, Wall PG, Noone A. Undernotification of tuberculosis in patients with AIDS. *Int J STD AIDS* 1996; 7: 58-60.
41. Ferguson A, Bennet D, Conning S. Notification of tuberculosis in patients with AIDS. *J Public Health Med* 1998; 20: 218-20.
42. Guérin B, Joly, Vallée E. La déclaration de la tuberculose dans un hôpital Parisien. *Bull Epid Hebdo* 1992; 53: 249.
43. Decludt B, Vaillant V, Chambaud L. Evaluation de la qualité de la déclaration obligatoire de la tuberculose dans 16 Départements Français. *Bull Epid Hebdo* 1995;12:51-53.
44. Denic L, Lucet JC, Pierre J, Deblangy C, Kosmann MJ, Carbonne A, Bouvet E. Notification of tuberculosis in a university hospital. *Eur J Epidemiol* 1998; 14: 339-42.
45. Migliori GB, Spabevello A, Ballardini L, Neri M, Gambarini C, Moro ML, Trnka L, Raviglione MC. Validation of the surveillance system for new cases of tuberculosis in a province of northern Italy. Varese Tuberculosis Study Group. *Eur Respir J* 1995; 8: 1252-8.
46. Buiatti E, Acciai S, Ragni P, Tortoli E, Barbieri A, Cravedi B, Santini MG. [The quantification of tuberculous disease in an Italian area and the estimation of underreporting by means of record linkage]. *Epidemiol Prev* 1998; 22: 237-41.
47. Moro ML, Malfait P, Salamina G, D'Amato S. [Tuberculosis in Italy: available data and open questions]. *Epidemiol Prev* 1999; 23: 27-36.
48. Gallo G, Majori S, Poli A, Pascu D, Zolin R, Piovesan C, Gazzola B. [Evaluation of the underreporting of tuberculosis through the linkage of 5 different information sources]. *Ann Ig* 2000; 12: 365-71.
49. World Health Organization. *Global Tuberculosis Control. WHO Report 2003*. Geneva: World Health Organization, 2003.
50. Gutiérrez M A, Castilla J, Noguer I, Díaz P, Arias J, Guerra L. [Anti-tuberculosis drug consumption as an indicator of the epidemiological situation of tuberculosis in Spain]. *Gac Sanit* 1999; 13: 275-81.
51. Criado-Alvarez JJ, Sanz Cortes J. [Use of pyrazinamide as an indicator of tuberculosis epidemiology in Castilla-La Mancha]. *Rev Clin Esp* 2004; 204: 298-302.
52. Loenhout-Rooijackers JH van, Leufkens HG, Hekster YA, Kalisvaart NA. Pyrazinamide use as a method to estimate under-reporting of tuberculosis. *Int J Tuberc Lung Dis* 2001; 5: 1156-60.
53. Baily GV. Tuberculosis Prevention Trial, Madras. *Ind J Med Res* 1980; 72: S1-74.
54. Medical Research Council Tuberculosis and Chest Diseases Unit. National survey of tuberculosis notifications in England and Wales 1978-9. *BMJ* 1980; 281: 895-8.
55. Medical Research Council Tuberculosis and Chest Diseases Unit. National survey of notifications of tuberculosis in England and Wales in 1983. *BMJ* 1985; 291: 658-61.
56. Medical Research Council Cardiothoracic Epidemiology Group. National survey of notifications of tuberculosis in England and Wales in 1988. *Thorax* 1992; 47: 770-5.
57. Kumar D, Watson JM, Charlett A Nicholas S, Darbyshire JH. Tuberculosis in England and Wales in 1993: results of a national survey. *Thorax* 1997; 52: 1060-7.
58. Rose AM, Watson JM, Graham C, Nunn AJ, Drobniewski F, Ormerod LP, Darbyshire JH, Leese J; Public Health Laboratory Service/British Thoracic Society/Department of Health Collaborative Group. Tuberculosis at the end of the 20th century in England and Wales: results of a national survey in 1998. *Thorax* 2001; 56: 173-9.
59. Styblo K. The relationship between the risk of tuberculosis infection and the risk of developing infectious tuberculosis. *Bull Int Union Tuberc Lung Dis* 1985; 60: 117-9.
60. Murray CJ, Styblo K, Rouillion A. Tuberculosis in developing countries: burden, intervention and cost. *Bull Int Union Tuberc Lung Dis* 1990; 65: 6-24.

61. Rieder HL. Methodological issues in the estimation of the tuberculosis problem from tuberculin surveys. *Tuberc Lung Dis* 1995; 76: 114-21.
62. Maggini M, Salmoaso S, Alegiani SS, Caffari B, Raschetti R. Epidemiological use of drug prescriptions as markers of disease frequency: an Italian experience. *J Clin Epidemiol* 1991; 44: 1299-1307.
63. Van Buynder P. Enhanced surveillance of tuberculosis in England and Wales: circling the wagons? *Commun Dis Public Health* 1998; 1: 219-20.
64. Trepka MJ, Beyer TO, Proctor ME, Davis JP. An evaluation of the completeness of tuberculosis case reporting using hospital billing and laboratory data; Wisconsin 1995. *Ann Epidemiol* 1999; 7: 419-23.
65. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-3.
66. San Gabriel P, Saiman L, Kaye K, Silin M, Onorato I, Schulte J. Completeness of pediatric TB reporting in New York City. *Public Health Rep* 2003; 118: 144-53.
67. Shanks NJ, Lambourne A, Kuhaymi RA, Humphries M, Sanford JRA. A new approach to tuberculosis notification. *J Epidemiol Community Health* 1984; 38: 331-4.
68. Yokoe DS, Subramanyan GS, Nardell E, Sharnprapai S, McCray E, Platt R. Supplementing tuberculosis surveillance with automated data from health maintenance organizations. *Emerg Infect Dis* 1999; 5: 779-87.
69. Yokoe DS, Coon SW, Dokholyan R, Iannuzzi MC, Jones TF, Meredith S, Moore M, Philips L, Ray W, Schech S, Shatin D, Platt R. Pharmacy data for tuberculosis surveillance and assessment of patient management. *Emerg Infect Dis* 2004; 10: 1426-31.
70. Centers for Disease Control and Prevention. Surveillance of tuberculosis and AIDS co-morbidity – Florida 1981-1993. *MMWR Morb Mortal Wkly Rep* 1996; 2: 39-41.
71. Ferrer Evangelista D, Ballester Diez F, Perez-Hoyos S, Igual Adell R, Fluixa Carrascosa C, Fullana Monllor J. [Incidence of pulmonary tuberculosis: application of the capture-recapture method]. *Gac Sanit* 1997; 11: 115-21.
72. Iváñez Gimeno L, Martínez Navarro JF. [Evaluation of epidemiological surveillance of respiratory tuberculosis in the province of Seville]. *Bol Epidemiol Sem* 1997; 5: 241-4.
73. Sanghavi DM, Gilman RH, Lescano-Guevara AG, Checkley W, Cabrera LZ, Cardenas V. Hyperendemic pulmonary tuberculosis in a Peruvian shantytown. *Am J Epidemiol* 1998; 148: 384-9.
74. Pérez Ciorda I, Castanera Moros A, Ferero Cáncer. [Tuberculosis in Huesca. Use of the capture-recapture method]. *Rev Esp Salud Publica* 1999; 73: 403-6.
75. Mayoral Cortes JM, Garcia Fernandez M, Varela Santos MC, Fernandez Merino JC, Garcia Leon J, Herrera Guibert D, Martinez Navarro F. Incidence of pulmonary tuberculosis and HIV co-infection in the province of Seville, Spain, 1998. *Eur J Epidemiol* 2001; 17: 737-42.
76. Iglesias Gozalo MJ, Rabanaque Hernández MJ, Gómez Lópes LI. [Tuberculosis in the Zaragoza province. Estimation by means of capture-recapture method]. *Rev Clin Esp* 2002; 202: 249-54.
77. Tejero Encinas S, Asensio Villahoz P, Vaquero Puerta JL. [Epidemiological surveillance of pulmonary tuberculosis treated at the specialized care level based on 2 data sources, Valladolid; Spain]. *Rev Esp Salud Publica* 2003; 77: 211-20.
78. Iñigo J, Arce A, Martin-Moreno JM, Herruzo R, Palenque E, Chaves F. Recent transmission of tuberculosis in Madrid: application of capture-recapture analysis to conventional and molecular epidemiology. *Int J Epidemiol* 2003; 32: 763-9.
79. Cailhol J, Che D, Jarlier V, Decludt B, Robert J. Incidence of tuberculous meningitis in France, 2000: a capture-recapture analysis. *Int Tuberc Lung Dis* 2005; 9: 803-8.
80. Guernier V, Guégan JF, Deparis X. An evaluation of the actual incidence of tuberculosis in French Guiana using a capture-recapture model. *Microbes Infect* 2006; 8: 721-7.
81. Baussano I, Bugiani M, Gregori D, Van Hest R, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung Dis* 2006; 10: 415-21.
82. Van Hest NA, Smit F, Baars HW, De Vries G, De Haas PE, Westenend PJ, Nagelkerke NJ, Richardus JH. Completeness of registration of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiol Infect* 2006; Published on-line: 7 December 2006; doi:10.1017/S0950268806007540.
83. Van Hest NA, Story A, Grant A, Antoine D, Crofts JP, Watson JM. Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999 -2002 (in preparation).

2

Methodology of capture-recapture analysis and application for epidemiological studies

2.1 Methodology of capture-recapture analysis

2.1.1 Introduction

For epidemiologists, prevalence and incidence rates are fundamental components of their discipline but these data often are inaccurate due to classification errors, i.e. under-ascertainment of the true number of persons or events and the presence of false-positive cases. Standardised means to evaluate and to adjust prevalence and incidence rates for the degree of under-ascertainment enable a more accurate and meaningful presentation of figures, comparison of data from different settings and analysis of trends.[1] Wittes and colleagues transferred capture-recapture methods used by ecologists to adjust wildlife population estimates to epidemiology to estimate prevalence or incidence. The use of simple two-source capture-recapture models for epidemiological data is often limited by violation of the underlying capture-recapture assumptions, resulting in biased estimates, and log-linear models were developed for epidemiological applications that partly address this problem.[2-6] An overview of the application of capture-recapture analysis to human epidemiology is given elsewhere.[7,8]

2.1.2 Assumptions underlying capture-recapture analysis

For a capture-recapture estimate to be valid in human epidemiology a number of assumptions should be respected: perfect record-linkage (i.e. no misclassification of records), a closed population (i.e. no immigration or emigration in the time period studied), a homogeneous population (i.e. no subgroups with markedly different probabilities of being observed and re-observed) and independent registers (i.e. the probability of being observed in one register is not affected by being –positive dependence- [or not being –negative dependence-] observed in another).[8] In two-source capture-recapture analysis the last assumption is crucial because it is impossible to check independence mathematically and it relies on the users to make the plausibility judgement. Dependencies can cause under-estimation (in case of positive dependence) or over-estimation (in case of negative dependence).[1] Heterogeneity of the population and violation of the perfect record-linkage and closed population assumptions can also cause bias in both directions. In human conditions violation to some degree of most of these assumptions, especially that registers are independent and the population is homogeneous, is unavoidable and limitations of capture-recapture analysis are described.[1,7,9,10-14]

In addition to the assumptions mentioned, it is important that the various registers only contain individuals with the condition under study, i.e. the registers should not include false-positive records. In other words, the specificity and positive predictive value of the registers should ideally be 100%. A low positive predictive value results in overestimation of the true population size. Finally, the individuals under study should be captured within the time and space defined by the investigation.

Various efforts can reduce violation of the assumptions underlying capture-recapture studies on human epidemiology. Complete and good quality information on the personal identifiers of individuals in the different registers will limit violation of the perfect record-linkage assumption. Collection of data within a short period of time will

minimise violation of the closed population assumption. Violation of the homogeneity assumption can be handled by stratification of the population into more homogeneous strata, perform capture-recapture analysis for each of the distinct subgroups and subsequently add the results for the total estimate An alternative is to include covariates with a strong relationship to the probability of capture in a log-linear covariate capture-recapture model.[15,16] A third approach, if possible, is to model the heterogeneity, e.g. with logistic regression.[17,18] Violation of the independency assumption can be partially identified and controlled when more than two sources are linked, allowing for sources to be examined pair-wise, i.e. two at a time.[5] In the absence of source dependence the possible pair-wise capture-recapture estimates of the total number of cases should be reasonably similar. Positive dependence between two of the lists can be suspected when a pair-wise estimate is considerably lower than the other pair-wise estimates. In the three-source capture-recapture approach according to Fienberg, pair-wise dependencies can be incorporated in the log-linear model as interactions. In case of three sources, three-way interaction, i.e. dependence between all three registers, is assumed to be zero, or, in case of multiple sources, highest-order interaction, i.e. dependence between all sources, is assumed to be zero. When interactions are incorporated in the log-linear model, especially when all interactions are incorporated in a so-called 'saturated' model, with no degrees of freedom –df's– left, three-way interaction cannot be excluded and the assumption that this interaction is zero has been called an "act of faith".[10,12] The positive predictive value can be increased through an adequate and unambiguous case-definition, uniform for all sources, cross-validation through record-linkage with other related data sources and identification and exclusion of false-positive cases. Examination of a period of time before and after the study episode can correct for late registration

2.1.3 The two-source capture-recapture model

A two-source capture-recapture problem with registers A and B can be graphically presented as in table 2.1.

Table 2.1 The two-source capture-recapture problem

	Register B		
Register A	Not observed	Observed	Total register A
Not observed	$\hat{n}_{00}$	n_{01}	
Observed	n_{10}	n_{11}	N_A
Total register B		N_B	

The numbers of cases only on register A, only on register B, on both registers and on neither register, can be expressed as n_{10}, n_{01}, n_{11} and $\hat{n}_{00}$ respectively. The number of cases on register A, N_A, is $n_{10} + n_{11}$ and the number of cases on register B, N_B, is $n_{01} + n_{11}$. The total observed population on at least one register, the case-ascertainment, equals $n_{10} +$

$n_{01} + n_{11}$. The aim is to estimate the number of cases not observed in both registers, $\hat{n}_{00}$. The estimated total number of cases, $\hat{N}$, is the observed number of cases plus the estimated unobserved number of cases. When the basic assumptions outlined in section 2.1.2 hold $\hat{n}_{00}$ can be expressed as

$$\hat{n}_{00} = \frac{n_{10} \times n_{01}}{n_{11}} \qquad (2.1)$$

and $\hat{N}$ as

$$\hat{N} = \frac{N_A \times N_B}{n_{11}} \qquad (2.2)$$

Equation (2.2) is known as the Petersen estimator. Approximately unbiased estimates of $\hat{N}$ are expected when the registers are large. The correction for bias caused by small registers, the Nearly Unbiased Estimator proposed by Chapman, can be expressed as

$$\hat{N} = \left[\frac{(N_A + 1) \times (N_B + 1)}{(n_{11} + 1)} \right] - 1 \qquad (2.3)^{19,20}$$

The confidence interval is calculated as $\hat{N} \pm$ 1.96 times the standard error. The mathematical framework of the two-source capture-recapture model is explained in more detail elsewhere.[8] The underlying assumptions and the effect of violation of these assumptions are discussed in section 2.1.2. In epidemiology, due to the possible presence of uncontrolled dependence between the sources and heterogeneity of the population resulting in biased estimates, two-source capture-recapture analysis is regarded as rarely appropriate by some.[8] Others consider this method useful under certain circumstances, e.g. when the likely direction of the bias caused by violation of the underlying assumptions can be predicted and plausible lower and upper boundaries of the prevalence or incidence of a disease can be estimated.[7,21,22] A variation of the two-source method, using a single sample has been described.[23]

2.1.4 The log-linear capture-recapture model

In the context of multiple registers another capture-recapture approach called log-linear modelling allows for controlling specific forms of dependence and heterogeneity, making these multiple-source log-linear capture-recapture models more powerful.[7] The mathematical framework of the multiple-source log-linear capture-recapture model is explained in detail elsewhere.[8] The underlying assumptions and the effect of violation of these assumptions are discussed in section 2.1.2. Basically, the log-linear model transforms the two-source capture-recapture model in a model for the logarithms of the observed counts which is linear in a set of parameters. Possible interaction between registers A and B will change the log-linear model and this modification can be incorporated in the model as an interaction parameter λ^{AB}. This type of log-linear model has become the standard form of analysis for contingency tables and was proposed and

developed for capture-recapture analysis by Fienberg.[3] With three registers there are eight possible combinations of these registers in which cases do or do not appear. The general model uses eight parameters, the common parameter (the logarithm of the number expected to be in all lists), three 'main effects' parameters (the log odds ratios against appearing in each list for cases who appear in the others), three 'two-way interactions' or second order effect parameters (the log odds ratios between pairs of lists for cases who appear in the other), and a 'three-way' interaction parameter. For three registers, A with i levels, B with j levels, C with k levels, the natural logarithm (ln or $\log_e$) of expected frequency F_{ijk} for cell ijk, ln F_{ijk}, can be denoted as

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC} \qquad (2.4)$$

where θ is the common parameter , λ^A, λ^B, and λ^C are the main effect parameters, λ^{AB}, λ^{AC} and λ^{BC} are the second order effect (two-way interaction) parameters and λ^{ABC} is the highest order effect (three-way interaction) parameter. The value of this last three-way interaction parameter can not be tested from the study data and is assumed to be zero. Assumptions about the other parameters can be tested, although these tests may not be very powerful for small samples.

Three types of log-linear models can be recognised. Firstly, the 'independent model' which assumes that all registers are independent. Secondly, models that are equivalent to two independent registers or two independent subsets of registers. Finally, a 'saturated' model that incorporates all possible interactions, including possible three-way interaction. To assess how the various log-linear models fit the data (model fitting) the log likelihood-ratio test, also known as G^2 or deviance, is used, denoted as

$$G^2 = -2\sum \text{Obs}_j \ln[\text{Obs}_j / \text{Exp}_{ji}] \qquad (2.5)$$

where Obs_j is the observed number of individuals in each cell j, and Exp_{ji} is the expected number of individuals in each cell j under model i. The lower the value of G^2 the better is the fit of the model. In the log-linear estimation procedure after model fitting follows model selection, i.e. to identify the models that are clearly wrong and select from a number of acceptable models the most appropriate. For model selection, apart from previous knowledge and expectations about dependencies between registers and heterogeneity of the population, formal procedures based upon likelihood-ratio tests, known as information criteria, can be used. One of these procedures is Akaike's Information Criterion (AIC)[24] which can be expressed as

$$\text{AIC} = G^2 - 2\,[\text{df}] \qquad (2.6)$$

The first term, G^2, is a measure of how well the model fits the data and the second term, 2 [df], is a penalty for the addition of parameters (and hence model complexity). Another information criterion is the Bayesian Information Criterion (BIC)[25] which can be expressed as

$$\text{BIC} = G^2 - [\ln N_{\text{obs}}]\,[\text{df}] \qquad (2.7)$$

where N_{obs} is the total number of observed individuals. Relative to the AIC, the BIC penalises complex models more heavily. In general, in the log-linear capture-recapture estimation procedure the least complex, i.e. the least saturated (in other words the most

parsimonious) model, whose fit appears adequate, is preferred.[11] Since the G^2 of the saturated model is zero and has no degrees of freedom left, the AIC and BIC are also zero and models with a negative AIC and BIC are preferred although this does not necessarily mean that the estimate is correct. When the saturated model is selected by any criterion the investigator should be particularly cautious about using the associated outcome.[7,26] However, when external considerations do not justify the presumption of plausible interactions of sources in the simpler models, some advocate the saturated model.[27] The confidence interval around log-linear estimates can be constructed based on likelihood-ratio statistics.[28,29] However, any confidence interval only adjusts for sampling fluctuation but it does not adjust for any uncertainty as to whether the underlying assumptions are violated.[27] With an increasing number of registers, the number of possible capture-recapture models rapidly increases. Programs for the analysis of log-linear models exist in most large statistical computer packages, such as S+, SAS and SPSS, and some have been specially developed for capture-recapture analysis, e.g. GLIM,[30] MARK[31] and CARE.[32]

2.1.5 Truncated models

As an alternative to the more conventional two-source and log-linear multiple-source capture-recapture analysis, so-called truncated models have been employed, assuming a specific distribution of the observed data, e.g. Poisson, binomial or a mixture of different distributions.[7] Truncated models, such as Zelterman's truncated Poisson mixture model and Chao's heterogeneity and bias-corrected homogeneity models[33-35] can be applied to frequency counts of observations of cases in a single register or multiple registers, with the aim to estimate the number of unobserved persons in the (truncated) zero-frequency class, based upon information of the lower frequency classes. These models have been used in genetic epidemiology[36] and social sciences, e.g. to estimate the size of hidden populations of illicit drug users and homeless persons.[37-40] The simple estimators do not need statistical packages, Zelterman's model supposedly allows for greater flexibility and applicability on real life data and the Zelterman and Chao models are arguably more robust to violation of the homogeneity assumption because they are partly based upon the lower frequency classes, assumed to have more resemblance to the zero frequency class. Despite obvious violation of other underlying assumptions, especially the independent registers assumption in case of multiple sources or the constant individual probability of re-observation assumption in case of a single source[39,40], truncated models have performed well when compared to log-linear capture-recapture estimates.[41] An overview of a range of truncated models is given elsewhere.[42]

2.2 Application and limitations of capture-recapture analysis for epidemiological studies

2.2.1 Introduction

Several steps are important in planning, applying, presenting and evaluating capture-recapture techniques in epidemiological studies:[43,44]

1. The purpose of the study and the required accuracy of the data should be described.
2. Appropriate sources for capturing cases should be selected.
3. Possible relationships between the selected sources should be investigated and described as well as their influence on the capture-recapture results.
4. An unambiguous and uniform case-definition for the various sources should be used.
5. The accuracy of diagnosis and disease classification in each of the sources should be examined.
6. The accuracy of record-linkage should be described.
7. The case-ascertainment, the distribution of the cases over the various registers, the selected capture-recapture model and the estimate of the number of missing cases, and thus the total number of cases, should be given, from which prevalence or incidence rates can be calculated.
8. The limitations of the capture-recapture methodology in epidemiology should be addressed.

Preferably preparatory fieldwork and explorative research should be performed, if possible including a small scale pilot capture-recapture study investigating feasibility, limitations and costs, ideally with the possibility of validation of the estimates against known data from a census or survey.[59] This preparation is still required while planning a capture-recapture study on a particular disease when such a study was performed successfully in the past or elsewhere, using similar data sources and resulting in a fitting capture-recapture model that produced a credible estimate. In the past and abroad, the characteristics of disease registers, legislation or guidelines and organisation of disease diagnosis and surveillance may be different. Consultation of a statistician, either a bio-statistician or social scientist, with experience in capture-recapture analysis should be part of the planning of a capture-recapture study. In case of investigating large data sources collaboration with a data manager with experience in sophisticated computerised record-linkage is required.

2.2.2 Purpose and required accuracy of the study

The aim of the study and the intended use of the reported estimates should be described, including the likely consequences for utilisation of under-estimates or over-estimates. Some objectives need accurate estimates while others accept under-estimates or over-estimates serving as the upper or lower bound of the true population size.[7]

2.2.3 Source-selection and number of sources

Selection of suitable data sources from the available information sources is needed in order to obtain valid capture-recapture estimates.[45,46] Potential data sources can be identified from the literature or the experience from other researchers. All cases should have a chance to be recorded in each one of the selected data sources. All selected data

sources should contain sufficient identifiers of the cases for reliable record-linkage. The data sources should be able to provide overlap information among the sources, as this is the key component of capture-recapture analysis, and therefore complementary or mutually exclusive data sources should not be used.[45] As later outlined in section 2.2.6 ideally the positive predictive value of each of the data sources should be 100%. However, in epidemiological capture-recapture studies often existing data sources, not designed for capture-recapture analysis, are used, possibly containing poor quality data, leading to poor capture-recapture outcomes.[10]

The number of available data sources may be limited, restricting the opportunity of choice in source selection. The use of only two data sources prevents mathematical assessment of possible interdependence, potentially causing under-estimation or over-estimation.[45] The use of at least three data sources allows log-linear models to incorporate specific forms of source interdependence. It is neither practical to have as many data sources as possible, e.g. for budgetary constraints as additional sources cost money and time, nor beneficial, e.g. an increasing number of sources causes decreasing overlap, resulting in increasing variation of the estimates and cells in the multi-way contingency table may even contain zero cases. To reduce the number of data sources, one can collapse multiple sources into for example three logical sources, although this is at the expense of overlap information. In general, the use of three to five data sources is recommended,[7,32,44,45,47] although some argue that the use of more than three data sources does not substantially alter the absolute value of or confidence in the obtained capture-recapture estimate.[46]

2.2.4 Relationships between the selected sources

Possible relationships between the selected sources should be explored because for reliable capture-recapture estimates absence of specific forms of dependence between the data sources is one of the crucial underlying assumptions. As outlined in section 2.1.2 possible positive or negative dependence between the data sources results in under-estimation and over-estimation respectively. If data sources are likely to be dependent three choices can be made: to discard the sources, to combine them or to use log-linear modelling.

2.2.5 Case-definition

A clear standardised case-definition, which cannot be changed throughout the investigation period, is needed and should be uniform and consistent among each of the sources selected, such as laboratory-based sources and clinician-based sources.[45,48] Different specificity of the case-definition results in different rates for false-positive cases between data sources, causing invalid estimates of the total number of cases.[9,49,50] Also the target population, the geographical area and the time interval from which the selected sources capture their cases, should be identical.

2.2.6 Accuracy of diagnosis and disease classification

Ideally the positive predictive value of each of the data sources should be 100% but in epidemiology not many sources will meet this condition.[11] Apart from under-ascertainment of cases, routine systems of disease surveillance, such as morbidity and

mortality statistics, are prone to errors in disease diagnosis or disease classification and coding, resulting in registration of individuals with diseases other than the disease of interest (or none) by one or several sources of case-ascertainment (false-positive diagnosis or misclassification). False-positive cases as a result of misdiagnosis or misclassification lead to over-estimation of cases by capture-recapture analysis.[51] One approach to minimise bias caused by misclassification is to combine diagnoses between which misclassification is common (e.g. colon and rectum cancer or cervical and endometrium cancer). Another approach is to identify and exclude false-positive cases through record-linkage with other related sources. Alternatively, the positive predictive value reported in previous validation studies, e.g. post-mortem examinations, can be used to correct for the expected over-estimation.

2.2.7 Record-linkage

The record-linkage procedure should be reliable because it performs one of the most important steps in the capture-recapture application, i.e. accurate determination of the number of overlap cases, and is one of the underlying capture-recapture assumptions as outlined in section 2.1.2. Inaccurate record-linkage can substantially alter the size of the observed and unobserved fractions.[13]

Imperfect record-linkage can cause incorrect linking of different individuals (false-positive links, also called homonym errors or "mis-link") or failure to identify the same individual in different sources (false-negative links, also called synonym errors, or "missed link"). False-positive links will lead to under-estimation of the case counts and false-negative links will lead to over-estimation of the case counts.[52,53] When both linkage errors are present, the antagonistic effects of false-positive and false-negative links on estimated case counts may partly or fully cancel out.[54] Imperfect record-linkage can result from incomplete registration of personal identifiers in different data sources, imperfect registration (e.g. due to clerical errors such as typing mistakes) or due to changes in the variables used for record-linkage, such as family name or address. Imperfect record-linkage is of particular concern in situations in which restrictive confidentiality and privacy rules allow only very limited registration of personal identifiers.[11,54] Record-linkage may even be impossible for ethical or legal reasons.[9] As outlined in section 2.2.3 sufficient information for reliable record-linkage should be a source selection criterion.

Like a numbered bird ring in animal population studies, in human conditions, ideally a unique identifier should be used such as a social security number that appears on all records of an individual. In the absence of such a 'mark', variables such as the name of a person, date of birth, age, sex or (part of the) postal code can be used to perform record-linkage. Considered individually these are proxy identifiers and record-linkage often depends on their use in combination. It has been suggested that the combination of first name, family name and date of birth is sufficient for adequate linkage.[46] The use of names for adequate record-linkage is only possible in countries or areas where the literacy rate is very high and record-keeping is reliable. But even in these countries, when a disease, such as malaria or tuberculosis, is common among persons with a foreign or ethnic background, variation in spelling of unfamiliar and complicated names, e.g. double-barrelled names, can cause imperfect linkage. A problem will also arise when one source registers married women under the husband's name and another source under the maiden name. Sophisticated cryptographic computer software has been developed for the

refinement of record-linkage techniques. These programmes 'translate' family names in a phonetic expression or code, a procedure called name or phonetic compression, increasing the probability of linking all possible variants of names.[55] Also to meet privacy restrictions names can be reduced into a fixed format code, e.g. by eliminating vowels, regarding certain consonants as silent and regarding others as equivalent. Examples are the Soundex code, the Dolby code, the New York State Information and Intelligence code and the Oxford name compression algorithm. In the absence of names adequate record-linkage often can be done using the other variables mentioned earlier.[56]

There are several record-linkage techniques: deterministic or exact record linkage, relaxing exact record-linkage, probabilistic record-linkage, and a combination of these. Exact record-linkage only links individuals from different data sources with exactly the same field values ('all-or-none') while relaxing record-linkage allows a minimal degree of discrepancy. Purely exact record-linkage is not recommended, because even for 'perfect' data sources inaccuracies in recording, transcription and keying, such as coding mistakes or typing errors, occur during data collection, resulting in false-negative record-linkage. The approach of probabilistic record-linkage is to assign different probabilities or weights to the variables as some provide more information and are more reliable than others. The perceived discriminating power of the various personal identifiers results in differential weighting of the amount of agreement or disagreement.[56-58] Computer software has been developed for exact and probabilistic record-linkage.

2.2.8 Case-ascertainment and capture-recapture analysis

After record-linkage the case-ascertainment (i.e. the number of cases known to at least one of the registers) should be presented, including the distribution of the cases over the different sources, in total and possibly stratified by important covariates. The estimates and associated confidence intervals of each of the eight log-linear capture-recapture models between independent and saturated model should be presented with degrees of freedom, goodness-of-fit criterion and information criteria. Coherence of the estimates with previous knowledge, e.g. census data, surveys or other capture-recapture study estimates, should be discussed and the selection of the preferred model explained. In case of three-source capture-recapture analysis the internal consistency of the log-linear capture-recapture estimate should be examined through comparison of all two-source estimates and those of each source versus all others pooled.[12] The outcome of the capture-recapture analysis should be discussed in the context of the aim of the study.

2.2.9 Limitations of capture-recapture analysis in epidemiology

Most of the limitations of capture-recapture analysis pertain to the possibility, almost the certainty, of violation of the underlying assumptions.[7] In contrast to animal population studies the assumptions as outlined in section 2.1.2 are unlikely to be satisfied in epidemiological applications.[12,14] In epidemiology capture-recapture analysis often uses existing administrative registers, not designed for capture-recapture analysis, instead of random surveys of the population according to a common protocol.[10] The accuracy of the registers used, such as correct diagnosis and coding and sufficient information for appropriate record-linkage, is important. In capture-recapture analysis, errors are highly likely to have a more than additive effect on estimates. Registers containing poor quality

data lead to poor capture-recapture outcomes.[7,10] Dependence of sources is often a problem in epidemiological capture-recapture applications. Such dependence can result from co-operation between the agencies that keep the different registrations, exchange of information or a more or less predictable flow of patients along various institutions due to referral. The probability of ascertainment by any particular source should be equal but in epidemiological settings often it is not, due to the intrinsic nature of human variation, e.g. socioeconomic differences or variation of severity of disease. Also human populations are rarely closed.

It has been argued that estimates from capture-recapture studies in epidemiology are wholly unreliable unless supported by a wide variety of sensitivity analyses, and by careful medical and social discussion of variability between individuals, and the reasons why a particular individual may fail to be recorded in a particular register. Matching of only two registers has even been called "mostly an exercise in self-deception, an unscientific, uncritical act of faith in the absolute truth of untested and implausible assumptions".[10] Also for the application of multiple-source log-linear estimators for any particular observed data on real populations some claim that "in no sense there is any proof or re-assurance that this results in a valid estimate, or even necessarily produces an estimate closer to the true value than some alternative approach".[7] Although many apparently successful capture-recapture studies have been published, only few have been reported to have failed and 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.[47]

2.3 References

1. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995; 6: 42-8.
2. Wittes J, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. *J Chronic Dis* 1968; 21: 287-301.
3. Fienberg SE. The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
4. Wittes JT. Applications of a multinominal capture-recapture method to epidemiological data. *J Am Stat Assoc* 1974; 69: 93-7.
5. Wittes JT, Colton T, Sidel VW. Capture-recapture models for assessing the completeness of case ascertainment when using multiple information sources. *J Chronic Dis* 1974; 27: 25-36.
6. Bishop YM, Fienberg SE, Holland PW. *Discrete multivariate analysis*. Cambridge: MIT-Press, 1975.
7. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-64.
8. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
9. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *Int J Epidemiol* 1994; 23: 1322-3.
10. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *J Clin Epidemiol* 1999; 52: 909-14.
11. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1999; 25: 474-8.
12. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* 2000; 152: 771-9.
13. Jarvis SN, Lowe PL, Avery A, Levene S, Cormack RM. Children are not goldfish: mark-recapture techniques and their application to injury data. *Inj Prev* 2000; 6: 46-50.

14. Tilling K. Capture-recapture methods-useful or misleading? Int J Epidemiol 2001;30:12-4.
15. Hook EB, Regal RR. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. *Amer J Epidemiol* 1993; 137: 1148-66.
16. Tilling K, Sterne JA. Capture-recapture models including covariate effects. *Am J Epidemiol* 1999; 149: 392-400.
17. Alho JM. Logistic regression in capture-recapture models. *Biometrics* 1990; 46: 623-35.
18. Alho M, Mulry MH, Wurdeman K, Kim J. Estimating heterogeneity in the probabilities of enumeration for the dual-system estimation. *J Amer Stat Assoc* 1993; 88: 1130-36.
19. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses. *U California Public Stat* 1951; 1: 131-60.
20. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture-recapture estimate. *Biometrics* 1972; 28: 592-97.
21. Hook EB, Regal RR. Capture-recapture methods. *Lancet* 1992; 339: 742.
22. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol* 1992; 135: 1060-7.
23. Laska EM, Meisner M. A plant-capture method for estimating the size of a population from a single sample. *Biometrics* 1993; 49: 209-20.
24. Sakamoto Y, Ishiguru M, Kitigawa G. *Akaike information criterion statistics*. Tokio: KTK Scientific, 1986: pp 1-24.
25. Agresti A. *Categorical data analysis*. New York: John Wiley and Sons, 1990: p 251.
26. Regal RR, Hook EB. Marginal versus conditional versus structural source models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Stat Med* 1998; 17: 69-74.
27. Regal RR, Hook EB. The effects of model selection on confidence intervals for the size of a closed population. *Stat Med* 1991; 10: 717-21.
28. Regal RR, Hook EB. Goodness-of-fit based confidence intervals for estimates of the size of a closed population. *Stat Med* 1984;3: 287-91.
29. Cormack RM. Interval estimation for mark-recapture studies of closed populations. *Biometrics* 1992; 48: 567-76.
30. Aitken D, Anderson D, Hinde J, Francis B. *Statistical Modelling in GLIM*. Oxford: Oxford University Press, 1994.
31. White GC, Burnham KP. Program MARK: Survival estimation from populations of marked animals. *Bird Study* 1999: 46; S120-38.
32. Chao A, Tsay PK, Lin S-H, Shau W-Y, Chao D-Y. The applications of capture-recapture models to epidemiological data. *Stat Med* 2001; 20: 3123-57.
33. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; 43: 783-91.
34. Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J Statistic Plan Inf* 1988; 18: 225-37.
35. Chao A. Estimating animal abundance with capture frequency data. *J Wildl Manage* 1988; 52: 295-300.
36. Stene J. The incomplete multiple ascertainment model: assumptions, applications and alternative models. *Genet Epidemiol* 1989; 6: 247-51.
37. 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 1997: EMCDDA, pp 47-66.
38. Bohning D, Suppawattanabodee B, Kusolvitkul, W, Viwatwongkasem C. Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *Eur J Epidemiol* 2004; 19: 1075-83.
39. Smit F, Reinking D, Reijerse M. Estimating the number of people eligible for health service use. *Evaluat Prog Plan* 2002; 25: 101-5.
40. Hay G, Smit F. Estimating the number of hard drug users from needle-exchange data. *Addiction Res Theory* 2003; 11: 235-43.
41. Hook EB, Regal RR. Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *Am J Epidemiol* 1982; 116: 168-76.
42. Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; 79: 543-53.
43. Hook EB, Regal RR. Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology. *J Clin Epidemiol* 1999; 52: 917-26.
44. Hook EB, Regal RR. On the need for a 16th and 17th recommendations for capture-recapture analysis. *J Clin Epidemiol* 2000; 53: 1275-77.

45. Chang YF, LaPorte RE, Aaron DJ, Songer TJ. The importance of source selection and pilot study in the capture-recapture application. *J Clin Epidemiol* 1999; 52: 927-8.
46. Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture-recapture? *Int J Epidemiol* 2000; 29: 536-41.
47. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; 46: 515-20.
48. Buster MC, Van Brussel GH, Van den Brink W. Estimating the number of opiate users in Amsterdam by capture-recapture: the importance of case definition. *Eur J Epidemiol* 2001; 17: 935-42.
49. Nanan DJ, White F. Capture-recapture: reconnaissance of a demographic technique in epidemiology. *Chronic Dis Can* 1997; 18: 144-8.
50. Borgdorff MW, Glynn JR, Vynnycky E. Using capture-recapture methods to study recent transmission of tuberculosis. *Int J Epidemiol* 2004; 33: 905-6.
51. Brenner H. Effects of misdiagnoses on disease monitoring with capture-recapture methods. *J Clin Epidemiol* 1996; 11: 1303-07.
52. Brenner H, Schmidtmann I. Determinants of homonym and synonym rates of record linkage in disease registration. *Methods Inform Med* 1996; 35: 19-24.
53. Brenner H, Schmidtmann I. Effects of record linkage errors on disease registration. *Methods Inform Med* 1998; 37: 69-74.
54. Brenner H. Application of capture-recapture methods for disease monitoring: potential effects of imperfect record linkage. *Methods Inform Med* 1994; 33: 502-6.
55. Gill L, Goldacre M, Simmons H, Bettley G, Griffith M. Computerised linking of medical records: methodological guidelines. *J Epidemiol Community Health* 1993; 47: 316-9.
56. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. *Stat Med* 1995; 14: 499-509.
57. Newcombe HB. Handbook of Record Linkage. *Methods for Health and Statistical Studies, Administration and Business*. Oxford: Oxford University Press, 1988.
58. Jaro MA. Probabilistic linkage of large public health data files. *Stat Med* 1995; 14: 491-8.

3

Synopsis of capture-recapture studies on infectious diseases, 1997 - 2006

3.1 Application of capture-recapture analysis in studies on human diseases

An overview of the applications of capture-recapture methods in human diseases before 1997 is given elsewhere.[1,2] Apart from injury-related capture-recapture studies, these reports were categorised into four different disease groups, often diseases with a chronic character. Apparently the characteristics of most of these diseases, their patients and their registers fulfil criteria for feasibility as well as for agreement with the assumptions underlying capture-recapture analysis best. In a first group more than 30 capture-recapture studies were performed for monitoring the incidence of insulin-dependent diabetes mellitus and in this field capture-recapture analysis has become a standard method to correct case-ascertainment. A second group consists of capture-recapture studies on the frequency of birth defects, such as congenital rubella, cleft lip and cleft palate, spina bifida, Down's syndrome and other congenital anomalies. A third group of capture-recapture studies focussed on cancer, e.g. to estimate the completeness of cancer registries or to estimate breast cancer screening sensitivity. Apart from these three disease categories, capture-recapture methods were used to estimate the incidence or prevalence of various diseases such as haemophilia, myocardial infarctions, Huntington's disease, mental disease, Rett's syndrome, and vaccine-associated poliomyelitis. A brief discussion of these early capture-recapture studies published before 1997 can be found at the website http://www.pitt.edu/~yuc2/iddm.html, http://www.pitt.edu/~yuc2/birth.html, http://www.pitt.edu/~yuc2/cancer.html and http://www.pitt.edu/~yuc2/other.html for diabetes, birth defects, cancer and other diseases respectively (accessed 1 May 2007).

After 1997 more than 100 capture-recapture studies on human diseases other than infectious diseases or tuberculosis have been published in peer-reviewed journals. Most of the diseases studied are again chronic conditions and largely follow the same categories as described above. Approximately half the studies are related to diabetes mellitus. Birth defects such as neural tube defects, Down's syndrome, congenital ocular anomalies, tuberous sclerosis, brain arteriovenous malformations, heart malformations and other congenital disorders were studied and several reports were related to estimating cancer incidence or completeness of cancer registries. New groups of diseases are neurological diseases with a number of studies reporting on Parkinson's disease, multiple sclerosis, epilepsy, amyotrophic lateral sclerosis, myasthenia gravis, hemiplegic migraine and stroke, and rheumatological disorders, such as rheumatoid arthritis, systemic sclerosis, polyarteritis nodosa, dermatomyositis and systemic lupus erythematodes.

Apparently the characteristics of most of these chronic diseases, their patients and their registers fulfil criteria for feasibility as well as for agreement with the assumptions underlying capture-recapture analysis best. Perhaps with the exemption of some neurological and rheumatological conditions, the case-definition is likely unambiguous and uniform over the various registers. Arguably, for these categories of diseases sufficient registers are available and possible relationships between these registers, e.g. clinical registers, laboratory registers, health insurance registers or patient support and advocacy group registers, violating the independence assumption, are limited and perhaps avoidable by source selection. The permanent character of most of these conditions may reduce violation of the closed population assumption.

3.2 Application of capture-recapture analysis in studies on infectious diseases

An increasing number of capture-recapture studies have infectious diseases as their subject. Before 1997 studies using capture-recapture analyses in the context of infectious disease epidemiology were performed predominantly for HIV/AIDS.[3-12] Other infectious diseases studied were cryptococcosis,[13] measles,[14,15] meningococcal infection,[16,17] pertussis,[18] sexually transmitted diseases other than HIV/AIDS[19] and tetanus.[20] The majority of these studies were two-source capture-recapture studies. An overview of these studies can be found at http://www.pitt.edu/~yuc2/infec.html (accessed 1 May 2007). A synopsis of capture-recapture studies performed after 1997 is given in Table 3.1 Capture-recapture studies on tuberculosis have been discussed separately in section 1.2. The capture-recapture studies involved 19 different infectious diseases or groups of infectious diseases. The four infectious diseases mostly studied with capture-recapture analysis over the past ten years are HIV/AIDS, malaria, meningitis and pertussis. The capture-recapture studies on infectious diseases were performed all over the world but mostly in Europe or the USA. Again, the majority are two-source capture-recapture studies. Half were national studies and half were regional or local studies. The aims varied from estimating the completeness of registers, the completeness of ascertainment of linked registers, under-notification, the number of patients or disease incidence, the number of fatal cases or mortality rates, to estimating the number of outbreaks of an infectious disease. Registers used for capture-recapture estimates were notification, surveillance, hospital episode, laboratory, death and school registers. Sometimes one data source was created by performing a survey and incidentally self-reports were used as a second source. Reported completeness of registration or case-ascertainment could be as high as 99.9 % or as low as 9%.

Table 3.1 Published capture-recapture studies of infectious diseases, excluding tuberculosis, 1997-2006.

Disease	Authors	Objective	Method	Data-source	Outcome
HIV[a]/AIDS[b]	Bernillon *et al*[21]	To estimate the completeness of the French mandatory AIDS surveillance system, 1990-1993	Two-source CRC[c] model	1. Mandatory AIDS surveillance system 2. Hospital database on HIV infection	The completeness of the mandatory AIDS surveillance system was estimated at 83.6% (95%CI[d] 82.9-84.3%)
HIV/AIDS	Jara *et al*[22]	To estimate the completeness of the Massachusetts AIDS registry in 1994	Two-source CRC model	1. Mandatory AIDS surveillance system 2. Hospital discharge database or Medicaid database	The completeness of the mandatory AIDS surveillance system was estimated at 92.6% (95%CI 91.6-93.5%) using the hospital discharge database and at 94.5% (95%CI 93.7-95.3%) using the Medicaid database
HIV/AIDS	Acin *et al*[23]	To estimate the completeness of the prison AIDS register in Spain in 2000	Three-source log-linear CRC model	1. Prison register of AIDS patients 2. Prison register of TB[e] patients 3. Prison register of hospital admissions	The completeness of the prison register of AIDS patients was estimated at 55.3% (95%CI 51.7-58.0%)
HIV/AIDS	Pezotti *et al*[24]	To estimate the number of people infected with HIV in the Venetto region in Italy and the completeness of case-ascertainment, 1983- 2000	Four-source log-linear CRC model	1. National mandatory AIDS registry 2. Regional HIV registry 3. Regional death registry 4. Regional hospital discharge registry	The number of people infected with HIV was estimated at 11 281 (95%CI 10 981-11 621) and the completeness of case-ascertainment at 77.3% (95%CI 75.1-79.4%)
Dengue	Dechant *et al*[25]	To estimate the number of people hospitalised for suspected dengue in Puerto Rico, 1991-1995	Two-source CRC model	1. Laboratory–based surveillance system 2. Hospital–based surveillance system	In non-epidemic years an estimated average of 2791 (95%CI 1553-3481) suspected dengue patients were hospitalised and in an epidemic year 9479 (95%CI 9076-9882) persons. Inclusion of the hospital-based system improved case-ascertainment with 12.6%

Disease	Authors	Objective	Method	Data-source	Outcome
Giardiasis	Hoque *et al*[26]	To estimate the level of under-notification of giardiasis in the adult population of Auckland, New Zealand, 1998-1999	Two-source CRC model	1. Disease notification database 2. Giardiasis study database	The level of under-notification of giardiasis was estimated at 51%
Hepatitis B	Christensen *et al*[27]	To describe a hepatitis B outbreak among injecting drug users in Funen, Denmark, 1992-1998	Two-source CRC model	1. Hospital admissions 2. Notifications	The estimated number of hepatits B cases was 334 (95CI 283-385), resulting in an estimated under-notification of 38% (95%CI 33-45%)
Hepatitis C	Wu *et al*[28]	To estimate the mortality rate of hepatitis C in New York State excluding New York City, USA, in 1997	Two-source CRC model	1. Regional hospital discharge database 2. Regional multiple cause-of-death database	The number of people who had died of hepatitis C was estimated at 491 (95%CI 355-627) and the completeness of case-ascertainment at 55.3% (95%CI 43.2-76.3%)
Influenza	Grijalva *et al*[29]	To estimate influenza-associated hospitalisations in children in Davidson County, TN, USA, in 2003-2004	Two-source CRC model	1. Vaccine surveillance network 2. Emerging infections program	The overall sensitivity of the vaccine surveillance network and the emerging infections program for influenza hospitalisations was 73% and 38% respectively
Leprosy	Van den Broek *et al*[30]	To estimate the prevalence of disabled leprosy patients in four states in Northern Nigeria, 1997 - 1998	Two-source CRC model	1. Sample survey in clinics 2. Self-reported leprosy referral hospital admissions	The number of disabled leprosy patients was estimated at 1262 (95%CI 991-1533) and the completeness of case-ascertainment at 39.4% (95%CI 32.4-50.2%)

Disease	Authors	Objective	Method	Data-source	Outcome
Legionnaires' disease	Infuso *et al*[31]	To estimate the level of underreporting of Legionnaires' disease and evaluate the feasibility of a laboratory-based reporting system in France in 1995	Three-source log-linear CRC model	1. National notification system 2. Reference laboratory database 3. Hospital laboratory survey	The number of patients with Legionnaires' disease (culture-positive and/or four-fold increase antibody titre) was estimated at 256 and also including single antibody titres ≥ 256 IU at 528 (95%CI 509-547). Sensitivity of notification was estimated at 13% and 9% respectively
Legionnaires' disease	Nardone *et al*[32]	To evaluate improvements made to the mandatory notification system for Legionnaires' disease in France in 1998	Three-source log-linear CRC model	1. National notification system 2. Reference laboratory database 3. Hospital laboratory survey	The number of patients with Legionnaires' disease was estimated at 1124 (95%CI 973-1275) and sensitivity of notification had increased to 33% (95%CI 29-38%)
Leishmaniasis	Yadon *et al*[33]	To estimate the completeness of the cutaneous leishmaniasis surveillance system in four districts of the Santiago del Estero province, Argentine, 1990-1993	Two-source CRC model	1. National leishmaniasis surveillance system 2. Hospital records and case-control study records	The number of patients with leishmaniasis was estimated at 210 (95%CI 202-218) and completeness of reporting at 44.8% (95%CI 43.2-46.4%)
Leptospirosis	Brum *et al*[34]	To estimate the completeness of surveillance data on human leptospirosis in the health district of Santa Maria, Brazil, 2001-2002	Three-source log-linear CRC model	1. District surveillance database 2. Regional laboratory database 3. Regional hospital database	Record-linkage revealed over 20 times more cases than the official district estimate. Capture-recapture analysis revealed that an insignificant number of cases were missed after record-linkage

Disease	Authors	Objective	Method	Data-source	Outcome
Malaria	Deparis *et al*[35]	To estimate the incidence of malaria in the French Army in 1994	Two-source CRC model	1. Passive military disease surveillance system 2. Active military malaria surveillance system	The number of military servicemen with malaria was estimated at 854 (95%CI 803-905) and the completeness of case-ascertainment at 78.0% (95%CI 73.6-82.9%)
Malaria	Barat *et al*[36]	To determine the completeness of a passive malaria surveillance system in Arizona, California, New Mexico and Texas, USA, in 1995	Two-source CRC model	1. Passive malaria surveillance system 2. Active laboratory-based malaria detection survey	The number of malaria patients was estimated at 62 (95%CI 60-63) and the completeness of case-ascertainment at 98.4% (95%CI 95.3-100%)
Malaria	Wang *et al*[37]	To adjust the number of malaria cases from reporting data	Unknown	Unknown	Unknown
Malaria	Van Hest *et al (this thesis)*[38]	To estimate the completeness of notification of malaria by physicians and laboratories in the Netherlands in 1996	Three-source log-linear CRC model	1. Passive national notification register 2. Active laboratory survey 3. National hospital admission registration	The number of malaria patients was estimated at 774 (95%CI 740-821) and the completeness of notifications by physicians, laboratory reports and case-ascertainment at 40.2%, 69.1% and 86.2% respectively
Malaria	Klein *et al*[39]	To estimate the completeness of malaria notification in the Netherlands, 1999-2003	Two-source CRC model	1. National notification registration 2. National hospital admission registration	The completeness was until 1999 estimated at 35.5% by physicians and after 1999 at 36.1% by laboratories
Measles	Van den Hof *et al*[40]	To estimate the number of hospitalised measles patients in the Netherlands, 1999-2000	Two-source CRC model	1. Hospital discharge code statistics 2. Hospital chart review through national measles notification register	The number of hospitalised measles patients was estimated at 157 (95%CI 145-179), resulting in a completeness of notification of hospitalised measles patients of 47%

Disease	Authors	Objective	Method	Data-source	Outcome
Measles	Gindler *et al*[1]	To estimate the number of deaths due to measles and the efficiency of two reporting systems in the USA, 1987-2002	Two-source CRC model	1. National centre for health statistics (death certificates) 2. National measles surveillance system	The number of measles deaths was estimated at 259 (95%CI 244-274), resulting in a completeness of reporting of 64% for the health statistics and of 71% for the surveillance system
Meningitis, bacterial	Faustini *et al*[2]	To estimate the incidence of bacterial meningitis and to assess the quality of the surveillance systems in the Lazio Region, Italy, 1995-1996	Three-source log-linear CRC model	1. Mandatory notifiable disease surveillance system 2. Voluntary hospital laboratory-based surveillance system 3. Hospital discharge code registry	The number of bacterial meningitis cases was estimated at 236 (95%CI 206-306), resulting in a completeness of reporting for the notification, laboratory and hospitals discharge systems of 57%, 77% and 40% respectively
Meningitis, meningococcal	Izquierdo Carreno *et al*[3]	To evaluate the exhaustiveness of three information sources on meningococcal disease in Tenerife, Spain, 1999-2001	Three-source log-linear CRC model	1. Mandatory notifiable disease surveillance system 2. Laboratory survey 3. Hospital information registry	The sensitivity of the notification system was 84.9% after record-linkage and after CRC analysis case-ascertainment was estimated at 98.1%
Meningitis, meningococcal	Breen *et al*[4]	To assess the completeness of meningococcal disease notification in South Cheshire Health Authority, UK, 1999-2001	Three-source log-linear CRC model	1. Notification database 2. Laboratory reports database 3. Hospital discharge codes database	The completeness of meningococcal disease notification was estimated at 94.8% (95%CI 93.2-96.2%)
Meningitis, meningococcal	Berghold *et al*[5]	To assess the completeness of national surveillance data on invasive meningococcal disease in Austria in 2002	Two-source CRC model	1. National reference centre for meningococci 2. Hospital episode database	The completeness of notifications at the national reference centre for meningococci was estimated at 87.4%

Disease	Authors	Objective	Method	Data-source	Outcome
Meningitis, meningococcal	De Greeff *et al*[46]	To estimate the completeness of three data sources for meningococcal disease, after correction for false-positive diagnoses, in the Netherlands, 1993-1999	Three-source log-linear CRC model	1. Notification register 2. Hospital episode statistics 3. Reference laboratory for bacterial meningitis records	The completeness of the notification, hospital and laboratory registers was estimated at 49%, 67% and 58% respectively before correction of false-positive diagnoses, and at 52%, 70% and 62% respectively afterwards
Meningitis, pneumococcal	Gjini *et al*[47]	To improve estimates of disease incidence and deaths from pneumococcal meningitis among adults in England, 1996-1999	Two-source CRC model (2x)	1. Hospital episode statistics (HES) 2. Public health laboratory reports (PHLS) 3. Office for national statistics (ONS)	The sensitivity for incidence of HES and PHLS was estimated at 46% (95%CI 42-50%) and 40% (95%CI 37-44%) respectively. The sensitivity for mortality of HES and ONS was estimated at 48% (95%CI 41-55%) and 49% (95%CI 42-56%) respectively
Pertussis	Devine *et al*[48]	To estimate completeness of notification of whooping cough in the north west of England, 1994-1996	Three-source log-linear CRC model	1. Notification data from office for national statistics 2. Hospital admission data 3. Public health laboratory reports	The total number of cases was estimated at 2420, resulting in case-ascertainment after record-linkage of 51.2% and completeness of notification of 37.7%
Pertussis	Crowcroft *et al*[49]	To improve estimates of pertussis deaths in England and to identify reasons for under-ascertainment, 1994-1999	Three-source log-linear CRC model	1. Hospital episode statistics 2. Enhanced laboratory pertussis surveillance 3. Office for national statistics	The total number of pertussis deaths was estimated at 46 (95%CI 37-71). Case-ascertainment after record-linkage was estimated at 72% (95%CI 46-89%)
Pertussis	Vitek *et al*[50]	To estimate the total number of pertussis deaths in the USA, 1990-1999	Two-source CRC model	1. National notifiable disease surveillance system 2. Mortality records from national centre for health statistics	The total number of pertussis deaths was estimated at 159 (95%CI 145-181). Case-ascertainment after record-linkage was estimated at 83% (95%CI 73-91%)

Disease	Authors	Objective	Method	Data-source	Outcome
Rocky Mountain spotted fever (RMSF)	Paddock *et al*[51]	To estimate the magnitude of fatal RMSF and to evaluate the completeness of reporting of recognised RMSF deaths in the USA, 1983-1998	Two-source CRC model	1. Reported cases of fatal RMSF 2. Mortality records from national centre for health statistics	An estimated 612 (95%CI 548-675) cases of fatal RMSF were diagnosed between 1983-1998. The completeness of reporting of fatal RMSF and registration on death certificates was estimated at 36.6% and 49.7% respectively
STD's[f]	Reintjes *et al*[52]	To estimate the sensitivity of case-finding of syphilis and gonorrhoea (GO) in two national STD surveillance systems in the Netherlands in 1995	Two-source CRC model	1. STD register at Municipal Health Services (MHS) 2. Statutory notification by physicians (NNS)	The estimated sensitivities for the MHS were 31% (95%CI 27-35%) and 15% (95%CI 14-18%) for syphilis and GO respectively against 64% (95%CI 56-71%) and 22% (95%CI 19-25%) for the NNS. The estimated sensitivities after record-linkage were 76% and 34% for syphilis and GO respectively
Streptococcal disease	Trijbels-Smeulders *et al*[53]	To estimate the incidence rate of neonatal group B streptococcal disease and completeness of an active paediatric surveillance system in the Netherlands, 1997-1998	Two-source CRC model	1. Active paediatric surveillance system 2. Survey among readers of a national parent's magazine	The neonatal group B streptococcal disease incidence rate was estimated at 1.9 (95%CI 1.0-2.7) per 1000 live births. The completeness for neonatal group B streptococcal disease of the active paediatric surveillance system was estimated at 47% (95%CI 33-90%)
Salmonella infection	Gallay *et al*[54]	To assess the number of food-borne Salmonella outbreaks in France in 1995	Three-source log-linear CRC model	1. Public health notification system 2. Veterinary notification system 3. National Salmonella and Shigella reference centre	The estimated number of food-borne Salmonella outbreaks was 757, resulting in an estimated "case"-ascertainment of 94.6% after record-linkage

Disease	Authors	Objective	Method	Data-source	Outcome
Salmonella infection	Perez-Ciordia *et al*[55]	To estimate the incidence of Salmonella infection in Huesca, Spain, 1996-1999	Two-source CRC model	1. National diseases surveillance system 2. Microbiological hospital laboratory register	The estimated number of Salmonella infections was 1145, resulting in an estimated case-ascertainment of 83.8% after record-linkage. The estimated completeness for the laboratory register was 68% against 49% for the national disease surveillance system
Varicella	Goldman *et al*[56]	To estimate the completeness of the Varicella reporting data in Antelope Valley, California, USA, in 1995	Two-source CRC model	1. Varicella reporting from healthcare providers 2. Varicella reporting from schools	The estimated number of Varicella cases was 4498, resulting in an estimated case-ascertainment of 46% after record-linkage
Varicella-zoster	Bonhoeffer *et al*[57]	To determine the epidemiology of severe Vricella-zoster virus infections in hospitalised paediatric patients in Switzerland, 2000-2003	Two-source CRC model	1. Paediatric surveillance reports 2. Hospital episode database	The estimated completeness of the paediatric surveillance reports and the hospital episode database was 56% and 79% respectively
Various communicable diases	Jansson *et al*[58]	To assess the sensitivity of the Swedish surveillance system for Salmonellosis, meningococcal infection, tularaemia and penicillin-resistant pneumococci disease, 1998-2002	Two-source CRC model	1. Clinical notifications 2. Laboratory notifications	The estimated sensitivity of the Swedish surveillance system for notifiable communicable diseases was 99.9% for Salmonellosis, 98.7% for meningococcal infection, 98.5% for tularaemia and 93.4% for penicillin-resistant pneumococci. disease

a: HIV = human immunodeficiency virus; b: AIDS = Acquired Immunodeficiency Syndrome; c: CRC = capture-recapture; d: CI = Confidence Interval; e: TB= tuberculosis; f: STD = sexually transmitted disease

3.3 References

1. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-64.
2. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
3. Hardy AM, Starcher ET, Morgan WM, Druker J, Kristal A, Day JM, Kelly C, Ewing E, Curran JW. Review of death certificates to assess completeness of AIDS case reporting. *Public Health Rep* 1987; 102: 386-91.
4. Drucker E, Vermund SH. Estimating population prevalence of human immunodeficiency virus infection in urban areas with high rates of intravenous drug users: a model of the Bronx in 1988. *Am J Epidemiol* 1989; 130: 133-42.
5. Modesitt SK, Julman S, Fleming D. Evaluation of active versus passive AIDS surveillance in Oregon. *Am J Public Health* 1990; 80: 463-64.
6. Van Haastrecht HJ, Van den Hoek JA, Bardoux C, Leentvaar-Kuypers A, Coutinho RA. The course of the HIV epidemic among intravenous drug users in Amsterdam, The Netherlands. *Am J Public Health* 1991; 81: 59-62.
7. Frischer M, Green ST, Goldberg DJ, Haw S, Bloor M, McKeganey N, Covell R, Taylor A, Gruer LD, Kennedy D. Estimate of HIV-infection among injecting drug users in Glasgow, 1985-1990. *AIDS* 1992; 6: 1371-75.
8. McKeganey N, Barnard M, Leyland A, Coote I, Follet E. Female streetworking prostitutes and HIV infection in Glasgow. *BMJ* 1992; 305: 801-4.
9. Frischer M, Leyland A, Cormack R, Goldberg DJ, Bloor M, Green ST, Taylor A, Covell R, McKeganey N, Platt S. Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in Glasgow, Scotland. *Am J Epidemiol* 1993; 138: 170-81.
10. Hser YI. Population estimates of intravenous drug users and HIV infection in Los Angeles County. *Int J Addict* 1993; 28: 695-709.
11. Abeni DD, Brancato G, Perucci CA. Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection. *Epidemiology* 1994; 5: 410-4.
12. Mastro TD, Kitayaporn D, Weniger BG, Vanichseni S, Laosunthorn V, Uneklabh T, Uneklabh C, Choopanya K, Limpakarnjanarat K. Estimating the number of HIV-infected injection drug users in Bangkok: a capture-recapture method. *Am J Public Health* 1994; 84: 1094-9.
13. Dromer F, Mathoulin S, Dupont B, Laporte A. Epidemiology of cryptococcosis in France: a 9-year survey (1985-1993). *Clin Infect Dis* 1996; 23: 82-90.
14. Davis SF, Strebel PM, Atkinson WL, Markowitz LE, Sutter RW, Scanlon KS, Friedman S, Jadler SC. Reporting efficiency during a measles outbreak in New York City, 1991. *Am J Public Health* 1993; 83: 1011-15.
15. McGilchrist CA, McDonnell LF, Jorm LR, Patel MS. Loglinear models using capture-recapture methods to estimate the size of a measles epidemic. *J Clin Epidemiol* 1996; 49: 293-296.
16. Ackman DM, Birkhead G, Flynn M. Assessment of surveillance for meningococcal disease in New York State, 1991. *Am J Epidemiol* 1996; 144: 78-82.
17. Hubert B, Desenclos JC. [Evaluation of the exhaustiveness and representativeness of a surveillance system using the capture-recapture method: application to the surveillance of meningococcal infections in France in 1989 and 1990]. *Rev Epidemiol Sante Publique* 1993; 41: 241-49.
18. Sutter RW, Cochi SL. Pertussis hospitalizations and mortality in the United States, 1985-1988. *JAMA* 1992; 267: 386-91.
19. Rubin G, Umbach D, Shyu SF, Castillo-Chavez C. Using mark-recapture methodology to estimate the size of a population at risk for sexually transmitted diseases. *Stat Med* 1992; 11: 1533-49.
20. Sutter RW, Cochi SL, Brink EW, Sirotkin BI. Assessment of vital statistics and surveillance data for monitoring tetanus mortality, United States, 1979-1984. *Am J Epidemiol* 1990; 131: 132-42.
21. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D. Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990-1993. The Clinical Epidemiology Group from Centres d'Information et de Soins de l'Immunodeficience Humaine. *Int J Epidemiol* 2000; 29: 168-74.
22. Jara MM, Gallagher KM, Schieman S. Estimation of completeness of AIDS case reporting in Massachusetts. *Epidemiology* 2000; 11: 209-13.
23. Acin E, Gomez P, Hernando P, Corella I. Incidence of AIDS cases in Spanish penal facilities through the capture-recapture method, 2000. *Euro Surveill* 2003; 8: 176-81.
24. Pezzotti P, Piovesan C, Michieletto F, Zanella F, Rezza G, Gallo G. Estimating the cumulative number of human immunodeficiency virus diagnoses by cross-linking from four different sources. *Int J Epidemiol* 2003; 32: 778-83.

25. Dechant EJ, Rigau-Perez JG. Hospitalizations for suspected dengue in Puerto Rico, 1991-1995: estimation by capture-recapture methods. The Puerto Rico Association of Epidemiologists. *Am J Trop Med Hyg* 1999; 61: 574-8.
26. Hoque ME, Hope VT, Scragg R, Graham J. Under-notification of giardiasis in Auckland, New Zealand: a capture-recapture estimation. *Epidemiol Infect* 2005; 133: 71-9.
27. Christensen PB, Krarup HB, Niesters HG, Norder H, Schaffalitzky de Muckadell OB, Jeune B, Georgsen J. Outbreak of Hepatitis B among injecting drug users in Denmark. *J Clin Virol* 2001; 22: 133-41.
28. Wu C, Chang HG, McNutt LA, Smith PF. Estimating the mortality rate of hepatitis C using multiple data sources. *Epidemiol Infect* 2005; 133: 121-5.
29. Grijalva CG, Craig AS, Dupont WD, Bridges CB, Schrag SJ, Iwane MK, Schaffner W, Edwards KM, Griffin MR. Estimating influenza hospitalizations among children. *Emerg Infect Dis* 2006; 12: 103-9.
30. Van den Broek J, Van Jaarsveld T, de Rijk A, Samson K, Patrobas P. Capture-recapture method to assess the prevalence of disabled leprosy patients. *Lepr Rev* 2001; 72: 292-301.
31. Infuso A, Hubert B, Etienne J. Underreporting of Legionnaires‘ disease in France: the case for more active surveillance. *Euro Surveill* 1998; 3: 48-50.
32. Nardone A, Decludt B, Jarraud S, Etienne J, Hubert B, Infuso A, Gallay A, Desenclos JC. Repeat capture-recapture studies as part of the evaluation of the surveillance of Legionnaires' disease in France. *Epidemiol Infect* 2003; 131: 647-54.
33. Yadon ZE, Quigley MA, Davies CR, Rodrigues LC, Segura EL. Assessment of Leishmaniasis notification system in Santiago del Estero, Argentina, 1990-1993. *Am J Trop Med Hyg* 2001; 65: 27-30.
34. Brum L, Kupek E. Record linkage and capture-recapture estimates for underreporting of human leptospirosis in a Brazilian health district. *Braz J Infect Dis* 2005; 9: 515-20.
35. Deparis X, Pascal B, Baudon D. [Evaluation of the completeness of the epidemiological surveillance systems for malaria by the capture-recapture system in the French armies in l994]. *Trop Med Int Health* 1997; 2: 433-9.
36. Barat LM, Barnett BJ, Smolinski MS, Espey DK, Levy CE, Zucker JE. Evaluation of malaria surveillance using retrospective, laboratory-based active case detection in four southwestern states, 1995. *Am J Trop Med Hyg* 1999; 60: 910-4.
37. Wang WM, Gao Q, He HZ. [Application of capture-recapture method to adjust the number of malaria cases from reporting data]. *Zhongguo Ji Sheng Chong Xue Yu Ji Sheng Chong Bing Za Zhi* 2000; 18: 376.
38. Van Hest NA, Smit F, Verhave JP. Improving malaria notification in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-77.
39. Klein S, Bosman A. Completeness of malaria notification in the Netherlands 1995-2003 assessed by capture-recapture method. *Euro Surveill* 2005; 10: 244-6.
40. Van den Hof S, Smit C, Van Steenbergen JE, De Melker HE. Hospitalizations during a measles epidemic in the Netherlands, 1999 to 2000. *Pediatr Infect Dis J* 2002; 21: 1146-50.
41. Gindler J, Tinker S, Markowitz L, Atkinson W, Dales L, Papania MJ. Acute measles mortality in the United States, 1987-2002. *J Infect Dis* 2004; 189: S69-77.
42. Faustini A, Fano V, Sangalli M, Ferro S, Celesti L, Contegiacomo P, Renzini V, Perucci CA. Estimating incidence of bacterial meningitis with capture-recapture method, Lazio Region, Italy. *Eur J Epidemiol* 2000; 16: 843-8.
43. Izquierdo Carreno A, Matute Cruz P, Martinez Navarro F. [The use of the capture- recapture method in evaluating the epidemiological meningococcal disease monitoring system in Tenerife, Spain (1999-2000)]. *Rev Esp Salud Publica* 2003; 77: 701-11.
44. Breen E, Ghebrehewet S, Regan M, Thomson AP. How complete and accurate is meningococcal disease notification? *Commun Dis Public Health* 2004; 7: 334-8.
45. Berghold C, Berghold A, Fulop G, Heuberger S, Strauss R, Zenz W. Invasive meningococcal disease in Austria 2002: assessment of completeness of notification by comparison of two independent data sources. *Wien Klin Wochenschr* 2006; 118: 31-5.
46. De Greeff SC, Spanjaard L, Dankert J, Hoebe CJ, Nagelkerke N, De Melker HE. Underreporting of meningococcal disease incidence in the Netherlands: Results from a capture-recapture analysis based on three registration sources with correction for false-positive diagnoses. *Eur J Epidemiol* 2006; 21: 315-21.
47. Gjini A, Stuart JM, George RC,Nichols T, Heyderman RS. Capture-recapture analysis and pneumococcal meningitis estimates in England. *Emerg Infect Dis* 2004; 10: 87-93.
48. Devine MJ, Bellis MA, Tocque K, Syed Q. Whooping cough surveillance in the north west of England. *Commun Dis Public Health* 1998; 1: 121-5.
49. Crowcroft NS, Andrews N, Rooney C, Brisson M, Miller E. Deaths from pertussis are underestimated in England. *Arch Dis Child* 2002; 86: 336-38.
50. Vitek CR, Pascual FB, Baughman AL, Murphy TV. Increase in deaths from pertussis among young infants in the United States in the 1990s. *Pediatr Infect Dis J* 2003; 22: 628-34.
51. Paddock CD, Holman RC, Krebs JW, Childs JE. Assessing the magnitude of fatal Rocky Mountain spotted fever in the United States: comparison of two national data sources. *Am J Trop Med Hyg* 2002; 67: 349-54.

52. Reintjes R, Termorshuizen F, van de Laar MJ. Assessing the sensitivity of STD surveillance in the Netherlands: an application of the capture-recapture method. *Epidemiol Infect* 1999; 122: 97-102.
53. Trijbels-Smeulders M, Gerards LJ, M PC, de Jong P, van Lingen RA, Adriaanse AH, de Jonge GA, Kollee LA. Epidemiology of neonatal group B streptococcal disease in The Netherlands 1997-98. *Paediatr Perinat Epidemiol* 2002; 16: 334-41.
54. Gallay A, Vaillant V, Bouvet P, Grimont P, Desenclos JC. How many foodborne outbreaks of Salmonella infection occurred in France in 1995? Application of the capture-recapture method to three surveillance systems. *Am J Epidemiol* 2000; 152: 171-7.
55. Perez-Ciordia I, Ferrero M, Sanchez E, Abadias M, Martinez-Navarro F, Herrera D. [Salmonella enteritis in Huesca. 1996-1999]. *Enferm Infecc Microbiol Clin* 2002; 20: 16-21.
56. Goldman GS. Using capture-recapture methods to assess varicella incidence in a community under active surveillance. *Vaccine* 2003; 21: 4250-5.
57. Bonhoeffer J, Baer G, Muehleisen B, Aebi C, Nadal D, Schaad UB, Heininger U. Prospective surveillance of hospitalisations associated with varicella-zoster virus infections in children and adolescents. *Eur J Pediatr* 2005; 164: 366-70.
58. Jansson A, Arneborn M, Ekdahl K. Sensitivity of the Swedish statutory surveillance system for communicable diseases 1998-2002, assessed by the capture-recapture method. *Epidemiol Infect* 2005; 133: 401-7.

4

Underreporting of malaria incidence in the Netherlands: results from a capture-recapture study

N.A.H. VAN HEST[1,2], F. SMIT[3] and J.P. VERHAVE[4]

1 Department of Tuberculosis Control, Municipal Public Health Service Rotterdam Area, Rotterdam
2 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam
3 Monitoring and Epidemiology Unit, Trimbos Institute of Mental Health and Addiction, Utrecht
4 Department of Medical Microbiology, University Hospital Nijmegen, Nijmegen
the Netherlands

Epidemiol Infect 2002; 129: 371-7

Abstract

The aim of this study was to estimate the completeness of notification of malaria by physicians and laboratories in the Netherlands in 1996. We used a capture-recapture analysis of three incomplete, partially overlapping registers of malaria cases: a laboratory survey, the Notification Office and the hospital admission registration. The response of the laboratories was 83.2%. In 1996 the laboratories microscopically identified 535 cases of malaria, 330 patients with malaria were admitted to hospital and physicians notified 311 malaria cases. 667 malaria cases were recorded in at least one register. Capture-recapture analysis estimated the total number of malaria cases at 774 (95%CI 740-821). This implies a completeness of notification of 40.2% for physicians and 69.1% for the laboratories. It can be concluded that laboratory-based notification can considerably increase the number of officially reported malaria cases as compared to notification by physicians. However, possibly one-third of the cases may still go unreported.

Introduction

Malaria is one of the most frequently imported diseases in the Netherlands. The number of notified malaria cases increased over 25 years from 19 patients in 1972 to 311 in 1996. This increase was mainly the result of a rise in the number of reported cases of Plasmodium falciparum malaria, a potentially fatal disease. A similar trend has recently been described in 23 European countries, Australia, Canada, New Zealand and the United States.[1]

Under previous legislation regarding infectious diseases in the Netherlands, malaria was placed in category B. This group of infectious diseases had to be notified nominally (that is with the name and other particulars of the patient) within 24 h to the Municipal Health Service by the diagnosing physician. The Municipal Health Service forwarded this information to the Register of Notifiable Infectious Diseases (RNID) at the Office of the Chief Medical Officer where national data were aggregated for analysis, monitoring, public health intervention or policy making. Meaningful surveillance of imported malaria, such as trends in number of patients or type of plasmodium, identification of groups at risk (e.g. immigrants from malaria endemic countries or last-minute tourists with tropical destinations), evaluation of chemoprophylaxis advice, and implementation of adequate interventions, should preferably be based on data without bias due to incompleteness or underreporting. However, substantial underreporting of malaria was suspected.[2] After comparing hospital admission and notification data this was estimated at 59% over the years 1988-1992.[3] To reduce underreporting, laboratory-based notification was recommended because of the laboratory's crucial role in the diagnosis of malaria. On 1 April 1999 a new Contagious Diseases Act came into force in the Netherlands. Under this law malaria and nine other infectious diseases (brucellosis, yellow fever, leptospirosis, anthrax, ornithosis/psittacosis, Q fever, rubella, *E. coli*-infection and trichinosis) are placed into category C, which introduces mandatory notification by the head of the diagnosing laboratory instead of the physician.

The concept of underreporting (i.e. incomplete coverage) is often mentioned in the literature but seems to be based upon different definitions and correspondingly involves different calculations. Instead of quantifying underreporting in one register relative to other registers a more accurate picture is portrayed by assessing the completeness of the different registrations relative to an estimated total number of cases (i.e. the number of registered cases plus an approximated number of unobserved cases). The unobserved cases can be estimated with a statistical technique known as 'capture-recapture analysis'. Capture-recapture analysis has been used to assess the completeness of registration of various infectious diseases,[4-14] including malaria.[15,16] We have performed a capture-recapture analysis using three malaria registrations and estimated the completeness of notification by physicians and laboratories, followed by separate analyses for each type of plasmodium, because of a special interest in the underreporting of the most severe form, falciparum malaria.

Methods

Nearly all Dutch laboratories involved in parasitology participate in the national quality assessment for parasitological diagnosis. In January 1996 these laboratories (n = 107) were asked to report to us all microscopically confirmed cases of malaria found in that year through standardised questionnaires, with specific identifiers for patient (date of birth, sex and postal code) and parasite (*Plasmodium* species). Checks were carried out to exclude the possibility that a number of malaria cases would be diagnosed outside the laboratories in the survey, but notified to the RNID. Information from the Centres for Asylum-seekers, the Central Military Hospital, the Medical Service for Merchant Sailors and the Occupational Health Service of Amsterdam Airport and KLM Royal Dutch Airlines assured us that all these institutions perform malaria diagnosis through the regular laboratories.

Using the individual identifiers, the laboratory survey data were matched to two other national registers for malaria in the Netherlands: the RNID and the hospital admission data (ICD-9 code for malaria) from the National Morbidity Registration, after elimination of duplicate reports within each of the registrations. Two authors matched the data files by hand and in case of doubt consensus was sought.

The total number of individuals present in one or more registrations does not necessarily reflect a reliable approximation of the true number of cases. The purpose of capture-recapture analysis is to assess, on the basis of the available information, the number of cases that are not registered. In an article published in 1972, Stephen Fienberg demonstrated how this number of unobserved cases could be estimated, using log-linear analysis.[17] For capture-recapture analysis according to Fienberg the availability of data from at least three different, possibly incomplete, partially overlapping and preferably, but not necessarily, independent sources is needed.[6,17-20] The data can be put in a $2 \times 2 \times 2$ contingency table, indicating the absence or presence of a case in each of the registers. This table has one empty cell, corresponding to the number of cases never registered. Capture-recapture analysis aims at obtaining an estimate of the unregistered number of patients in the empty cell from the available data in the other cells. This estimate can be found under the best fitting and most parsimonious log-linear model. Finally, the total number of individuals is the number of registered cases plus the estimated number of non-registered patients.

Starting from a saturated model non-significant interaction terms were eliminated one after the other until the best fitting, most parsimonious, log-linear model was obtained by stepwise analysis as implemented in SPSS® (version 8.0. for Windows), with the procedure for hierarchical log-linear analysis. The coefficients and thus the final estimates were calculated with the SPSS® procedure for generalized log-linear analysis. The 95% confidence interval around the estimated number of malaria cases was calculated assuming a Poisson distribution. Four assumptions must be met for the three-sample capture-recapture model to be valid. We will return to these assumptions in the discussion. After estimating the total number of malaria cases we performed a stratified capture-recapture analysis by type of plasmodium to find out if the four malaria types have different capture-recapture probabilities. This was done to assess whether

underreporting occurred to a lesser or greater extent in relationship with the dangerous falciparum malaria or the more benign types.

Results

The response rate of the laboratories in the survey in 1996 was 83.2%. Some of the participating laboratories reported not performing microscopic diagnosis of malaria (4.7%) or did not identify any malaria case (5.6%), resulting in 72.9% of the laboratories reporting at least one case of malaria. In the laboratory survey *P. falciparum* accounted for 57.0% of the malaria cases. The distribution of the different malaria parasites in the laboratory survey is shown in Table 4.1 In the RNID 60% *P. falciparum* could be found against 69% among the hospital admissions. In the participating laboratories 535 cases of malaria were microscopically identified in 1996, while physicians officially notified 311 malaria cases and 330 malaria patients were admitted to a hospital. To increase the validity of the capture-recapture analysis, the matched data file was corrected for 12 cases notified to the RNID in 1997 but found to be diagnosed in the laboratories in 1996 and 15 cases notified in 1996 but actually diagnosed in the laboratories in 1995. After this correction for late notification a total of 667 malaria patients were known in at least one of the registers (Table 4.2). For two cases in the laboratory survey insufficient identifiers for perfect matching were available and these patients were excluded from the capture-recapture analysis. Figure 4.1 shows the distribution of the 665 malaria patients over the different malaria registrations and the overlap between these lists, as used in the capture-recapture analysis. A substantial number of malaria patients are only known to the RNID or the hospital admission register and do not appear in the laboratory survey.

Table 4.1 Distribution of diagnosed malaria parasites (*Plasmodium* species) and their percentage of the total number of malaria cases identified in the Netherlands in 1996.

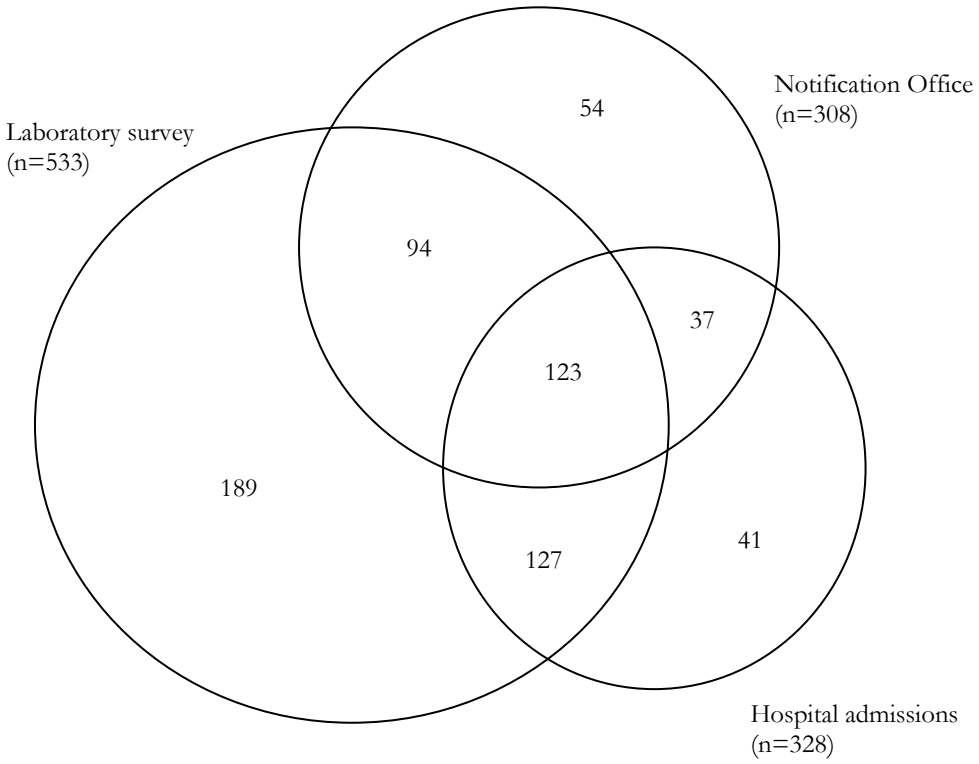
Plasmodium species	Malaria patients (%)	
P. falciparum	305	(57.0 %)
P. vivax	165	(30.8 %)
P. ovale	43	(8.0 %)
P. malariae	7	(1.3 %)
Parasite unknown	15	(2.8 %)

Table 4.2 The number of malaria patients identified in each of the three malaria registrations and the number of malaria patients registered in at least one of malaria registers in 1996.

Notified to the RNID*	311
Hospital admissions*	330
Diagnosed in the laboratories*	535
Registered in at least one of the registers* †	667

* After correction for duplicate reports
† After correction for late notification

Figure 4.1 Distribution of malaria cases in the Netherlands over three registers in 1996.



A log-linear model with two 2-way interactions, N*H, L*H (L = laboratory survey, N = Notification Office and H = hospital admissions) was obtained. These interactions represent pair-wise dependencies between the different registers [N and H] and [L and H]. The small likelihood ratio, G^2, compared to the number of degrees of freedom (df), shows that this model fitted the data well (G^2 = 0.741; df = 1; P = 0.785) and gave an estimate of 774 (95%CI 740 – 821) malaria cases. The completeness of notification in 1996 can now be estimated at 40.2% for physicians (311/774 cases) and 69.1% for the laboratories (535/774 cases).

The case-ascertainment (the number of malaria patients known in at least one of the three malaria registers) for 1996 can be estimated at 86.2% (95%CI 81.2-90.1%). The stratified capture-recapture analysis by type of plasmodium (Table 4.3) resulted in a slightly higher total number of estimated malaria cases of 788 patients (within 95%CI of original estimate). The detection rates of patients with different plasmodia do not show very much variation.

Discussion

This study confirms that more malaria cases occur in the Netherlands than are reflected by the numbers officially notified by physicians in the past. Furthermore it is demonstrated that the three different malaria registers are all substantially incomplete. Of particular interest is the observation that a considerable number of patients could only be found in the records of the RNID and/or the hospital admission register. They were unknown to the laboratories, although malaria diagnosis by thick or thin smear is often considered as the gold standard, especially at the time of this study when antigen tests were only used for research purposes. These cases could partly be explained by non and incomplete reporting of laboratories and cases could also have occurred in the (few) laboratories not participating in the national quality assessment of parasitological

Table 4.3 Stratified capture-recapture analysis of malaria cases in the Netherlands in 1996, according to the type of plasmodium. The detection rate is calculated as obs(N)/est(N).

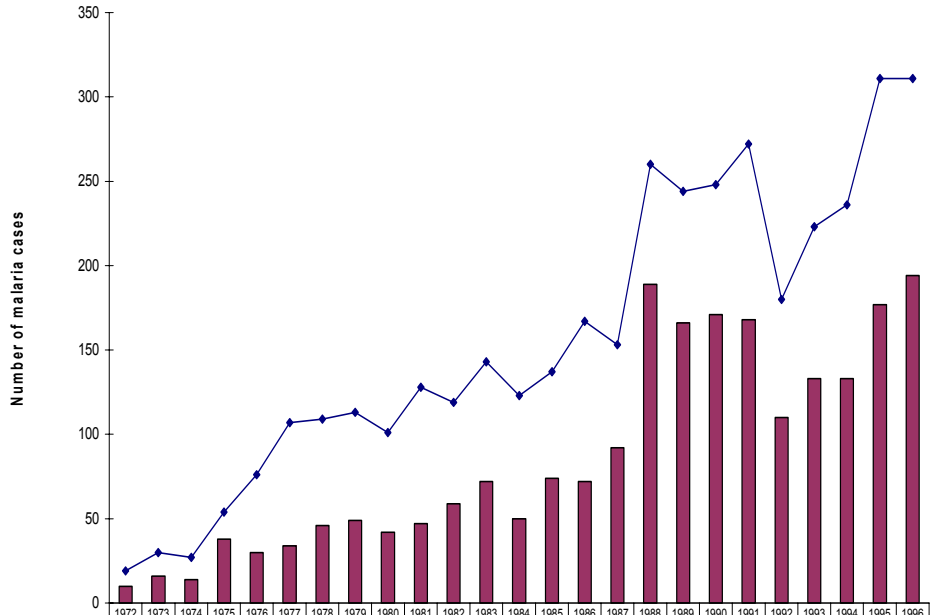
Plasmodium species	Cases observed*	Cases estimated	Detection rate
P. falciparum	383	438	0.87
P. vivax	195	222	0.88
P. ovale	50	56	0.89
P. malariae	8	8	1.00
Parasite unknown	29	64	0.45
Total malaria cases	665	788	0.84

* after exclusion of 2 laboratory cases with insufficient identifiers for perfect matching

diagnosis. Other patients might have been diagnosed abroad and started with anti-malarials before their arrival in the Netherlands, clearing the parasites from the peripheral blood and subsequently reported to us as negative by the Dutch laboratories. A number of patients may have been notified or admitted solely on clinical grounds, without laboratory verification. Although unlikely, some physicians (for example former tropical doctors) could still prepare and microscopically examine the blood films themselves.

The number of malaria notifications in the Netherlands showed an increasing trend until 1996 (Figure 4.2). According to Wetsteyn and De Geus,[21] the incidence of imported malaria is determined by the level of endemicity in the malarious areas visited, the exposure to infected *Anopheles* mosquitoes (in turn, related to duration of stay, way of travelling and practising anti-mosquito methods) and the success of chemoprophylaxis (determined by compliance and prophylactic drug resistance). The increase of imported malaria in the Netherlands in the second half of the 1970s could be explained by growing tourism to tropical Africa and a further rise during the 1980s is expected to be the result of the spread of resistance against chloroquine and other commonly used anti-malarial drugs. Apart from that, malaria transmission itself seems to have increased in certain

Figure 4.2 The total number of notified malaria cases and the number of malaria cases caused by *P. falciparum* in the Netherlands between 1972 and 1996.



	1972	1973	1974	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996
P.falciparum	10	16	14	38	30	34	46	49	42	47	59	72	50	74	72	92	189	166	171	168	110	133	133	177	194
Notified malaria	19	30	27	54	76	107	109	113	101	128	119	143	123	137	167	153	260	244	248	272	180	223	236	311	311

areas, such as West-Africa.[22] Participation in peace-keeping operations or elections,[23,24] the number and nationalities of immigrants and asylum-seekers[24-26] or the extent of certain marginalised groups[25,27] can also alter the incidence of imported malaria cases over a certain period of time, as well as influence the proportion of the different malaria parasites. In the Netherlands an increase could be observed in the proportion of malaria caused by *P. falciparum*. Around 1980 falciparum malaria was responsible for approximately 40% of all notified malaria cases but 10 years later this had increased to almost 70%.[21] In the 1990s the proportion of falciparum malaria stabilised around 60%.

Estimates of underreporting are frequently derived from different settings. They can be based upon surveys performed at the national level[28] or among small groups.[25,29] The background of the data that are compared with the official notification register may be different, and can vary between hospital admission data,[3] laboratory-based information,[25,28] physicians consulting a Reference Laboratory[29] or travellers.[30] The different registers were sometimes matched at the individual level,[25] at times in a stratified manner[3] or in another way.[28] In this study we used a well-described and replicable method and estimated the completeness of notification of three different malaria registers through capture-recapture analysis.

For the three-sample capture-recapture technique to be valid, four assumptions must hold.[17-20] First, overlap between registers must be established without erroneously misclassifying people as observed in only one, two or all three registers. This can be achieved when cases can be uniquely identified. We used individual identifiers for each of the patients and only two patients could not be identified beyond doubt due to (partially) missing identifiers. It is important that only true cases are counted. Ideally both the positive predictive value and the negative predictive value of the registrations should be 100%. None of our registrations will meet this condition, although in the case of malaria specifically, we assume that the positive predictive value will be high. The large overlap of the hospital records and the notification data with one or two of the other registrations also supports this view. When the positive predictive value of registrations is low capture-recapture analysis will result in overestimating the number of cases.

Second, the registers should preferably, but not necessarily, be 'independent', meaning that the probability of being recorded in one register is not affected by being (or not being) registered in another. Such dependence can result from co-operation between the agencies that keep the different registrations, exchange of information or a more or less predictable flow of patients along various institutions due to referral. In two-sample capture-recapture methods this assumption is crucial and dependencies can cause under- or overestimation. In the three-sample capture-recapture approach pair-wise dependencies between registers can be handled analytically. In the log-linear model they can be identified as interactions in the model. Since we could not rule out pair-wise dependencies, we decided not to rely on the two-sample capture-recapture analysis but instead to use Fienberg's method.

Third, the population should be "homogeneous" meaning that the population under consideration should not be composed of segments that have markedly different capture-recapture probabilities. One way of handling the homogeneity assumption is to

stratify the population into more homogeneous strata and then to carry out capture-recapture analyses for each of the distinct subgroups. We performed a stratified capture-recapture analysis by type of plasmodium. This resulted in a slightly, but not significantly, higher total number of estimated malaria cases of 788 patients. The detection rates of patients with different plasmodia do not show considerable variation. This may indicate the absence of a violation of the homogeneity assumption. However, we cannot exclude the possible presence of other (but unmeasured) sources of heterogeneity.

Finally, the population should be 'closed' such that the true population size is neither affected by people entering the population (e.g. through in-migration of cases and disease onset) nor by people leaving the population (e.g. through out-migration, recovery or mortality). The closed population assumption should be given critical attention because the aim of this study was to obtain an estimate of the incidence of imported malaria cases and violation might have resulted in overestimation (because incident cases may be late entries who have, therefore, a smaller probability of being captured more than once). When an open population is assumed, this could be handled in two different ways. One method is to perform the analysis of the different registrations within a short period of time and therefore the population could be considered as 'closed' during this interval. For imported malaria, a relatively rare disease with a short course, this approach does not seem feasible. An alternative is to use more complicated models, allowing for migration, birth and death to take place, such as the Jolly-Seber model.[31] The design of capture-recapture studies, the data requirements, the validity of the outcome of the different analyses and the selection of the most appropriate model to estimate the incidence of imported malaria and other infectious diseases should be given thought in further studies. In the context of these considerations our results suggest that laboratory-based notification can considerably increase the number of reported malaria cases as compared to notification by physicians. Since we actively approached the laboratories their level of underreporting found in this study cannot necessarily be extrapolated to the level of underreporting for laboratory notification. However, malaria was notified 571 times in 2001. Assuming a similar number of cases of imported malaria, this figure lies well within the range of our laboratory results for 1996. But possibly one-third of the malaria cases may still go unreported.

Acknowledgements

This study was performed in co-operation with the Office of the Chief Medical Officer in the Netherlands, the Netherlands Society of Parasitology (NVP), the Foundation for Parasitology Laboratory Diagnosis (SPLD), the Foundation for Quality Assessment in Medical Microbiology (SKMM) and the Centre for Communicable Disease Epidemiology of the National Institute of Public Health and the Environmental (RIVM). Permission was obtained from the Inspector-General for Infectious Diseases at the Office of the Chief Medical Officer and the Privacy Commission of the National Morbidity Registration (LMR). We thank the heads of the participating laboratories for their co-operation.

References

1. Muentener P, Schlagenhauf P, Steffen R. Imported malaria (1985-95): trends and perspectives. *Bull World Health Organ* 1999; 77: 560-6.
2. Sprenger MJW, Schrijnemakers PM. [More information on infectious diseases provided by a national information system]. *Ned Tijdschr Geneesk* 1998; 142: 1923-6.
3. Reep-van den Bergh CMM, Docters van Leeuwen WM, Kessel RPM van, Lelijveld JLM. [Malaria: under-notification and risk assessment for travellers to the tropics]. *Ned Tijdschr Geneesk* 1996; 140: 878-82.
4. Hubert B, Desenclos JC. [Evaluation of the exhaustiveness and representativeness of a surveillance system using the capture-recapture method. Application to the surveillance of meningococcal infections in France in 1989 and 1990]. *Rev Epidemiol Sante Publique* 1993; 41: 241-9.
5. Abeni DD, Brancato G, Perucci CA. Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection. *Epidemiology* 1994; 5: 410-4.
6. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. II: Applications in human diseases. *Am J of Epidemiol* 1995; 142: 1047-68.
7. Dromer F, Mathoulin S, Dupont B, Laporte A. Epidemiology of cryptococcosis in France: a 9-year survey (1985-1993). French Cryptococcosis Study Group. *Clin Infect Dis* 1996; 23: 82-90.
8. Infuso A, Hubert B, Etienne J. Underreporting of Legionnaires' disease in France: the case for more active surveillance. *Eurosurveillance* 1998; 3: 48-50.
9. Devine MJ, Bellis MA, Tocque K, Syed Q. Whooping cough surveillance in the north west of England. *Commun Dis Public Health* 1998; 1: 121-5.
10. Reintjes R, Termorshuizen F, van de Laar MJ. Assessing the sensitivity of STD surveillance in the Netherlands: an application of the capture-recapture method. *Epidemiol Infect* 1999; 122: 97-102.
11. Dechant EJ, Rigau-Perez JG. Hospitalizations for suspected dengue in Puerto Rico, 1991-1995: estimation by capture-recapture methods. The Puerto Rico Association of Epidemiologists. *Am J Trop Med Hyg* 1999; 61: 574-8.
12. Bernillon P, Lievre L, Pillonel J, Laporte A, Costagliola D. Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990-1993. The Clinical Epidemiology Group from Centres d'Information et de Soins de l'Immunodeficience Humaine. *Int J Epidemiol* 2000; 29: 168-74
13. Gallay A, Vaillant V, Bouvet P, Grimont P, Desenclos JC. How many foodborne outbreaks of Salmonella infection occurred in France in 1995? Application of the capture-recapture method to three surveillance systems. Am *J Epidemiol* 2000; 152: 171-7.
14. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4:141-3.
15. Deparis X, Pascal B, Baudon D. [Evaluation of the completeness of the epidemiological surveillance systems for malaria by the capture-recapture system in the French armies in l994]. *Trop Med Int Health* 1997; 2: 433-9.
16. Barat LM, Barnett BJ, Smolinski MS et al. Evaluation of malaria surveillance using retrospective, laboratory-based active case detection in four southwestern states, 1995. *Am J Trop Med Hyg* 1999; 60: 910-4.
17. Fienberg SE. The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
18. Bishop YMM, Fienberg SE, Holland PW. *Discrete multivariate analysis.* Cambridge: MIT-Press, 1975.
19. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-63.
20. LaPorte RE, Dearwater SR, Chang YF, Songer TJ, Aaron DJ, Anderson RL, Olsen T. Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring. *Am J Epidemiol* 1995; 142: 1069-77.
21. Wetsteyn JCFM, De Geus A. Falciparum malaria, imported into the Netherlands, 1979-1988. I. Epidemiological aspects. *Trop Geogr Med* 1995; 47: 53-60.
22. Philips-Howard PA, Porter J, Behrens RH, Bradley DJ. Epidemic alert: malaria infections in travellers from West Africa. *Lancet* 1990; 335: 119-120.
23. Kachur SP, Reller ME, Barber AM, Barat LM, Koumans EH, Parise ME, Robert J, Ruebush TK, Zucker JR. Malaria surveillance-United States, 1994. *MMWR CDC Surveill Summ* 1997; 46: 1-18.
24. Wetsteyn JCFM, Kager PA, Van Gool. The changing pattern of imported malaria in the Academic Medical Centre, Amsterdam. In: Wetsteyn JCFM. *Imported malaria in the Netherlands, an uninvited guest* (dissertation). Amsterdam: University of Amsterdam, 1993: pp. 117-133.

25. Lambeth, Southwark and Lewisham Health Authority. *The surveillance of communicable disease and non-infectious environmental hazards in Lambeth, Southwark and Lewisham.* London: Directorate of Health Policy and Public Health, 1996: 1996 Surveillance Report.
26. Bradley DJ, Warhurst DC. Malaria imported into the United Kingdom during 1991. *Commun Dise Rep CDR Rev* 1993; 3: R25-8.
27. Centers for Disease Control and Prevention. Summary of notifiable diseases United States 1996. *MMWR Morb Mortal Wkly Rep* 1997; 45: 1-87.
28. Legros F, Fromage M, Ancelle T, Burg E, Danis M. *Enquête nationale de recensement des cas de paludisme d'importation en France métropolitaine pour l'année 1997.* Paris: Centre National de Référence pour les Maladies d'Importation (CNMRI), 1998; Bulletin No 14.
29. Davidson RN, Scott JA, Behrens RH, Warhurst D. Underreporting of malaria, a notifiable disease in Britain. *J Infect* 1993; 26: 348-9.
30. Steffen R, Heuser R, Machler R, Bruppacher R, Naef U, Chen D, Hofman AM, Somaini B. Malaria chemoprophylaxis among European tourists in tropical Africa: use, adverse reactions, and efficacy. *Bull World Health Organ* 1990; 68: 313-22.
31. Cormack RM. Loglinear models for capture-recapture experiments in open populations. In: Hiorns RW, Cooke D, eds. *The mathematical theory of the dynamics of biological populations II*. London: Academic Press, 1981.

5

Incidence and completeness of notification of Legionnaires' disease in the Netherlands: covariate capture-recapture analysis acknowledging geographical differences

N.A.H. VAN HEST[1,2], C.J.P.A. HOEBE[3], J.W. DEN BOER[4], J.K. VERMUNT[5], E.P.F.IJZERMAN[6], W.G. BOERSMA[7] and J.H. RICHARDUS[1,2]

1 Department of Infectious Disease Control, Municipal Public Health Service Rotterdam-Rijnmond, Rotterdam
2 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam
3 Department of Infectious Diseases, South Limburg Public Health Service, Geleen
4 Department of Infectious Diseases and the Environment, Kennemerland Public Health Service, Haarlem
5 Department of Methodology and Statistics, Tilburg University, Tilburg
6 Regional Public Health Laboratory Kennermerland, Haarlem
7 Department of Pulmonary Diseases, Medical Centre Alkmaar, Alkmaar
the Netherlands

Epidemiol Infect (in press)

Abstract

To estimate incidence and completeness of notification of Legionnaires' disease in the Netherlands in 2000 and 2001, we performed a capture-recapture analysis using three registers: Notifications, Laboratory results and Hospital admissions. After record-linkage, of the 780 Legionnaires' disease patients identified 373 were notified. Ascertained under-notification was 52.2%. Because of expected and observed regional differences in the incidence rate of Legionnaires' disease, alternatively to conventional log-linear capture-recapture models, a covariate (region) capture-recapture model, not previously used for estimating infectious disease incidence, was specified and estimated 886 Legionnaires' disease patients (95% confidence interval (CI) 827-1022). Estimated under-notification was 57.9%. Notified, ascertained and estimated average annual incidence rates of Legionnaires' disease were 1.15, 2.42 and 2.77 per 100 000 inhabitants respectively, with the highest incidence in the southern region of the Netherlands. Covariate capture-recapture analysis acknowledging regional differences of Legionnaires' disease incidence appears to reduce bias in the estimated national incidence rate.

Introduction

Any surveillance system is concerned with the quality of the data collected, including the degree of ascertainment of affected individuals.[1] A conventional surveillance system is notification, possibly containing false-positive cases and often incomplete for true-positive cases, as described for Legionnaires' disease.[2,3]

Legionnaires' disease is a serious, possibly fatal, pneumonia caused by *Legionella* species, occurring in sporadic cases and outbreaks.[4,5] Under the present legislation regarding infectious diseases in the Netherlands, Legionnaires' disease is placed in category B. This group of infectious diseases has to be notified within 24 h to the Municipal Public Health Service by the diagnosing physician. The Municipal Public Health Service forwards this information to the Register of Notifiable Infectious Diseases at the Office of the Health Care Inspectorate where national data are aggregated for analysis, monitoring, public health intervention or policy making. Since 1999 on average 230 Legionnaires' disease patients were notified in the Netherlands annually. The average national annual incidence rate was 1.4 Legionnaires' disease patients per 100 000 inhabitants, almost three times higher than the average annual incidence rate in the United States and the United Kingdom.[6,7] However, the incidence rate based on notifications varies considerably per province.[8] Under-diagnosis and under-notification are likely. This can obscure the true burden of Legionnaires' disease, hamper the detection of clusters of Legionnaires' disease patients and hinder good investigations into the possible source of legionella infections. The Dutch Health Council estimated an annual number of 800 Legionnaires' disease patients. This number is based on the annual number of cases of pneumonia in the Netherlands (110 000) of whom 15% needs hospital admission (16 000) of which 5% is caused by *Legionella* species (800).[9]

Record-linkage is important for assessing the quality and completeness of infectious disease registers, i.e. comparing patient data across multiple registers.[10] Completeness of notification can be assessed by comparison with the case-ascertainment, i.e. the total number of patients observed in at least one register, or the estimated total number of patients through capture-recapture analysis. The total number of individuals present in one or more registrations does not necessarily reflect a reliable approximation of the true number of cases. The purpose of capture-recapture analysis is to assess the number of cases that are not registered. In an article published in 1972, Stephen Fienberg demonstrated how this number of unobserved cases could be estimated, using log-linear analysis.[11] For capture-recapture analysis, according to Fienberg, the availability of data from at least three different, possibly incomplete, partially overlapping and preferably, but not necessarily, independent sources is needed.[12-16] The data can be put in a $2 \times 2 \times 2$ contingency table, indicating the absence or presence of a case in each of the registers. This table has one empty cell, corresponding to the number of cases never registered. Based on certain assumptions, which will be discussed later, capture-recapture analysis aims at obtaining an estimate of the unregistered number of patients in the empty cell from the available data in the other cells. This estimate can be found under the best fitting and most parsimonious log-linear model, as explained later. Finally, the total number of individuals is the number of registered cases plus the estimated number of non-registered

patients. Capture-recapture methods have been used to estimate the total number of patients with Legionnaires' disease and other infectious diseases.[2,3,13]

The validity of capture-recapture analysis depends on possible violation of the underlying assumptions and one focus is to establish which method is most appropriate for specific datasets.[15] Usually, log-linear modelling of data from at least three linked registers is the preferred capture-recapture method because it can reduce bias due to interdependencies between two registers.[13,17] Stratified capture-recapture analysis according to categorical covariates associated with the probability of capture in a register can further reduce bias.[11,12,14,16] An alternative is to include these covariates, e.g. demographic, diagnostic or prognostic variables, in a log-linear covariate capture-recapture model but these models have rarely been used to estimate human disease incidence.[18,19]

This study aims to estimate incidence and completeness of notification of Legionnaires' disease in the Netherlands in 2000 and 2001 through record-linkage of three data sources and capture-recapture analysis.

Methods

Data sources and patient identifiers

Three Legionnaires' disease data sources were used:

1. *Notification.* Patients notified by their physician to the Health Care Inspectorate. A uniform questionnaire collected additional information from local Public Health Services processing the notifications.
2. *Laboratory.* Patients with a specified positive laboratory test result reported by the clinical microbiologists in a survey among all clinical microbiology laboratories after obtaining permission for this survey from the Dutch Society for Microbiology and supported by the Inspector-General for Infectious Diseases of the Health Care Inspectorate. Positive laboratory test results were classified as either confirmed (culture, urine antigen test or a fourfold rise in antibody titre [≥128 IU] against *Legionella* species in paired acute and convalescent serum samples) or probable (PCR, a high titre [≥256 IU] against *Legionella* species in one serum sample, or direct fluorescent antibody staining), according to the European Working Group for Legionella Infections (EWGLI) definitions. Patients with Legionnaires' disease only known to the Hospital register were classified as cases with unknown laboratory verification.
3. *Hospital.* Hospitalised patients recorded in the National Morbidity Registration by Prismant, covering all hospitals in the Netherlands with:
   a. an International Code for Diseases (ICD-9 code) for all forms of pneumonia (ICD-9 codes 480.0–487.0) for individuals known to Notification and/or Laboratory
   b. ICD-9 code 482.8 for individuals only known to Hospital.

ICD-9 has no specific code for Legionnaires' disease and, as reported from other countries, in the Netherlands ICD-9 code 482.8 (pneumonia due to other specified bacteria) is used for Legionnaires' disease patients.[20] Hospital records coded as 482.8 can therefore include

false-positive cases, mainly patients with *Escherichia coli* pneumonia, a rare nosocomial disease, predominantly occurring among intensive care patients. Data on the annual number of *E. coli* pneumonia patients in the Netherlands are not available. Based upon an estimated annual number of 60 000 intensive care admissions and an estimated *E. coli* pneumonia incidence of 1 per 1000 intensive care admissions (derived from a random survey among intensive care consultants in the Netherlands), the estimated annual number of *E. coli* pneumonia patients is 60. This number is used to correct the number of patients only known to Hospital. Because proxy code 482.8 is used, for cross-validation and collection of additional information, uniform questionnaires requested all chest physicians to report hospitalised Legionnaires' disease patients in 2000 and 2001.

For all patients in each register it was attempted to collect date of birth, postal code digits or town of residence, sex and date of notification (and first day of illness), first laboratory sample or hospital admission as personal identifiers to be used in all record-linkage procedures. Duplicate entries in each register were deleted.

Case-definition and study period

Legionnaires' disease patients are defined as all ascertained (notified, laboratory-reported or hospitalised) and un-ascertained Legionnaires' disease patients. Notified Legionnaires' disease patients with a first day of illness in 2000 and 2001 were included in the study. For inclusion of patients known to Laboratory and/or Hospital the laboratory sample date, hospital admission date or first known of both dates were used as proxy for first day of illness. Through examining the registers one month before and after the study period, all registers were corrected for late notification or laboratory results, as described previously.[17]

Record-linkage and stratification

Record-linkage was performed manually using the patient identifiers, proximity of dates and geographical information found in the three registers. In case of doubt consensus was sought between two investigators. Because of expected geographical differences in incidence of Legionnaires' disease, after record-linkage, on the basis of the provinces of the Netherlands, ascertained Legionnaires' disease patients were stratified into four regions: North (1 671 534 inhabitants), East (4 467 527 inhabitants), West (5 955 299 inhabitants) and South (3 892 715 inhabitants) (Figure 1) Correction for the estimated number of *E. coli* pneumonia patients in the different regions was proportional to the regional division of the total number of patients only ascertained in Hospital.

Coverage rates and capture-recapture analysis

The ascertained register-specific coverage rate is defined as the number of Legionnaires' disease patients in each register divided by the case-ascertainment, expressed as percentage. The total number of un-ascertained Legionnaires' disease patients was estimated on the basis of the distribution of the ascertained cases over the three registers. For internal validity analysis we used two-source capture-recapture analysis, as explained elsewhere.[21] Briefly, by two-source capture-recapture analysis the estimated total number of cases, N_{est}, equals the number of cases on register A, N_A, times the number of cases on register B, N_B, divided by the overlap of the two registers, N_{both} ($N_{est} = N_A \times N_B / N_{both}$, also known as the Petersen estimator equation). Approximately unbiased estimates of N_{est}

Figure 5.1 Four regions of the Netherlands

Region North: provinces Groningen, Friesland and Drenthe; Region East: provinces Gelderland, Utrecht, Overijssel and Flevoland; Region West: provinces North-Holland and South-Holland; Region South: provinces Zeeland, North-Brabant and Limburg

are expected when the registers are large. To correct for bias caused by small registers Chapman proposed the Nearly Unbiased Estimator, which can be expressed as N_{est} = $[(N_A + 1) \times (N_B + 1)/(N_{both} + 1)] - 1$.[13,22,23]

The independence of registers and other assumptions underlying capture-recapture analysis were described previously.[17] Specific interdependencies between the three registers, causing bias in two-source capture-recapture estimates, are probable. Using SPSS statistical software (version 13.0; SPSS Inc, Chicago, IL), conventional total and stratified three-source log-linear capture-recapture analysis was employed taking possible interdependencies and heterogeneity into account, as previously described.[17]

Alternatively to capture-recapture analysis stratified by region, a log-linear covariate capture-recapture model with one covariate, region, was specified.[18,19,24] Other covariates considered will be discussed later. The best-fitting models were identified using the likelihood ratio test (G^2). The null hypothesis in the likelihood-ratio goodness-of-fit test is that the specified model holds and the alternative is that it does not hold. If the null hypothesis does not need to be rejected (e.g. $P > 0.05$) this means that there is no evidence that the specified model is in disagreement with the data. The lower the value of G^2 the better is the fit of the model. In the log-linear estimation procedure model selection follows model fitting, i.e. to identify the models that are clearly wrong and select from a number of acceptable models the most appropriate. For model selection we used Akaike's Information Criterion (AIC) which can be expressed as AIC = G^2 – 2[degrees of freedom (df)].[25] The first term, G^2, is a measure of how well the model fits the data and the second term, 2[df], is a penalty for the addition of parameters (and hence model complexity). A second information criterion used was the Bayesian Information Criterion (BIC) which can be expressed as BIC = G^2 – [ln N_{obs}][df], where N_{obs} is the total number of observed individuals.[26] Relative to the AIC, the BIC penalises complex models more heavily. In general, in the log-linear capture-recapture estimation procedure the least complex, i.e. the least saturated (in other words the most parsimonious) model, whose fit appears adequate, is preferred.[13] Since the G^2 of the saturated model is zero and has no degrees of freedom left, the AIC and BIC are also zero and models with a negative AIC and BIC are preferred, although this does not necessarily mean that the estimate is correct. The estimated register-specific coverage rate is defined as the number of Legionnaires' disease patients in each register divided by the estimated total number of Legionnaires' disease patients, expressed as percentage.

Results

Notification system

In the notification register from the Health Care Inspectorate 358 Legionnaires' disease patients were recorded. An additional 15 patients were reported through the questionnaires from local Public Health Services processing the notifications, giving a total of 373 notified Legionnaires' disease patients.

Laboratory survey

Questionnaires were received from 36 out of the 48 laboratories (response rate 75%). Based on population estimates the co-operating laboratories served 81.2% of the Dutch population. A total of 261 patients with a positive test for *Legionella* species were reported. Of these patients 186 (71.3%) were notified. Additional information on laboratory diagnosis was available for another 127 patients through Public Health Service or chest physician questionnaires, bringing the total number of patients with known laboratory results to 388

Hospital records

From 385 chest physicians in the Netherlands 179 replies were received (response rate 46%), the majority indicating that the requested information could not be retrieved or no Legionnaires' disease patients were admitted. Chest physicians reported 44 Legionnaires' disease patients, all of them also known to Notification and/or Laboratory.

Out of 448 Legionnaires' disease patients in Notification and/or Laboratory, 331 (73.9%) could be linked to the National Morbidity Registration pneumonia records. Of the linked Legionnaires' disease patients 79 (23.9%) were classified as either 'pneumonia not specified' (ICD-9 code 486; 63 cases), 'pneumonia due to other specified organism' (ICD-9 code 483; 9 cases) or 'pneumococcal pneumonia' (ICD-9 code 481; 7 cases). The remaining 252 linked patients (76.1%) had ICD-9 code 482.8, the assigned code for Legionnaires' disease. Another 452 patients, unknown to Notification and/or Laboratory, were identified in Hospital with ICD-9 code 482.8. This number was adjusted to 332 Legionnaires' disease patients after deduction of an estimated number of 120 *E. coli* pneumonia patients in the two years studied, also recorded under ICD-9 code 482.8.

Epidemiological results

Table 5.1 shows the epidemiological characteristics of 447 Legionnaires' disease patients in Notification and/or Laboratory (one patient had insufficient data). The mean age was 54 years (standard deviation 14 years). The recorded case-fatality rate was 5.6%. The mean duration between onset of disease and microbiological diagnosis was 12 days (median 6 days). The mean duration of hospital admission was 19 days (median 13 days).

Table 5.2 shows the number and proportion per region of the different laboratory tests for *Legionella* species. There are differences between the four Dutch regions in laboratory diagnostic approach. In region North no culture results were reported. In region West a low proportion of fourfold rise in antibody titre and PCR results were reported and more patients had unknown test results, probably the result of non-participation of some larger laboratories. In region South a high proportion of fourfold rise in antibody titre and PCR results were reported, probably the result of a major reference laboratory in that region.

Case-ascertainment

Table 5.3 shows the distribution of the 780 ascertained Legionnaires' disease patients over the three registrations after record-linkage, in total and stratified by region The ascertained register-specific coverage rate of Notification, Laboratory and Hospital was 47.8% (373/780), 33.5% (261/780) and 85.0% (663/780) respectively. The ascertained under-notification was 52.2%. Table 5.4 shows the number of notified and ascertained Legionnaires' disease patients, the average annual incidence rate by notification and by case-ascertainment and the proportion of the ascertained patients notified, in total and stratified per region. The average national annual incidence rate by notification was 1.15/100 000 and by case-ascertainment 2.42/100 000. The regional annual incidence rates differ, with a 100% difference between the highest and lowest regional incidence rate based on notification, reducing to 50% difference after record-linkage. Based upon the notification data the low incidence rate in region North partly results from under-

notification but the notified and ascertained incidence rates in region South were higher than in the rest of the Netherlands ($P < 0.0001$).

Capture-recapture analysis

Internal validity analysis by two-source capture-recapture analysis on Notification and Hospital and on Laboratory and Hospital both estimate 865 Legionnaires' disease patients through Chapman's Nearly Unbiased Estimator. The considerable lower capture-recapture estimate obtained with Notification and Laboratory (523 Legionnaires' disease patients) indicates a larger positive association between this pair than between the other pairs, resulting in an estimate more biased downwards.

The best-fitting three-source log-linear capture-recapture model was the saturated model, i.e. the model including all two-variable associations and assuming absent three-way interaction, which yielded an estimate of 1253 Legionnaires' disease patients (95% confidence interval (CI) 1019-1715). Estimated under-notification was 70.2%. To acknowledge the geographical differences capture-recapture analysis stratified by region was performed. For all regions apart from region East a more parsimonious model, containing only one two-way interaction (between Notification and Laboratory), was selected as best-fitting model, with totals of 78, 327 and 277 Legionnaires' disease patients and incidence rates of 2.33, 2.75 and 3.56 per 100 000 inhabitants for region North, West and South respectively. For region East a saturated model was selected that estimated an unexpectedly high number of 650 Legionnaires' disease patients with a wide 95%CI of 283-2382 patients.

As an alternative to the stratified capture-recapture analysis we specified a log-linear covariate (region) capture-recapture model. The covariate model that served as a starting point contained, apart from the main effects for Region and the three registers, the Region-Notification, Region-Laboratory, Region-Hospital, Notification-Laboratory, Notification-Hospital, Laboratory-Hospital two-variable terms. In this model we allow for regional differences in the number of cases in the three registers, but not for interaction with other effects per stratum, as the association between the registers is assumed equal across regions. This model fits the data reasonably well ($G^2 = 22.1$; df = 9; $P = 0.009$) and estimates 932 Legionnaires' disease patients with a narrower CI of 851-1106, reducing statistical uncertainty. Inspection of the misfit for individual cells showed a large adjusted residual for Legionnaires' disease patients only known to Laboratory in region East. After including a separate parameter for this single cell we obtain a good fitting model ($G^2 = 5.7$; df = 8; $P = 0.686$). The estimated number of Legionnaires' disease patients was 886 (95%CI 827-1022), similar to the two internal validity estimates with least assumed interdependence.

The estimated register-specific coverage rate of Notification, Laboratory and Hospital was 42.1% (373/886), 29.5% (261/886) and 74.9% (663/886) respectively. The estimated under-notification was 57.9%. The estimated average annual incidence rate of Legionnaires' disease was 2.77/100 000.

Table 5.1 Epidemiological characteristics of 447 Legionnaires' disease patients*

	Male (*N* = 319)		Female (*N* = 128)		Total (*N* = 447)	
Age category						
0-19 years	0.3%	(1/318)	4.7%	(6/128)	1.6%	(7/446)
20-39 years	11.9%	(38/318)	18.0%	(23/128)	13.7%	(61/446)
40-59 years	55.3%	(176/318)	43.8%	(56/128)	52.0%	(232/446)
60-79 years	28.9%	(92/318)	30.5%	(39/128)	29.4%	(131/446)
> 80 years	3.5%	(11/318)	3.1%	(4/128)	3.4%	(15/446)
Seasonal pattern: month of disease onset						
Jan-Feb	7.8%	(25/319)	10.2%	(13/128)	8.5%	(38/447)
Mar-Apr	11.0%	(35/319)	10.9%	(14/128)	11.0%	(49/447)
May-Jun	19.4%	(62/319)	14.1%	(18/128)	17.9%	(80/447)
Jul-Aug	26.6%	(85/319)	25.0%	(32/128)	26.2%	(117/447)
Sep-Oct	21.9%	(70/319)	31.3%	(40/128)	24.6%	(110/447)
Nov-Dec	13.2%	(42/319)	8.6%	(11/128)	11.9%	(53/447)
Travel abroad during incubation period†						
Travel abroad: yes	53%	(169/319)	50%	(64/128)	52%	(233/447)
Countries involved:						
Turkey	20%	(33)	30%	(19)	22%	(52)
France	23%	(39)	8%	(5)	19%	(44)
Spain	12%	(21)	13%	(8)	12%	(29)
Italy	8%	(14)	11%	(7)	9%	(21)
Germany	7%	(12)	9%	(6)	8%	(18)
Portugal	2%	(4)	2%	(1)	2%	(5)
Greece	2%	(4)	2%	(1)	2%	(5)
Belgium	3%	(5)	0%		2%	(5)
Rest Europe	11%	(18)	13%	(8)	11%	(26)
America's	5%	(9)	6%	(4)	6%	(13)
Asia	3%	(5)	2%	(1)	3%	(6)
Africa	0%		3%	(2)	1%	(2)
Unknown	3%	(5)	3%	(2)	3%	(7)
Legionella species						
L. pneumophila serogroup 1	61.2%	(170/278)	54.5%	(60/110)	59.3%	(230/388)
L. pneumophila serogroups 2-12	2.5%	(7/278)	1.8%	(2/110)	2.7%	(9/388)
L. non-pneumophila	3.2%	(9/278)	0.9%	(1/110)	2.6%	(10/388)
unknown	31.7%	(88/278)	42.7%	(47/110)	34.8%	(135/388)
Laboratory confirmation‡						
At least two confirming tests	22.0%	(61/277)	17.3%	(19/110)	20.7%	(80/387)
One confirming test	56.0%	(155/277)	56.4%	(62/110)	56.1%	(217/387)
Only probable test	22.0%	(61/277)	26.4%	(29/110)	23.3%	(90/387)

* From 448 patients sufficient data was available for analysis; sometimes one or two variables are missing; † *Rest of Europe*: Austria, Croatia, Cyprus, England, Hungary, Ireland, Luxemburg, Moldavia, Poland, Slovakia, Switzerland, Czech Republic, Yugoslavia; *America's*: Netherlands Antilles, Brazil, Canada, Dominican Republic,

Mexico, Peru, USA, Venezuela; *Asia*: China, Indonesia, Japan, Kazakhstan, Malaysia. *Africa*: Morocco and Tunis; ‡ confirmed laboratory diagnosis: positive culture, positive urine antigen test or a fourfold rise in antibody titre against *Legionella* species in paired acute and convalescent serum samples, ≥128 IU; probable laboratory diagnosis: positive PCR, a high titre in one serum sample against *Legionella* species, ≥256 IU, or direct fluorescent antibody staining of the organism.

A sensitivity analysis, assuming double or half the number of false-positive cases due to *E. coli* pneumonia only known to Hospital, estimated the number of Legionnaires' disease patients to range between 727 (95%CI 689-813) and 966 (95%CI 896-1126).

Discussion

After record-linkage and log-linear covariate capture-recapture analysis of three registers of Legionnaires' disease in 2000 and 2001 in the Netherlands we found a notified, ascertained and estimated annual incidence rate of 1.15, 2.42 and 2.77 cases per 100 000 inhabitants respectively. Ascertained and estimated under-notification was 52.2% and 57.9% respectively. This indicates the need for more consistent notification, e.g. through treatment of Legionnaires' disease by a limited group of clinicians, familiar with notification. The southern part of the Netherlands had a higher notified, ascertained and estimated incidence rate of Legionnaires' disease.

Legionella pneumonia might be responsible for 0%-14% of all nosocomial pneumonia's and for 2%-16% of all community-acquired pneumonias.[27] In the Netherlands legionella pneumonia is reportedly responsible for 7% of all nosocomial pneumonias and 2%-8% of all community-acquired pneumonias in hospitalised patients.[28-30] Under-notification of Legionnaires' disease is estimated at 67% in France, 90% in England and 95% in the United States.[3,31-33] At 57.9% we estimated lower under-notification in the Netherlands, possibly influenced by increased awareness after a major outbreak or increased use of the urine antigen test (although this use is proportionally still low compared to the average EWGLI data for Europe).[4,31] Among patients in the laboratory survey with positive legionella results under-notification was 28.7%, much lower than reported in France.[2] Parallel to mandatory notification by clinicians, many Dutch laboratories report positive results voluntarily to the Public Health Services, which reduces under-notification of Legionnaires' disease and other infectious diseases. The ascertained and estimated register-specific coverage rates for the laboratories would be higher with a better response. Record-linkage improved completeness of information in the linked dataset but, unlike laboratories, clinicians are not a useful source of additional information.

Several assumptions must be met for valid results of three-source log-linear capture-recapture models and limitations of capture-recapture analysis are described.[13,16,34-39] Violation of the closed population assumption is assumed limited for Legionnaires' disease as opportunities for notification, laboratory-verification or hospitalisation are largely determined within a short period of time, but could result in overestimation of the number of patients. Due to lack of a unique patient identification number used in all registrations and incomplete information on personal identifiers in

Table 5.2 Number and proportion of the laboratory test results for *Legionella* species in the Netherlands in 2000 and 2001, in total and stratified per region

	Confirmed laboratory test						Probable laboratory test						Unknown*	
	Culture (%)		Urine antigen test (%)		Fourfold rise in antibody titre (%)		Positive PCR (%)		High single titre (%)		DFA			
All legionella pneumonia (100% of population)	71	(100%)	216	(100%)	92	(100%)	33	(100%)	119	(100%)	0		56/441	(13%)
Region North (11%)	0	(0%)	15	(7%)	14	(15%)	2	(6%)	8	(7%)	0		3/34	(9%)
Region East (18%)	16	(22%)	61	(28%)	23	(25%)	5	(15%)	33	(28%)	0		14/123	(11%)
Region West (41%)	31	(44%)	78	(36%)	17	(19%)	3	(9%)	30	(25%)	0		31/149	(21%)
Region South (30%)	24	(34%)	62	(29%)	38	(41%)	23	(70%)	48	(40%)	0		8/135	(6%)

* for 441 patients information of Region was known

Table 5.3 Ascertained total number of Legionnaires' disease patients and number stratified by region of the Netherlands in three linked Legionnaires' disease registrations in 2000 and 2001, after proportional adjustment for false-positive Escherichia coli pneumonia patients only known to the Hospital register.

	*N*ascertained	Only NOT*	Only LAB†	Only HOSP‡	NOT and LAB	NOT and HOSP	LAB and HOSP	NOT and LAB and HOSP
All Legionnaires' disease patients	780	56	30	332	31	131	45	155
Region North§	69	3	2	35	2	6	8	13
Region East§	185	13	13	62	3	42	7	45
Region West§	286	23	5	136	7	55	14	46
Region South§	234	13	9	99	19	28	15	51

* NOT: Notification register (373 patients); † LAB: Laboratory register (261 patients); ‡ HOSP: Hospital admission register. The proportional correction for the Escherichia coli pneumonia patients in the regions North, East, West and South is 13, 22, 49 and 36 patients respectively (663 patients); § for 6 Legionnaires' disease patients the place of residence unknown

Table 5.4 Number of notified and ascertained Legionnaires' disease patients, the average annual Legionnaires' disease incidence (*N*/100 000) and the proportion of the ascertained Legionnaires' disease patients notified, in the Netherlands and stratified per region

	Notification (passive surveillance)		Record-linkage (case-ascertainment)		
	Number of notified Legionnaires' disease patients*	Average annual incidence (*N*/100 000)	Number of ascertained Legionnaires' disease patients	Average annual incidence (*N*/100 000)	Proportion notified
All Legionnaires' disease patients (1 598 7075 inhab†)	373	1.15	780	2.42	47.8%
Region North (1 671 534 inhab)	24	0.72	69	2.06	34.8%
Region East (4 467 527 inhab)	103	1.15	185	2.07	55.7%
Region West (5 955 299 inhab)	131	1.10	286	2.40	46.0%
Region South (3 892 715 inhab)	111	1.43	234	3.01	47.4%

* the information on region was missing for 4 Legionnaires' disease patients; † inhab: inhabitants

some records, imperfect record-linkage cannot be excluded but balanced misclassification can still result in unbiased numbers in each category. Limitations of capture-recapture studies due to lack of a uniform and unambiguous case-definition and variable specificity of registers are described by others.[36,40] The notification criteria in the Netherlands requires a clinical diagnosis of pneumonia and a confirmed or probable laboratory diagnosis. However, for 187 notified patients (50.1%) and 463 hospitalised patients (69.8%) no laboratory-verification was found, although part of these patients could be microbiologically diagnosed in a non-participating laboratory or abroad or, due to imperfect record-linkage, could not be linked to Laboratory. Likewise Laboratory may contain cases without pneumonia and cases diagnosed on a single high antibody titre, a test with a low positive predictive value.[3,29] The 79 linked patients in Hospital with another pneumonia ICD-9 code than 482.8 are likely miscoded but some could be false-positive cases. Violation of the perfect positive predictive value of the hospital episode registers is always a reason for concern in capture-recapture studies on infectious diseases and should be addressed critically, even when specific disease codes are used, e.g. for tuberculosis in ICD-9.[41-44] We have corrected for imperfect positive predictive value of Hospital. Possible bias as a result of correction for other hospitalised patients with ICD-9 code 482.8 is reflected in the confidence intervals of the sensitivity-analysis. Conventional log-linear capture-recapture analysis for the Netherlands and region East selected the saturated model, with an unexpectedly high estimate in region East. When saturated capture-recapture models are selected by any criterion investigators should be particularly

cautious about the associated outcomes.[16,44-46] We selected the three-source covariate capture-recapture model with equal two-way interactions across the regions as the best-fitting model. Internal validity analysis and analyses stratified by region indicate dependence between Notification and Laboratory as the dominant interaction. Positive three-way interaction across sources, causing underestimation of the number of Legionnaires' disease patients, cannot be incorporated in the selected model but is arguably limited. Regional heterogeneity in probability of being captured in the different registers was expected and observed.[3,8] Covariate capture-recapture models have been used only rarely to estimate disease incidence but appear to reduce bias due to heterogeneity and result in plausible estimates of the total number of cases, e.g. in simulations.[18,19] Inclusion of other covariates than region in the model, such as age or method of laboratory diagnosis, could have further reduced bias. In France, apart from region, method of diagnosis was identified as a variable with heterogeneity of capture.[3] However, proportional correction for *E. coli* pneumonia patients in Hospital, as performed for the regional stratification, was not feasible. Bias due to exclusion of these and unobserved possibly relevant covariates from the model can not be excluded.

Different characteristics of diseases, the patients and their registers can introduce various degrees of register interdependence and population heterogeneity into capture-recapture analysis, influencing model preference. This study shows that in the Netherlands for Legionnaires' disease there is considerable interdependence between Notification and Laboratory and confirms geographical heterogeneity. Log-linear covariate capture-recapture analysis with region as covariate appears to reduce bias in the estimated number of Legionnaires' disease patients. To our knowledge this is the first covariate capture-recapture study performed for infectious disease surveillance. Further research is needed into the causes of the geographical differences of Legionnaires' disease incidence rates.

Acknowledgements

We thank Dr Carol Joseph for reviewing an earlier manuscript. Permission for this study was obtained from the medical ethics committee of the Erasmus MC, University Medical Centre Rotterdam, Rotterdam, the Netherlands, and the data protection committees of the Legionnaires' disease registrations.

References

1. Nanan DJ, White F. Capture-recapture: reconnaissance of a demographic technique in epidemiology. *Chronic Dis Can* 1997; 18: 144-8.
2. Infuso A, Hubert B, Etienne J. Underreporting of Legionnaires' disease in France: the case for more active surveillance. *Euro Surveill* 1998; 3: 48-50.
3. Nardone A, Decludt B, Jarraud S, Etienne J, Hubert B, Infuso A, Gallay A, Desenclos JC. Repeat capture-recapture studies as part of the evaluation of the surveillance of Legionnaires' disease in France. *Epidemiol Infect* 2003; 131: 647-54.
4. Den Boer JW, Yzerman EP, Schellekens J, Lettinga KD, Boshuizen HC, Van Steenbergen JE, Bosman A, Van Den Hof A, Van Vliet HA, Peeters MF, Van ketel RJ, Speelman P, Kool JL, Conyn-Van Spaendonck MA. A large outbreak of Legionnaires' disease at a flower show, the Netherlands, 1999. *Emerg Infect Dis*

2002; 8: 37-43.
5. Lettinga KD, Verbon A, Weverling GJ, Schellekens JF, Den Boer JW, Yzerman EP, Prins J, Boersma WG, van Ketel RJ, Prins JM, Speelman P. Legionnaires' disease at a Dutch flower show: prognostic factors and impact of therapy. *Emerg Infect Dis* 2002; 8: 1448-54.
6. Ricketts KD, Joseph CA. Legionnaires' disease in Europe 2003-2004. *Euro Surveill* 2005; 10: 256-9.
7. Centers for Disease Control and Prevention (CDC); Jajosky RA, Hall PA, Adams DA, Dawkins FJ, Sharp P, Anderson WJ, Aponte JJ, Jones GF, Nitschke DA, Worsham CA, Adekoya N, Doyle T. Summary of Notifiable Diseases -- United States, 2004. *MMWR Morb Mortal Wkly Rep* 2006; 53: 1-79.
8. Den Boer JW, Friesema IH, Hooi JD. [Reported cases of Legionnaires' disease in the Netherlands, 1987-2000]. *Ned Tijdschr Geneesk* 2002; 46: 315-20.
9. Health Council of the Netherlands. *Controlling Legionnaire's Disease.* The Hague: Health Council of the Netherlands, 2003; publication no. 2003/12.
10. Migliori GB, Spanevello A, Ballardini L, Neri M, Gambarini C, Moro ML, Trnka L, Raviglione MC. Validation of the surveillance system for new cases of tuberculosis in a province of northern Italy. Varese Tuberculosis Study Group. *Eur Respir J* 1995; 8: 1252-8.
11. Fienberg SE. The multiple-recapture census for closed populations and the 2^k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
12. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis.* Cambridge: MIT-Press, 1975.
13. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
14. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation II: Applications in human diseases. *Am J Epidemiol* 1995; 142: 1059-68.
15. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med* 2001; 20: 3123-57.
16. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-63.
17. Van Hest NA, Smit F, Verhave JP. Underreporting of malaria incidence in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-7.
18. Tilling K, Sterne JA. Capture-recapture models including covariate effects. *Am J Epidemiol* 1999; 149: 392-400.
19. Tilling K, Sterne JA, Wolfe CD. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol* 2001; 30: 1351-9.
20. Slobbe LC, De Bruin A, Westert GP, Kardaun JW, Verwij GC. [*Classification of diagnoses and procedures and application in new hospital episode statistics*]. *Bilthoven, the Netherlands* .National Institute of Public Health and the Environment (RIVM), 2004. RIVM report 260201002/2004: p. 73. (http://www.cbs.nl/NR/rdonlyres /E9DC7CF9-0BDF-40EA-A52D-EE6FBEE1B904/0/ rivmrapport260201002.pdf). Accessed 18 April 2007.
21. Hook EB, Regal RR. Internal validity analysis: a method for adjusting capture-recapture estimates of prevalence. *Am J Epidemiol* 1995; 142: S48-52.
22. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses. *U California Public Stat* 1951; 1: 131-160.
23. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture-recapture estimate. *Biometrics* 1972; 28: 592-597.
24. Hope VD, Hickman M, Tilling K. Capturing crack cocaine use: estimating the prevalence of crack cocaine use in London using capture-recapture with co-variates. *Addiction* 2005; 100: 1701-8.
25. Sakamoto Y, Ishiguru M, Kitigawa G. *Akaike information criterion statistics.* Tokio: KTK Scientific, 1986: pp 1-24.
26. Agresti A. *Categorical data analysis.* New York: John Wiley and Sons, 1990: p 251.
27. Kool JL *Preventing Legionnaires' disease.* [Thesis]. Amsterdam: University of Amsterdam, 2000.
28. Bohte R, Van Furth, R, Van den Broek, PJ. Aetiology of community-acquired pneumonia: a prospective study among adults requiring admission to hospital. *Thorax* 1995; 50: 543-7.
29. Braun JJ, de Graaff CS, de Goey, Zwinderman AH, Petit PL. [Community-acquired pneumonia: pathogens and course in patients admitted to a general hospital]. *Ned Tijdschr Geneesk* 2004; 148: 836-40.
30. Van der Eerden MM, Vlaspolder F, De Graaff CS, Groot T, Bronsveld W, Jansen HM, Boersma WG. Comparison between pathogen directed antibiotic treatment and empirical broad spectrum antibiotic treatment in patients with community-acquired pneumonia: a prospective randomised study. *Thorax* 2005; 60: 672-8
31. Joseph CA. Legionnaires' disease in Europe 2000-2002. *Epidemiol Infect* 2004; 132: 417-24.
32. Marston BJ, Lipman HB, Breiman RF. Surveillance for Legionnaires' disease. Risk factors for morbidity and mortality. *Arch Intern Med* 1994; 154: 2417-22.
33. Marston BJ, Plouffe JF, File TM, Hackman BA, Salstrom SJ, Lipman HB, Kolczak MS, Breiman RF.

Incidence of community-acquired pneumonia requiring hospitalization - Results of a population-based active surveillance study in Ohio. *Arch Intern Med* 1997; 157: 1709-18.
34. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *Int J Epidemiol* 1994; 23: 1322-3.
35. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *J Clin Epidemiol* 1999; 52: 909-14.
36. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1999; 25: 474-8.
37. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* 2000; 152: 771-9.
38. Jarvis SN, Lowe PL, Avery A, Levene S, Cormack RM. Children are not goldfish-mark/recapture techniques and their application to injury data. *Inj Prev* 2000; 6: 46-50.
39. Tilling K. Capture-recapture methods-useful or misleading? *Int J Epidemiol* 2001; 30: 12-4.
40. Borgdorff MW, Glynn JR, Vynnycky E. Using capture-recapture methods to study recent transmission of tuberculosis. *Int J Epidemiol* 2004; 33: 905-6.
41. Tocque K, Bellis MA, Beeching N, Davies PD. Capture recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-143.
42. Baussano I, Bugiani M, Gregori D, Van Hest R, Borraccino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Intl J Tuberc Lung Dis* 2006; 10: 415-421.
43. Van Hest NA, Smit F, Baars HW, Vries G de, Haas P de, Westenend PJ, Nagelkerke N, Richardus JH. Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiol Infect* 2006; Published on-line: 7 December 2006; doi:10.1017 /S0950268806007540.
44. De Greeff SC, Spanjaard L, Dankert J, Hoebe CJ, Nagelkerke N, De Melker HE. Underreporting of meningococcal disease incidence in the Netherlands: Results from a capture-recapture analysis based on three registration sources with correction for false-positive diagnoses. *Eur J Epidemiol* 2006; 21: 315-21.
45. Regal RR, Hook EB. Validity of methods for model selection, weighing for model uncertainty and small sample adjustments in capture-recapture estimation. *Am J Epidemiol* 1997; 145: 1138-44.
46. Cormack RM, Chang YF, Smith GS. Estimating deaths from industrial injury by capture-recapture: a cautionary tale. *Int J Epidemiol* 2000; 29: 1053-9.

6

Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis?

N.A.H. VAN HEST[1,2], F. SMIT[3,4], H.W.M. BAARS[1], G. DE VRIES[1,2], P.E.W. DE HAAS[5], P.J. WESTENEND[6], N.J.D. NAGELKERKE[7] and J.H. RICHARDUS[1,2]

1 Department of Infectious Control, Municipal Public Health Service Rotterdam-Rijnmond, Rotterdam
2 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam
3 Monitoring and Epidemiology Unit, Trimbos Institute of Mental Health and Addiction, Utrecht
4 Department of Clinical Psychology, Vrije Universiteit Amsterdam, Amsterdam
5 Diagnostic Laboratory of Infectious Diseases and Prenatal Screening, Mycobacteria Reference Unit, National Institute of Public Health and the Environment, Bilthoven
6 Laboratory for Pathology Dordrecht, Dordrecht
the Netherlands
7 Department of Community Medicine, United Arab Emirates University, Al Ain
United Arab Emirates

Epidemiol Infect 2006; Published on-line: 7 December 2006;
doi:10.1017/S0950268806007540

Abstract

The aim of this study was to describe a systematic process of record-linkage, cross-validation, case-ascertainment and capture-recapture analysis to assess the quality of tuberculosis registers and to estimate the completeness of notification of incident tuberculosis cases in the Netherlands in 1998. After record-linkage and cross-validation 1499 tuberculosis patients were identified, of whom 1298 were notified, resulting in an observed under-notification of 13.4%. After adjustment for possible imperfect record-linkage and remaining false-positive hospital cases observed under-notification was 7.3%. Log-linear capture-recapture analysis initially estimated a total number of 2053 (95%CI 1871-2443) tuberculosis cases, resulting in an estimated under-notification of 36.8%. After adjustment for possible imperfect record-linkage and remaining false-positive hospital cases various capture-recapture models estimated under-notification at 13.6%. One of the reasons for the higher than expected estimated under-notification in a country with a well-organised system of tuberculosis control might be that some tuberculosis cases, e.g. extrapulmonary tuberculosis, are managed by clinicians less familiar with notification of infectious diseases. This study demonstrates the possible impact of violation of assumptions underlying capture-recapture analysis, especially the perfect record-linkage, perfect positive predictive value and absent three-way interaction assumptions.

Introduction

Surveillance of infectious diseases, including tuberculosis, is vital for public health. Mandatory notification is one of the mechanisms to carry out such surveillance but can be contaminated by false-positive cases while true-positive cases may be missed.[1,2] For correct interpretation of tuberculosis figures and the longitudinal trends therein the quality of tuberculosis registers and the completeness of notification should be assessed.[3]

Important in this assessment is record-linkage, i.e. comparing patient data across registers. Record-linkage not only improves completeness of registration but cross-validation with other registers also improves the quality of the data.[3,4] In the Netherlands multiple tuberculosis registers are available. Completeness of notification and other registers can then be assessed relative to the case-ascertainment, i.e. the total number of patients observed in at least one register, or relative to an estimated number of patients through capture-recapture analysis. Based on certain assumptions capture-recapture methods use information on the overlap between registers to estimate the number of cases unknown to all registers and thus the estimated total number of cases.[5] The preferred capture-recapture method entails log-linear modelling of at least three linked registers, less compromised by possible violation of the underlying assumptions compared to capture-recapture analysis based on two linked registers.[6-9] Capture-recapture analysis has been used to assess the completeness of notification and other registers of various infectious diseases,[10] including tuberculosis.[11-15]

The primary objective of this study is to describe a systematic process of record-linkage of different tuberculosis registers, cross-validation, case-ascertainment and capture-recapture estimation of incident tuberculosis cases in the Netherlands in 1998. The secondary objective is to assess the completeness of tuberculosis notification. Under-notification was expected to be low in a country with a well-organised system of tuberculosis control and with a previous estimate of 8% between 1995 and 1998.[16]

Methods

Permission for this study was obtained from the Medical Ethics Committee of the Erasmus Medical Centre in Rotterdam and the data protection committees of the tuberculosis registrations.

Data sources and patient identifiers

Three registers of tuberculosis cases in the Netherlands in 1998 were examined:

1. Patients notified by tuberculosis physicians to the Register of Notifiable Infectious Diseases of the Health Care Inspectorate (Notification).
2. Patients with a positive culture for *Mycobacterium tuberculosis* complex known to the Mycobacteria Reference Unit at the National Institute for Public Health and the Environment (Laboratory).

3. Hospitalised patients recorded by the National Morbidity Registration with an International Code for Diseases (ICD-9) for active tuberculosis (ICD-9 codes 010–018) (Hospital).

Duplicate entries in each register and laboratory contamination records were deleted. Three other tuberculosis-related registers used for cross-validation (exclusion of false-positive tuberculosis cases or verification of assumed true-positive tuberculosis patients among non-culture-confirmed tuberculosis cases) or acquisition of additional patient variables, will be discussed later. For each patient date of birth, postal code, sex, and date of notification, first culture sample or hospital admission were collected as personal identifiers to be used in all record-linkage procedures.

Study year

The reference year chosen was 1998 as of 1 April 1999 only the year of birth is recorded among the mandatory notification data, effectively ruling out reliable record-linkage between the Notification and other registers.[17] Patients with a date of notification, hospital admission or culture-sampling (in order of primacy) between 1 January 1998 and 1 January 1999 were included. To correct for misclassification due to late notification or positive bacteriological results, all three registers were examined between 1 July 1997 and 1 July 1999.

Case-definition

Tuberculosis cases are defined as all observed (by notification, culture-confirmation or hospital admission) and unobserved cases of active tuberculosis (excluding *Mycobacterium bovis* BCG infection). Culture-confirmed patients are assumed true-positive tuberculosis patients.

Record-linkage

Record-linkage was performed manually using the patient identifiers and proximity of date of notification, first culture sample or hospital admission. First the Notification and Laboratory registers were linked. For perfect linkage all patient identifiers should be identical and date of notification and first culture sample should differ by < 1 month. To avoid misclassification of near links with a minor discrepancy in one of the identifiers, e.g. due to clerical errors such as typing mistakes, near-links and cases with a date difference of > 1 month were checked using the surname of the patient. Since the researchers did not know the patients' names due to privacy regulations, a "trusted third party" ascertained match or mismatch. Finally, the Hospital register was linked to the two other registers, using human judgement and consensus in case of near-links.

Cross-validation of cases and collection of additional variables

To improve the positive predictive value of the linked tuberculosis registers, non-culture-confirmed cases were examined through record-linkage with three tuberculosis-related datasets in the Netherlands. Cross-validation was conducted in fours steps. First, cases with disease actually caused by non-tuberculous mycobacteria (NTM) were identified and excluded through record-linkage with the national register for NTM cultures at the Mycobacteria Reference Unit, after a representative check in a large regional laboratory

demonstrated that 80% (143/179) of the local NTM isolates could be found in the national NTM register. Second, patients later diagnosed with disease other than tuberculosis or NTM were identified and excluded through record-linkage with a dataset of such patients secondary to the Netherlands Tuberculosis Register (NTR), an extensive system of voluntary reporting by tuberculosis physicians.[18] Third, non-culture-confirmed patients possibly diagnosed by histopathology examination were verified through the Pathological Anatomy Laboratory Computerised Archive (PALGA), the nation-wide network and registry of histopathology and cytopathology results in the Netherlands. Excerpts of the histopathology reports of linked patients were reviewed by a pathologist and cases with inconsistent results discarded. Finally, the total set of linked tuberculosis registers was linked to the NTR for verification of the remaining non-culture-confirmed tuberculosis patients and collection of additional variables for cases in any of the linked registers: nationality (Dutch, non-Dutch), location of tuberculosis (pulmonary, extrapulmonary) and infectiousness (sputum smear-positive, sputum smear-negative). Although more complete in data the NTR was expected to have a complete overlap with the notification register (both registers are maintained by the same tuberculosis physicians) and was deliberately used for the purpose of validation of the conventional notification, laboratory and hospital tuberculosis registers.[3]

Case-ascertainment, capture-recapture analysis and observed and estimated register-specific coverage rates

The total and stratified observed register-specific coverage rates are defined as the number of tuberculosis patients in each register divided by the total or stratified case-ascertainment, expressed as percentage.

The total number of unobserved tuberculosis cases was estimated on the basis of the cross-validated distribution of the observed cases over the Notification, Laboratory and Hospital registers. The independence of registers and other assumptions underlying capture-recapture analysis have been described previously.[10] Interdependencies between the three tuberculosis registers 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.[12,15] Estimated register-specific coverage rates are defined as the number of tuberculosis patients in each register divided by the estimated total number of tuberculosis patients by capture-recapture analysis.

Results

Table 6.1 shows the initial number of cases, the number of cases excluded from the study before and after record-linkage and the final number of cases in the three tuberculosis registers in the Netherlands in 1998. The hospital admission of 12 cases in 1997 and 8 cases in 1999, all notified in 1998, was included in the data.

Among the 295 near-links between the Notification and Laboratory registers, the "trusted third party" confirmed 267 candidate-pairs as true links. Among the confirmed links, 133 candidate-pairs had administrative discrepancies, predominantly (63.8%) in the postal code.

Record-linkage of all 537 non-culture-confirmed cases to the NTM register and the subset of the NTR revealed that despite NTM infection or any other diagnosis than tuberculosis 26 out of 426 non-culture-confirmed cases on the Notification register (6.1%) were not de-notified and 25 out of 217 non-culture-confirmed cases on the Hospital register (11.5%) were still recorded with an ICD-9 tuberculosis code. Figure 6.1 shows the distribution of the final number of 1499 cases over the different tuberculosis registers. Of the 1006 culture-confirmed tuberculosis patients 108 patients (10.7%) could not be found in the Notification register.

Verification through PALGA of the remaining 493 non-culture-confirmed cases in the linked registers identified 117 patients (23.7%) with a histopathology report consistent with active tuberculosis. Verification through the NTR identified 385 patients (78.1%). Both exercises combined verified 407 patients (82.6%). Figure 6.2 shows the distribution of the PALGA and NTR verified non-culture-confirmed cases over the three linked tuberculosis registers. In total 94.3% (1413/1499) of all patients were culture-confirmed or verified but only 37.6% (35/93) of the unlinked hospital patients.

Record-linkage of patients observed in any of the three linked tuberculosis registers with the NTR resulted in a coverage of 91.1%, 84.7% and 78.9% of the Notification, Laboratory and Hospital registers respectively. Of the 108 culture-confirmed tuberculosis patients not found in the Notification register 38 (35%) were voluntarily reported to the NTR.

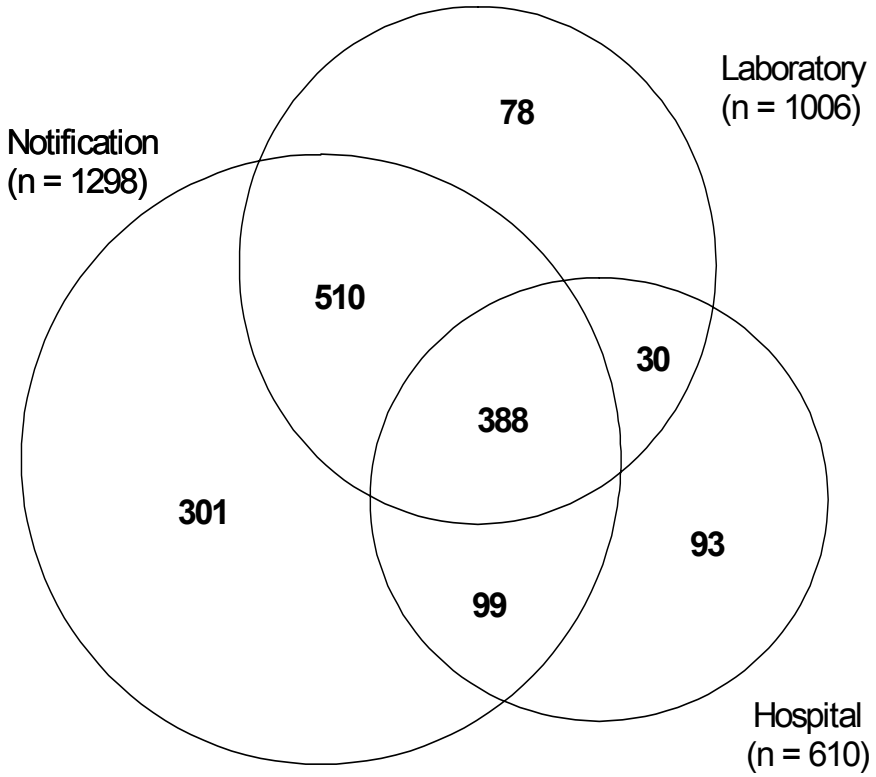
The total and stratified observed number of tuberculosis patients and register-specific coverage rates of the three tuberculosis registers are shown in Table 6.2. Observed completeness of notification, culture-confirmation and hospitalisation is 86.6%, 67.1% and 40.7% respectively. The completeness of the Notification register is consistent over the strata, with non-culture-confirmed patients least likely to be notified. The Laboratory and Hospital registers have higher proportions of sputum smear-positive patients and both registers show a trend of culture-confirmation and hospitalisation increasing with age. If only culture-confirmed or otherwise verified cases were included the verified observed completeness of the Notification register would be 89.9%. The observed and verified observed under-notification is 13.4% and 10.1% respectively. When all 58 non-verified unlinked hospital cases are considered false-positive and the 38 culture-confirmed patients reported to the NTR considered notified, the adjusted observed under-notification is 7.3% (105/1441).

Based on the Akaike Information Criterion the log-linear capture-recapture procedure initially selected the saturated model (see Discussion) as the best-fitting model which estimated 554 unobserved tuberculosis cases, resulting in an estimated total number of 2053 (95% confidence interval (CI) 1871-2443) tuberculosis cases. This translates into an estimated completeness of case-ascertainment of 73.0% (1499/2053) and estimated register-specific coverage rates of 63.2%, 49.0% and 29.7% for the Notification, Laboratory and Hospital registers respectively. The estimated under-notification is 36.8% (95%CI 30.6-46.9%).

Table 6.1 The initial number of cases, the number of cases excluded from the study before and after record-linkage and the final number of cases in the three tuberculosis registers in the Netherlands in 1998.

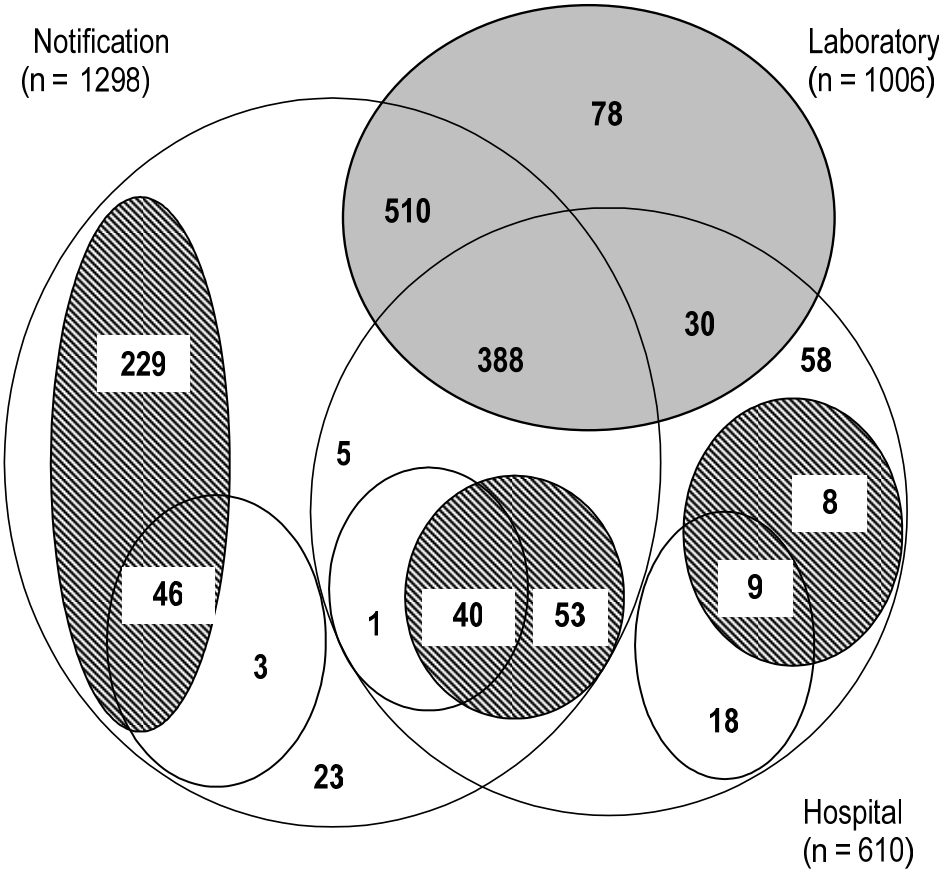
	Tuberculosis registers		
	Notification	Laboratory	Hospital
Patients initially found in the different tuberculosis registers in the Netherlands	1334	1074	658
Patients excluded from the analysis			
Patients lost during matching process	3	0	0
Duplicate entry laboratory register	0	1	0
Duplicate entry notification register	1	0	0
Laboratory contamination	6	19	2
Culture of *Mycobacterium bovis* BCG	0	14	1
Subtotal before record linkage	1324	1040	655
Patients with a laboratory sample date in 1998 but notified in 1997	0	3	2
Patients with a laboratory sample date in 1998 but notified in 1999	0	29	6
Patients only known to the hospital in 1998 but notified in 1997	0	0	10
Patients only known to the hospital in 1998 but notified in 1999	0	0	2
Patients not notified with a laboratory sample date in 1998 but admitted to the hospital in 1999	0	2	0
Patients with initial tuberculosis notification in 1998 but diagnosis later withdrawn because of non-tuberculous mycobacteria (n = 35; 7 patients appear in both registers)	19	0	23
Patients with initial tuberculosis notification in 1998 but diagnosis later withdrawn because of other reasons than non-tuberculous mycobacteria	7	0	2
Patients included in the capture recapture analysis	1298	1006	610

Figure 6.1 Schematic view of the distribution of observed number of tuberculosis patients in the Netherlands in 1998, after record-linkage of three tuberculosis registers (total number of observed cases is 1499).



After adjustment for the 58 possibly false-positive unlinked hospital cases and the 38 possibly misclassified laboratory patients (Figure 6.3) the selected, most parsimonious, log-linear capture-recapture model was the model with two two-way interactions between Notification and Laboratory and between Notification and Hospital. The small likelihood ratio, G^2, compared with the number of degrees of freedom (df), shows that this model fits the data well (G^2 = 0.053; df = 2; P = 0.974; Akaike Information Criterion = –3.95) and estimates 1547 (95%CI 1513-1600) tuberculosis patients. The completeness of case-ascertainment after the adjustment is 93.1% (1441/1547) and the estimated register-specific coverage rates are 86.4%, 65.0% and 35.7% for the Notification, Laboratory and Hospital registers respectively. Adjusted estimated under-notification is 13.6% (95%CI 11.7-16.5%).

Figure 6.2 Schematic view of the distribution of observed number of tuberculosis patients in the Netherlands in 1998, after record-linkage of three tuberculosis registers (light grey = culture-positive), and the number of validated tuberculosis patients among the culture-negative cases (dark grey = Netherlands Tuberculosis Register; white = Pathological Anatomy Laboratory Computerised Archive).



Discussion

Main findings

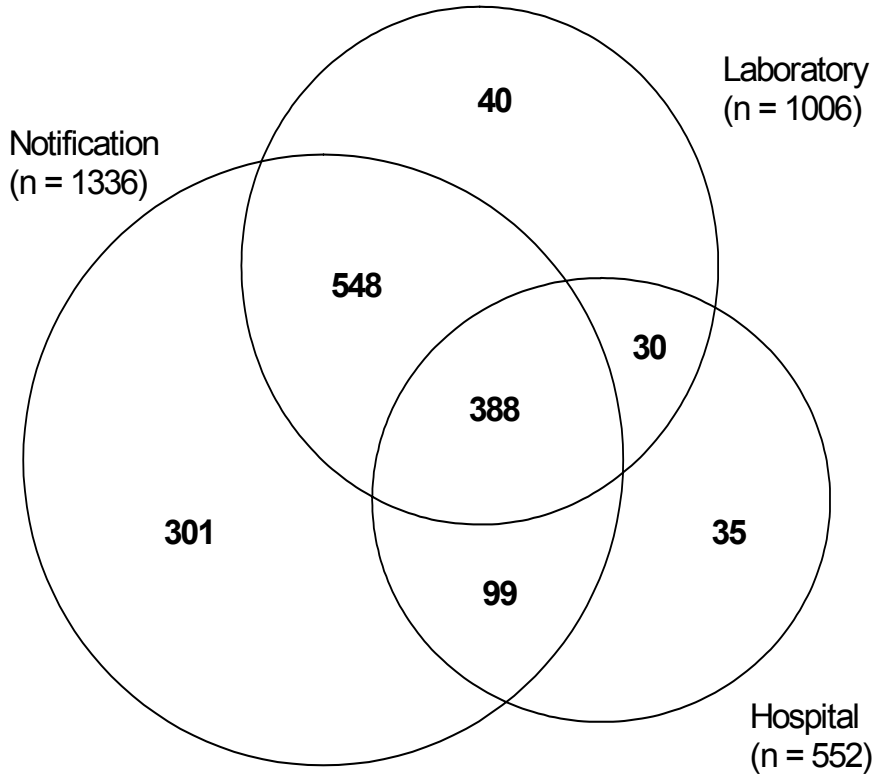
This study shows that, even in a country with a well-organised system of tuberculosis control, record-linkage and cross-validation improve the data quality of tuberculosis registration and case-ascertainment. These findings underscore the need for scrutiny of all tuberculosis registers, especially with regard to hospital-based data. Total and verified observed under-notification of tuberculosis in the Netherlands in 1998 was 13.4% and 10.1% respectively. The latter was slightly higher than a previously reported under-

Table 6.2 Total and stratified number of tuberculosis cases identified by three tuberculosis registers and observed register-specific fractions.

	Observed cases	Notification		Laboratory		Hospital	
	N (%)	Frequency	Fraction	Frequency	Fraction	Frequency	Fraction
Total	1499	1298	86.6%	1006	67.1%	610	40.7%
Male*	849 (57.2)	747	88.0%	580	68.3%	357	42.0%
Female*	635 (42.8)	541	85.2%	411	64.7%	251	39.5%
Dutch†	389 (32.0)	372	95.6%	250	64.3%	157	40.4%
Non-Dutch†	826 (68.0)	790	95.6%	588	71.2%	316	38.3%
Pulmonary tuberculosis‡	770 (62.2)	734	95.3%	545	70.8%	296	38.4%
Extra-pulmonary tuberculosis‡	467 (37.8)	448	95.9%	307	65.8%	185	39.6%
Sputum smear-positive§	276 (42.3)	265	96.0%	243	88.0%	149	54.0%
Sputum smear-negative§	376 (57.7)	358	95.2%	237	63.0%	105	28.0%
< 15yrs	101 (6.7)	89	88.1%	40	39.6%	37	36.6%
≥15yrs and <65yrs	1150 (76.7)	1000	90.0%	790	68.7%	450	39.1%
≥65yrs	248 (16.5)	209	84.3%	176	84.3%	123	49.6%
Culture-confirmed cases	1006 (67.1)	896	89.0%	1006	100%	418	41.6%
Non-Culture-confirmed cases	493 (32.9)	402	81.1%	0	0%	192	38.9%
Metropolitan	477 (31.8)	418	87.6%	333	69.8%	182	38.2%
Non-Metropolitan	1022 (68.2)	880	86.1%	673	65.9%	428	41.9%

* for 15 cases no information was available
† for 284 cases no information was available
‡ for 262 cases no information was available
§ for 847 cases no information was available or were non-pulmonary tuberculosis.

Figure 6.3 Schematic view of the distribution of observed number of tuberculosis patients in the Netherlands in 1998, after record-linkage of three tuberculosis registers, and correction for possible misclassification of culture-positive patients and remaining false-positive unlinked hospital cases (total number of observed cases is 1441).



notification of 8%. After correction for possibly misclassified laboratory patients and remaining false-positive hospital cases the adjusted observed under-notification of 7.3% is similar to this previous estimate. The 36.8% under-notification estimated by a log-linear capture-recapture model before adjustments were made is highly inconsistent with the prior report. Adjustment for possible misclassification of laboratory patients and remaining false-positive hospital cases had a considerable impact on the log-linear capture-recapture estimate.

Possible causes of poor data quality

The quality of the tuberculosis registers is mainly determined by the proportion of administrative discrepancies causing possible record-linkage misclassification (8.6% between Notification and Laboratory) and the proportion of false-positive cases (8.2% among non-culture-confirmed cases in this study after previous elimination of laboratory contamination records and exclusion of *M. bovis* BCG isolates). The majority of

administrative discrepancies were found in the postal code. Apart from clerical errors, this could be due for example to frequent transfers of asylum seekers, notification of home address of prisoners versus laboratory postcode of prison region or assigning a random local postal code to records with missing data in some registers. Patients with a culture of *M. bovis* BCG were excluded because of an expected low positive predictive value for systemic disease as all were either infants (with likely a post-BCG vaccination abscess) or older males (with probable urological *M. bovis* BCG instillation).

Despite maximum efforts to eliminate administrative discrepancies and false-positive records, our results still indicate imperfect record-linkage as, assuming a negligible number of lost reports, only 91.1% of all tuberculosis cases in the Notification register could be linked to the NTR. Since tuberculosis physicians report to both registers the expected overlap is 100%. A proportion of the tuberculosis cases in the final dataset not present in the Notification register could be explained by imperfect record-linkage because, remarkably, 38 culture-confirmed but not notified patients were voluntarily reported to the NTR, suggesting notification as well. After adjustment the number of patients in the Notification register (1336) is almost similar as the number reported by the NTR in 1998 (1341). Still 70 culture-confirmed patients may not have been notified, reflecting the most serious public health aspect of under-notification, i.e. preventing possibly indicated contact investigations around potentially infectious patients.

In almost one-quarter of the non-culture-confirmed patients histopathology examination contributed to the diagnosis tuberculosis. The majority of these patients were found in the Hospital register which is plausible because histopathology examination is more likely to be performed as part of a diagnostic work-up in patients with extrapulmonary tuberculosis requiring hospital admission. In the Netherlands, the contribution of PALGA to case-verification in addition to the NTR was limited.

Despite the availability of additional tuberculosis-related registers, the majority (62.4%) of unlinked hospital cases could not be verified, compared to 7.6% of the unlinked notified cases. Although often used as a third data source in capture-recapture studies on human disease incidence, in the case of tuberculosis the data quality of hospital registers should be judged critically. A local capture-recapture study in the United Kingdom found 27% of all tuberculosis cases in the hospital register to be false-positive and in a regional capture-recapture study in Italy this was even 80% among unlinked hospital tuberculosis cases.[12,15]

Limitations

The findings have to be placed in the context of the limitations of this study. The estimated coverage of the tuberculosis registers was based on three-source log-linear capture-recapture models. These models are 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 registers (i.e. the probability of being observed in one register is not affected by being (or not being) observed in another). In the three-source

capture-recapture approach dependencies between two registers can be identified and incorporated in the log-linear model.[5] The three-way interaction however, i.e. dependency between all three registers, cannot be incorporated in the model and its absence must be assumed. Nevertheless, violation of this assumption may occur, rendering capture-recapture analysis outcomes less valid. This and other limitations of capture-recapture analysis are described elsewhere in more detail.[8,19-25]

In this study, the possible remaining false-positive cases and violation of the perfect record-linkage assumption have already been discussed. Violation of the closed population assumption is presumably limited as with tuberculosis the opportunities for notification, culture-confirmation or hospitalisation are largely determined within a short period of time but could result in overestimation of the number of patients. More likely is violation of the absent three-way interaction assumption. Tuberculosis services in the Netherlands are organized around close collaboration between clinicians, microbiologists and public health professionals such as tuberculosis physicians and tuberculosis nurses. Examples of this collaboration are laboratory pre-notification, clinical isolation, contact-investigations and referrals, explaining the two two-way interactions identified in the final log-linear capture-recapture model. The initial log-linear capture-recapture model with the best goodness-of-fit was the saturated model, i.e. including all two-way interactions. Violation of the absent three-way interaction assumption, which biased our estimates of the true population size, cannot be ruled out.[8,21,23,26] Also more likely is violation of the homogeneity assumption: age, location of disease and infectiousness, among others, can account for different probabilities of being seen in a tuberculosis register. Although at least as vulnerable as log-linear models to violation of underlying assumptions, to investigate possible bias as a result of violation of the homogeneity assumption, we have examined the data again with alternative estimators, as described in the capture-recapture analysis literature.[8,27] These estimators reportedly perform well when compared to log-linear capture-recapture estimates,[28] are arguably more robust to violation of the homogeneity assumption[29] and have been used in social sciences to estimate the size of hidden populations such as illicit drug users and homeless persons.[29-32] We applied Chao's heterogeneity and bias-corrected homogeneity models on the adjusted observed distribution of tuberculosis patients.[33-35] Both models estimate a total of 1545 tuberculosis patients (95%CI 1519-1580), very similar to the log-linear model, with an estimated case-ascertainment of 93.3% (1441/1545) and an estimated under-notification of 13.5% (95%CI 12.0-15.4). The CI of the adjusted log-linear and alternative estimates does not contain the expected value of 8%.

Improving tuberculosis surveillance systems

Some ways of improving the performance of tuberculosis (and other infectious disease) surveillance systems could be:

- As an alternative to log-linear three-source capture-recapture analysis to estimate tuberculosis incidence, record-linkage, preferably web-based, between the two most relevant sources for tuberculosis surveillance, namely the Notification and Laboratory registers, both having a high positive predictive value, will improve timeliness of reporting, completeness of demographic, microbiological and epidemiological variables of

the patients, and completeness of the number of patients and hence observed tuberculosis incidence.

- Treatment of all tuberculosis patients, including extrapulmonary cases, by a limited group of experienced specialist physicians, such as tuberculosis physicians, chest-physicians or infectiologists, familiar with notification procedures, will improve completeness of notification.

- The introduction of pre-notification of positive laboratory test results for tuberculosis to the public health physicians responsible for processing the notifications from the local clinicians to the Health Care Inspectorate at the national level, with subsequent follow-up of unreported cases, as implemented in some regions of the Netherlands, will also improve completeness of notification.

Conclusion

Tuberculosis under-notification in the Netherlands in 1998 is probably around 8% and possibly around 13.6%. This study demonstrates the need for assessment of tuberculosis registers for quality of the data and completeness, and the importance of record-linkage.[22] It underscores that 'as for the results of all epidemiological investigations, the credibility of any capture-recapture estimate will be enhanced to the extent that the investigator may be able and confirm the accuracy of all information used, such as diagnosis, location of the case within the space-time interval analysed, and appropriate case matching, as with capture-recapture methods, errors are highly likely to have a more than additive effect on estimates'.[8, 36]

Acknowledgements

We thank Nico Kalisvaart of the KNCV Tuberculosis Foundation, Dr Bert Mulder and Karel Nolsen of the Regional Laboratory for Microbiology Twente, Matty Meijer of the Register of Notifiable Infectious Diseases of the Health Care Inspectorate, Willem Hoogen Stoevenbelt of the National Morbidity Registration, Dr Mariel Casparie of PALGA, and all Departments of Tuberculosis Control of the Public Health Services in the Netherlands for technical assistance and co-operation.

References

1. Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. Global burden of tuberculosis. *JAMA* 1999; 282: 677-86.
2. World Health Organization. *Global tuberculosis control: surveillance, planning, financing. WHO Report 2004*. Geneva: World Health Organization, 2004.
3. Migliori GB, Spanevello A, Ballardini L, Neri M, Gambarini C, Moro ML, Trinka L, Raviglione MC. Validation of the surveillance system for new cases of tuberculosis in a province of northern Italy. Varese Tuberculosis Study Group. *Eur Respir J* 1995; 8: 1252-8.
4. Mukerjee AK. Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data sources. *Commun Dis Public Health* 1999; 2: 143-4.
5. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
6. Fienberg SE. The multiple-recapture census for closed populations and the 2k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
7. Bishop YMM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis*. Cambridge: MIT-Press, 1975.

8. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-63.
9. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation II: Applications in human diseases. *Am J Epidemiol* 1995; 142: 1059-68.
10. Van Hest NA, Smit F, Verhave JP. Underreporting of malaria incidence in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-7.
11. Sanghavi DM, Gilman RH, Lescano-Guevara AG, Checkley W, Cabrera LZ, Cardenas V. Hyperendemic pulmonary tuberculosis in a Peruvian shantytown. *Am J Epidemiol* 1998; 148: 384-9.
12. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-3.
13. Mayoral Cortes JM, Garcia Fernandez M, Varela Santos MC, Fernandez Merino JC, Garcia Leon J, Herrera Guibert D, Martinez Navarro F. Incidence of pulmonary tuberculosis and HIV coinfection in the province of Seville, Spain, 1998. *Eur J Epidemiol* 2001; 17: 737-42.
14. Cailhol J, Che D, Jarlier V, Decludt B, Robert J. Incidence of tuberculous meningitis in France, 2000: a capture-recapture analysis. *Int Tuberc Lung Dis* 2005; 9: 803-8.
15. Baussano I, Bugiani M, Gregori D, Van Hest R, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung Dis* 2006; 10: 415-21.
16. Van Loenhout-Rooijackers JH, Leufkens HGM, Hekster YA, Kalisvaart NA. Pyrazinamide use as a method to estimate under-reporting of tuberculosis. *Int J Tuberc Lung Dis* 2001; 5: 1156-60.
17. Klein S, Bosman A. Completeness of malaria notification in the Netherlands 1995-2003 assessed by capture-recapture method. *Euro Surveill* 2005; 10: 244-6.
18. KNCV Tuberculosis Foundation. *Index Tuberculosis 2001-2002 – Netherlands*. The Hague: KNCV Tuberculosis Foundation, 2005.
19. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *Int J Epidemiol* 1994; 23: 1322-3.
20. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995; 6: 42-8.
21. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *J Clin Epidemiol* 1999; 52: 909-14.
22. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1999; 25: 474-8.
23. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* 2000; 152: 771-9.
24. Jarvis SN, Lowe PL, Avery A, Levene S, Cormack RM. Children are not goldfish-mark/recapture techniques and their application to injury data. *Inj Prev* 2000; 6: 46-50.
25. Tilling K. Capture-recapture methods-useful or misleading? *Int J Epidemiol* 2001; 30: 12-4.
26. Regal RR, Hook EB. Marginal versus conditional versus structural source models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Stat Med* 1998; 17: 69-74.
27. Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; 79: 543-53.
28. Hook EB, Regal RR. Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *Am J Epidemiol* 1982; 116: 168-76.
29. Smit F, Reinking D, Reijerse M. Estimating the number of people eligible for health service use. *Eval Program Plan* 2002; 25: 101-105.
30. 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.
31. Bohning D, Suppawattanabodee B, Kusolvitkul, W, Viwatwongkasem C. Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *Eur J Epidemiol* 2004; 19: 1075-83.
32. Hay G, Smit F. Estimating the number of hard drug users from needle-exchange data. *Addiction Res Theory* 2003; 11: 235-43.
33. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; 43: 783-91.
34. Chao A. Estimating animal abundance with capture frequency data. *J Wildl Manage*1988; 52: 295-300.
35. Chao A. Estimating population size for sparse data in capture-recapture experiments. *Biometrics* 1989; 45: 427-38.
36. Seber GA, Huakau JT, Simmons D. Capture-recapture, epidemiology, and list mismatches: two lists. *Biometrics* 2000; 56: 1227-32

7

Undetected burden of tuberculosis in a low-prevalence area

I. BAUSSANO[1], M. BUGIANI[2], D. GREGORI[3], R. VAN HEST[5,6], A. BORRACINO[3], R. RASO[4] and F. MERLETTI[1]

1 Cancer Epidemiology Unit, San Giovanni Battista Hospital and University of Turin, Turin
2 Local Health Unit 4, Tuberculosis Prevention Service, Turin
3 Department of Public Health and Microbiology, University of Turin, Turin
4 Epidemiology Unit, Local Health Unit 20, Alessandria
Italy
5 Department of Tuberculosis Control, Municipal Public Health Service Rotterdam Area, Rotterdam
6 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam
the Netherlands

Int J Tuberc Lung Dis 2006; 10: 415-21

Abstract

Setting: Under-ascertainment and underreporting of tuberculosis hampers surveillance and control. Case detection is improved by record-linkage of case registers and under-reporting can be estimated by capture-recapture analysis.

Objectives: To assess the completeness of the tuberculosis registration systems and estimation of tuberculosis incidence and underreporting in the Piedmont Region of Italy in 2001.

Methods: Record-linkage of the 'physician notification system', the tuberculosis laboratory register and the hospital records register, and subsequent three-sample capture-recapture analysis.

Results: Record-linkage identified 657 tuberculosis cases; capture-recapture analysis estimated 47 (95%CI 31-71) unrecorded cases. Underreporting of the 'physician notification system' was estimated at 21% (95%CI 20-23%). The overall estimated tuberculosis incidence rate was 16.7 cases per 100 000 population (95%CI 16.3-17.3), varying according to subset investigated: 12.7 for individuals from low tuberculosis prevalence countries and 214.1 for immigrants from high tuberculosis prevalence countries; 13.1 and 25.8 for persons < and ≥ 60 years respectively; and 32.1 in Turin, the regional capital versus 10.8 in the rest of the region.

Conclusions: When multiple recording systems are available, record-linkage and capture-recapture analysis can be used to assess tuberculosis incidence and the completeness of different registers, contributing to a more accurate surveillance of local tuberculosis epidemiology.

Introduction

Meaningful quantification and description of the distribution of tuberculosis within a community is an essential part of any tuberculosis control programme.[1,2] Underreporting by local surveillance systems in countries with high and low endemicity for tuberculosis leads to underestimation of the tuberculosis burden and makes descriptions and interpretation of spatial and temporal variations unreliable.[3,4] In 2003, the World Health Organization (WHO) estimated that underreporting of tuberculosis in Italy was 12%[5] but according to other reports it reached 37-54% in some areas of the country.[6,7]

Case detection can be improved by record-linkage, i.e. comparing patient data across multiple registers,[6] and underreporting can be estimated by capture-recapture analysis. The latter uses information after record-linkage of various datasets, evidenced by the observed overlap of the registers, to estimate the number of cases unknown to all sources.[8] Capture-recapture analysis was first used in studies of animal population biology and, more recently, in epidemiology.[8-10] It is now increasingly used to estimate the burden of both non-communicable[11,12] and communicable diseases,[13,14] including tuberculosis.[4,15,16] We undertook record-linkage of multiple information systems and subsequently conducted a capture-recapture analysis to estimate the tuberculosis incidence in the Piedmont Region of Italy in 2001 and to assess the performance of the surveillance system.

Study population and methods

Study population and case-definition

We focused the study on residents of the Piedmont Region, Italy, during 2001. According to the fourteenth national census in 2001, the total resident population of the Piedmont Region was 4 214 677, of whom 2 034 161 (48%) were men, 3 027 034 (72%) were age < 60 years, 865 263 (20%) lived in Turin, the capital, and 84 070 (2%) were immigrants from high tuberculosis burden countries (HTBCs), i.e. countries with an annual incidence > 80 cases per 100 000 population. About one third of the immigrants were from North Africa, one third from Eastern Europe or the former Soviet Union, and the remainder came from Asia, sub-Saharan Africa and Latin America.[17]

We included in the study all new cases of pulmonary tuberculosis and non-pulmonary tuberculosis, diagnosed in the Piedmont Region in 2001 and known to at least one of three tuberculosis registers. Tuberculosis cases were defined according to the guidelines of WHO and the European Region of the International Union Against Tuberculosis and Lung Disease Working Group for Uniform Reporting on Tuberculosis Cases.[1,18] Cases were classified as follows: confirmed (culture-confirmed or smear-positive) or probable cases (clinically, radiologically or empirically diagnosed); pulmonary tuberculosis or non-pulmonary tuberculosis; patients < or ≥ 60 years; resident in the Turin metropolitan area or in the remaining parts of Piedmont; and born in HTBCs or in low tuberculosis burden countries (LTBCs), i.e. countries with an annual incidence < 80 cases/100 000 population. Cases caused by environmental mycobacteria (21 records) were excluded to improve the specificity and the positive predictive value of each register.

The research was conducted on mandatory regional registries set up following regional and national law; therefore, according to national legislation, no informed consent is required to obtain and store the information for public health and research purposes. The authors of the paper were authorised by the regional public health authorities to keep and analyse the data and to produce reports.

Sources of cases and record-linkage

Three sources were used to identify tuberculosis cases between 2000 and the first half of 2002. The first was the 'physician notification system', including both notification and treatment outcome monitoring registers. The second source was the laboratory tuberculosis register, which collects reports of microscopic and culture identification of mycobacteria from the regional reference microbiology laboratories. The local public health service periodically checks these records for false-positive reports due to environmental mycobacteria and laboratory cross-contamination. Data from the 'physician notification system' and laboratory sources are not routinely merged and, according to national legislation, only the notification register contributes to the official national tuberculosis statistics. The third source of cases was the hospital discharge records register. Hospital discharge records including any form of tuberculosis (International Classification of Diseases-9 codes 0.10-0.18 and 647.3) were selected.

After correction for duplicate entries in each of the three registers, the records of tuberculosis cases were matched by a deterministic linkage procedure using the identifiers full name, date of birth and sex. Apparent matches were reviewed to avoid homonymous and synonymous errors. Prevalent cases diagnosed in 2000 were identified and were excluded from the study, whereas cases incident in 2001 were corrected for late reporting in the first half of 2002. A case-verification procedure was performed by inspecting the hospital charts of patients identified uniquely in this source to improve the positive predictive value of this register. A similar procedure was not performed for cases identified in the other sources, as case-verification is regularly performed by the public health care services. We defined observed source-specific sensitivities as the number of tuberculosis patients in each register divided by the total number of tuberculosis patients observed after record-linkage. As local tuberculosis surveillance and control guidelines advise to investigate the human immunodeficiency virus (HIV) status of adults with tuberculosis after obtaining consent, information on HIV status was also collected.

Capture-recapture analysis

To use log-linear models for capture-recapture analysis, data from at least three different, partially overlapping and preferably independent sources are necessary.[8,19] The annual incidence and the estimated source-specific sensitivity (i.e. the number of observed tuberculosis patients in each of the investigated sources divided by the estimated total number of tuberculosis patients by capture-recapture analysis) of the regional tuberculosis surveillance system were estimated by a three-sample capture-recapture analysis.[19] Pair-wise dependency between sources was incorporated into the log-linear models and possible capture heterogeneity was tested. Capture-recapture analysis was conducted on the full set of data and repeated for subsets defined according to geographical origin, location of tuberculosis, age group, bacteriological status and site of residence, as

previously specified. For bacteriological status, due to the availability of only two sources for culture-negative tuberculosis patients, a separate calculation was made for microbiologically confirmed and unconfirmed tuberculosis cases.

Statistical analyses were conducted using the STATA version 8 software package (Stata Corp, College Station, TX, USA) and the S-PLUS 2000 software package (Mathsoft Inc, Seattle, WA, USA) with the CARE library.[20] Model selection was based on three statistical criteria: deviance, the Akaike information criterion and the Bayesian information criterion, to limit the risk of selecting unstable or over-complex models. Point estimates and relative 95% confidence intervals (CIs) for the number of unrecorded cases were obtained using the method of Chao et al.[20]

Results

Overall, we identified 657 incident cases of tuberculosis in the Piedmont Region in 2001, with 557 cases from the surveillance system, 406 from hospital discharge records and 285 from laboratory records (69 microscopically identified and 216 confirmed by culture). Figure 7.1 shows the distribution of all identified cases by source and their overlap, whereas Figure 7.2 shows the distribution of microbiologically confirmed cases. A

Figure 7.1 Distribution of all cases of tuberculosis found in the investigated sources.

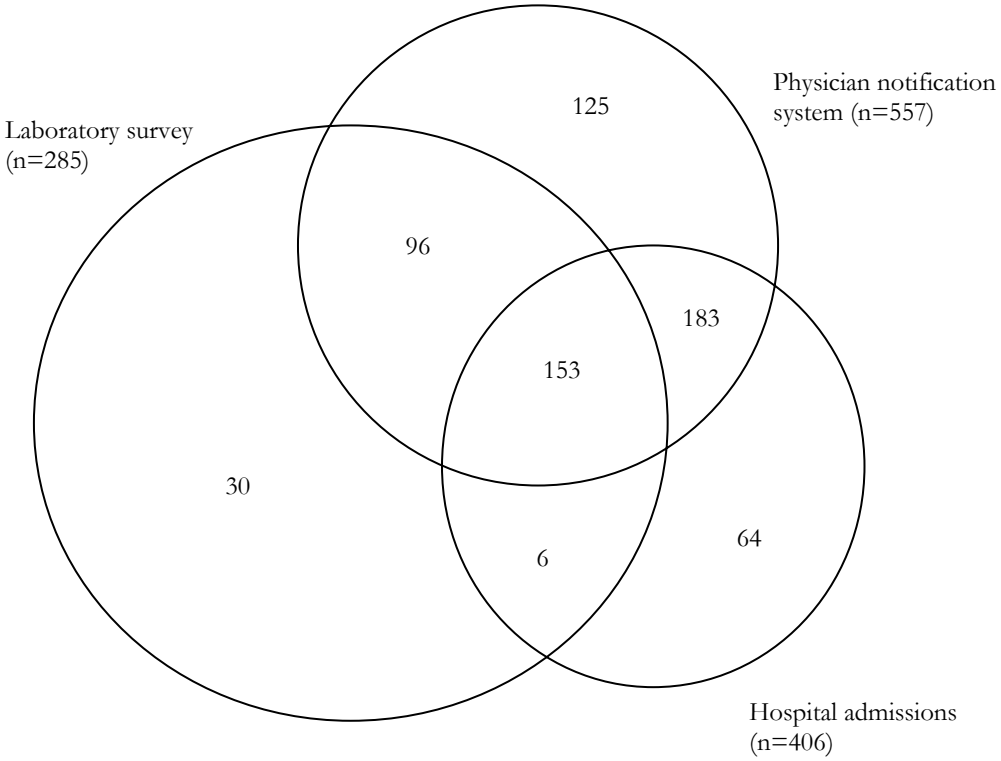
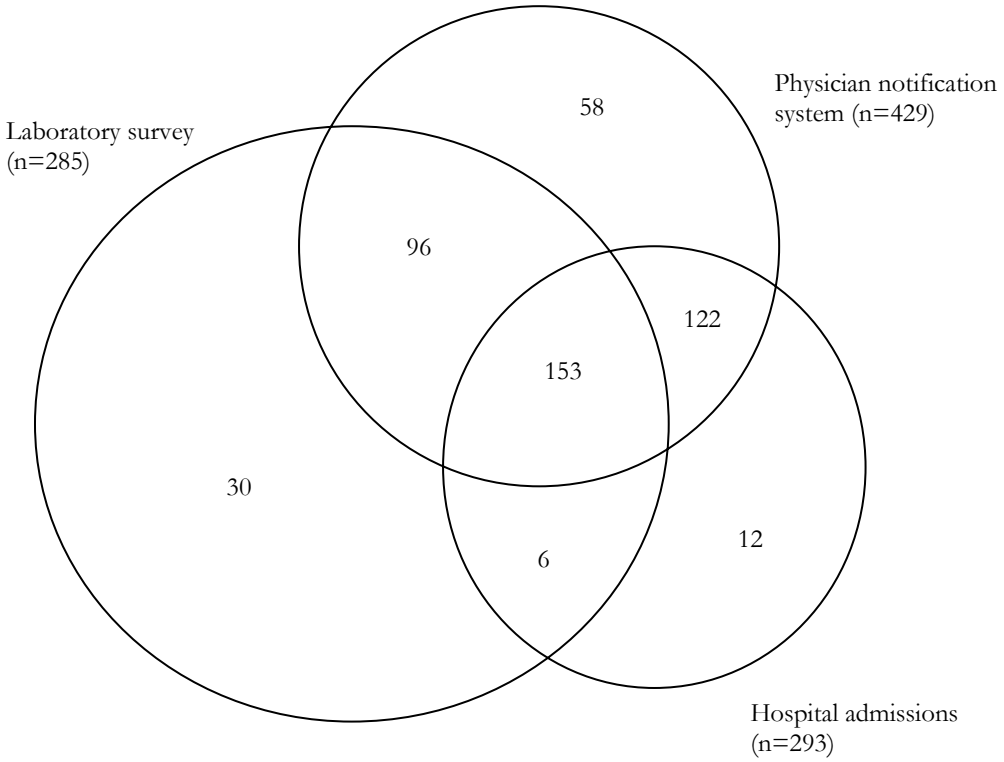


Figure 7.2 Distribution of microbiologically confirmed cases of tuberculosis found in the investigated sources.



verification procedure carried out for the 322 cases identified uniquely from hospital discharge records confirmed 64 cases, an overall positive predictive value of 63%. Table 7.1 shows the total and subset distribution of the tuberculosis cases identified. The sexes were equally represented in the different geographic areas: 286 males were from LTBCs (59% of all individuals from LTBCs) and 91 from HTBCs (54% of all individuals from HTBCs). The age distribution was bimodal, with a difference of > 30 years in average age at diagnosis between immigrants from HTBCs (median age 32 years; range 1-78) and those born in LTBCs (median age 63 years; range 1-101).

There was no difference in age distribution between males (median, 54 years; range, 1-101) and females (median 53 years; range 2-96). HIV status was known to be positive in 32 tuberculosis patients (5%), 23 (72%) of whom were from industrialised countries (data not shown).

Observed source-specific sensitivity

The overall proportion of cases detected in the 'physician notification system', was 84.1% (71.1% for probable cases; 89.9% for confirmed cases) (Table 7.1). This system was more sensitive for identifying persons from HTBCs than for those from LTBCs (91.2% vs.

Table 7.1 Total and subset numbers of cases of tuberculosis identified by investigated sources and source-specific observed sensitivities

	Observed cases		Physician notification system		Laboratory		Hospital discharge records	
	n (%)		Frequency	Sensitivity	Frequency	Sensitivity	Frequency	Sensitivity
Total	657		557	84.8	285	43.4	406	61.8
Sex								
Male*	377	(58)	322	85.4	168	44.6	226	59.9
Female*	278	(42)	235	84.5	115	41.4	180	64.7
Area of origin								
Low TB prevalence countries	486	(74)	401	82.5	198	40.7	319	65.6
High TB prevalence countries	171	(26)	156	91.2	87	50.9	87	50.9
Location of TB								
Pulmonary TB†	434	(66)	387	89.2	250	57.6	268	61.7
Extra-pulmonary TB‡	211	(32)	158	74.9	52	24.6	193	91.5
Age								
<60 years	374	(57)	328	87.7	155	41.4	232	62.0
≥60 years	283	(43)	229	80.9	130	45.9	174	61.5
Status								
Confirmed cases	477	(73)	429	89.9	285	59.7	293	61.4
Probable cases	180	(27)	128	71.1	0	0.0	113	62.8
Area of residence								
Turin‡	271	(41)	256	94.5	108	39.9	177	65.3
Piedmont, excluding Turin‡	331	(50)	276	83.4	147	44.4	229	69.2

TB = tuberculosis; * No information available for two cases; † No information available for 70 cases; ‡ No information available for 55 cases

82.5%), for pulmonary tuberculosis than for non-pulmonary tuberculosis (89.2% vs. 74.9%), for tuberculosis patients aged < 60 than for older patients (87.7% vs. 80.9%) and for persons from the Turin metropolitan area than from the rest of the Piedmont Region (94.5% vs. 83.4%). The sensitivity of hospital discharge records was 61.8%, ranging from 50.9% for persons from HTBCs to more than 90% for non-pulmonary cases. The laboratory source had the lowest overall sensitivity (43.3%) and the highest sensitivity levels for this source were for confirmed (59.7%) and pulmonary tuberculosis cases (57.6%).

Capture-recapture models

The estimates for each log-linear model are shown in Table 7.2. The selected model allowed for capture dependency between the surveillance and laboratory sources, and did not take into account heterogeneity (deviance 27.6; standard error (SE) 10). Three models with appealing goodness-of-fit criteria were rejected because their estimates were unstable, as reflected by the high standard error. The selected model estimated 47 (95%CI 31-71) tuberculosis patients unknown to all three sources, resulting in an estimated total of 704 (95%CI 688-728) incident cases of tuberculosis in the Piedmont Region in 2001. We then estimated the number of tuberculosis cases in various subsets (Table 7.3), using the same log-linear model for all grouping variables. The total number of microbiologically confirmed cases was estimated, using three sources, at 500 (95%CI 490–517). The number of probable tuberculosis cases, which by case-definition cannot be captured by the laboratory source, was estimated, using two sources, at 237 (95%CI 214-273). The 95%CI of the total number of tuberculosis cases estimated by geographic origin, location of tuberculosis, age group, and bacteriological status overlapped with the 95%CI of the non-stratified estimate.

Estimated source-specific sensitivity and incidence

The overall estimated ascertainment of tuberculosis cases (i.e. cases recorded in at least one of the registers examined) was 93.3%. Although notification of diagnosis and treatment of tuberculosis is mandatory, the estimated sensitivity of the 'physician notification system' system was 79.1% (95%CI 76.5-80.1%) (Table 7.3). The system performed better in the Turin metropolitan area (sensitivity 92.1%). The analysis showed that persons aged ≥ 60 years (sensitivity 74.8%) and non-pulmonary cases (sensitivity, 74.2%) are relatively underreported or underdetected. The system was more likely to capture cases in persons from HTBCs (sensitivity 86.1%) than in those from LTBCs (sensitivity 76.8%).

The estimated overall annual tuberculosis incidence rate was 16.7/100 000, with 11.9 cases per 100 000 population microbiologically confirmed. The incidence estimates varied widely according to the population subset being investigated. The estimated annual incidence rate was 12.6/100 000 among persons from LTBCs and 214.1/100 000 among immigrants from HTBCs. The estimated annual tuberculosis rate in the Turin metropolitan area (32.1/100 000) was nearly three times higher than in the rest of the Piedmont Region (11/100 000). The estimated annual incidence of pulmonary tuberculosis (10.8/100 000) was twice that of non-pulmonary (5.0/100 000), as was that

Table 7.2 Capture-recapture estimation models (deviance, degrees of freedom, standard error, cases estimated, upper and lower 95% confidence intervals) for three sources, obtained with the CARE library[20]

Estimation model		Cases estimated (95% CI)	Estimated number of unknown cases	Deviance (SE)	df	AIC*	BIC†
Independence between sources	Independent	689 (678–706)	32	44.32 (7)	3	52.32	70.27
One dependency between two sources, no catchments heterogeneity	Physician notification and laboratory	704 (688–728)	47	27.60 (10)	2	37.60	60.04
	Hospital and laboratory	686 (675–703)	29	42.15 (7)	2	52.15	74.59
	Physician notification and hospital	701 (684–728)	44	40.63 (11)	2	50.63	73.07
Two dependencies between two sources, no catchments heterogeneity	Physician notification and hospital; hospital and laboratory	696 (680–724)	39	39.55 (11)	1	51.55	78.48
	Physician notification and hospital; physician notification and laboratory	977 (790–1424)	320	0.24 (150)	1	12.24	39.17
	Laboratory and hospital; physician notification and laboratory	701 (685–725)	44	26.80 (10)	1	38.80	65.73
Heterogeneity (heterogeneous probability of capture among individuals)	Symmetric heterogeneity between sources	726 (695–782)	69	273.08 (21)	4	279.08	292.54
	Quasi-symmetric heterogeneity between sources	721 (692–775)	64	36.70 (20)	2	46.70	69.14
	Partial-quasi-symmetric heterogeneity between physician notification–hospital AND hospital–laboratory sources	726 (695–781)	69	25.09 (21)	1	37.09	64.02
	Partial-quasi-symmetric heterogeneity between surveillance–hospital AND physician notification–laboratory sources	1019 (802–1559)	362	0.20 (178)	1	12.20	39.13
	Partial-quasi-symmetric heterogeneity between physician notification–laboratory AND hospital–laboratory sources	743 (702–823)	86	33.49 (30)	1	45.49	72.42
Full dependency between sources	Saturated	1005	138	0.00 (174)	0	14.00	45.41

CI = confidence interval; CARE = capture-recapture; SE = standard error; df = degrees of freedom; AIC = Akaike information criterion; BIC = Bayesian information criterion

Table 7.3 Capture-recapture estimates: cases estimated (95% confidence intervals [CIs]); estimated sensitivities of 'physician notification system' and estimated crude annual tuberculosis incidence

	Estimated unknown cases (95% CI)	Estimated total cases (95% CI)	Estimated sensitivity of 'physician notification system' (95% CI)	Estimated tuberculosis incidence in 2001* (95% CI)
Total	47 (31–71)	704 (688–728)	79.1% (76.5–80.1)	16.7 (16.3–17.3)
Low tuberculosis prevalence countries	36 (22–57)	522 (508–543)	76.8% (73.8–78.9)	12.6 (12.3–13.1)
High tuberculosis prevalence countries	9 (3–22)	180 (174–193)	86.1% (80.8–89.6)	214.1 (207.0–229.6)
Pulmonary tuberculosis	19 (10–35)	453 (444–469)	85.4% (82.5–85.4)	10.7 (10.5–11.1)
Extra-pulmonary tuberculosis	2 (1–9)	213 (212–220)	74.2% (71.6–74.5)	5.0 (5.0–5.2)
< 60 years	24 (14–34)	398 (388–417)	82.4% (78.6–84.5)	13.1 (12.8–13.8)
≥ 60years	23 (13–41)	306 (296–324)	74.8% (70.7–77.4)	25.8 (24.9–27.3)
Turin	7 (3–18)	278 (274–289)	92.1% (88.5–93.4)	32.1 (31.7–33.4)
Piedmont excluding Turin	29 (17–49)	360 (348–380)	76.7% (72.6–79.3)	10.8 (10.4–11.3)

* cases per 100 000 population

of cases in persons aged ≥ 60 years (25.8/100 000) when compared with younger persons (13.1/100 000).

Discussion

The main findings of this study are that in Piedmont the reported tuberculosis incidence rates are largely underestimated. Although Piedmont remains a low-prevalence area, the burden of tuberculosis is higher than was previously thought. Record-linkage considerably improved the estimated case-ascertainment to 93.3%. The capture-recapture estimate of underreporting of 21% is almost twice that of the WHO for Italy as a whole.[5] The

estimated crude annual incidence of tuberculosis (16.7/100 000) was about twice that of all Italy (8/100 000) and was also higher than reported for the Piedmont region (12/100 000).[5,7,21] Record-linkage with additional capture-recapture analysis is a valuable means for quantifying underreporting and can provide relatively accurate estimates of the annual incidence of tuberculosis in areas where multiple recording systems are available.[22]

The incidence estimates found are representative of low tuberculosis prevalence areas. The overall crude annual incidence rate is similar to those of neighbouring countries such as Austria, France and Switzerland, which range from 11 to 16/100 000.[7]

Inaccurate estimates of the annual incidence of tuberculosis, particularly among high-risk subsets of the population such as immigrants from HTBCs and urban dwellers, vitiate the implementation of appropriate prevention and control measures. Our analyses for different subsets of the population in this study confirmed that persons from HTBCs have a much higher risk of developing tuberculosis than the local population. A similar phenomenon has been reported among immigrants and asylum seekers elsewhere.[23] The estimated annual tuberculosis rate in the Turin urban area is 32.1/100 000, which is three times higher than the rate in the rest of the Piedmont Region. A comparable trend has been reported in other metropolitan areas of Europe, such as Amsterdam, London and Rotterdam.[23,24] These rates reflect larger risk groups for tuberculosis in the population of large cities, such as certain ethnic groups, illegal immigrants, homeless persons and drug addicts.[25] The high rate among the elderly (97% of whom originated from LTBCs) is typical of tuberculosis epidemiology in low-incidence countries. Few young people are infected and develop active disease, while older persons can experience endogenous reactivation of latent tuberculosis infection or of a previous episode of tuberculosis in the era before effective chemotherapy. Although HIV infection is a well-known predisposing factor for active tuberculosis,[26] it makes a minor contribution to the burden of tuberculosis in Piedmont, as 5% of all cases were found to be HIV-seropositive. This proportion is consistent with the estimates of TB-HIV in the WHO European Region.[27]

Our finding of a sensitivity of around 85% for the 'physician notification system', currently implemented in the Piedmont Region, for detecting tuberculosis in immigrants from HTBCs, pulmonary tuberculosis and confirmed cases (the most important groups in terms of tuberculosis transmission control) is encouraging. This might reflect heightened awareness among both clinicians and public health authorities about the notification and surveillance of potentially infectious tuberculosis. Our observation confirms that tuberculosis patients aged ≥ 60 years are at risk of under-detection (25% for 'physician notification system').[28,29] The sensitivity of hospital discharge records, ranging from 50.9% for persons from HTBCs to more than 90% for extrapulmonary tuberculosis cases, is presumably affected by a different need for hospitalisation of the two groups, the former preferably being managed as out-patients and the latter usually being admitted to hospital for diagnostic workup.

We have described how assessment of tuberculosis incidence and case detection of tuberculosis in areas where multiple recording systems are available, such as the Piedmont Region in Italy, can be improved considerably by record-linkage of different data sources, such as the 'physician notification system' and laboratories. Implementation

of routine independent reporting from laboratories should be enforced to reduce under-ascertainment and improve the quality of information on diagnostic practices and criteria. Subsequent capture-recapture analysis, despite its inherent methodological limitations,[30] can be used to estimate total tuberculosis incidence and the completeness of registration, thus contributing to more accurate surveillance of local tuberculosis epidemiology. A detailed subset analysis can further identify gaps in the surveillance system and indicate adequate corrective interventions, such as improving the education of health care providers about reporting requirements or modifying reporting procedures.

Acknowledgments

The authors would like to thank Prof B Terracini, Prof A Biggeri, Prof N Pearce and Dr C Sacerdote for helpful comments and support in the research. The study was supported by 'Ricerca Finalizzata' Regione Piemonte/CIPE. The capture-recapture methods were developed within the framework of the Special Project 'Oncology', Compagnia San Paolo FIRMS, and of a grant from the Italian Association for Cancer Research

References

1. Rieder HL, Watson JM, Raviglione MC, Forssbohm M, Migliori GB, Schwoebel V, Leitch AG, Zellweger JP. Surveillance of tuberculosis in Europe. Working Group of the World Health Organization (WHO) and the European Region of the International Union Against Tuberculosis and Lung Disease (IUATLD) for uniform reporting on tuberculosis cases. *Eur Respir J* 1996; 9: 1097-104.
2. Migliori GB, Spanevello A, Ballardini L, Neri M, Gambarini C, Moro ML, Trnka L, Raviglione MC. Validation of the surveillance system for new cases of tuberculosis in a province of northern Italy. Varese Tuberculosis Study Group. *Eur Respir J* 1995; 8: 1252-8.
3. Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003; 3: 31.
4. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture-recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-3.
5. World Health Organization. *Global Tuberculosis Control: Surveillance, Planning, Financing. WHO Report 2003*. Geneva: WHO, 2003.
6. Moro ML, Malfait P, Salamina G, D'Amato S. [Tuberculosis in Italy: available data and open questions]. *Epidemiol Prev* 1999; 23: 27-36.
7. Buiatti E, Acciai S, Ragni P, Tortoli E, Barbieri A, Cravedi B, Santini MG. [The quantification of tuberculous disease in an Italian area and the estimation of underreporting by means of record-linkage]. *Epidemiol Prev* 1998; 22: 237-41.
8. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-64.
9. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. *Am J Epidemiol* 1995; 142: 1047-58.
10. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: Applications in human diseases. International Working Group for Disease Monitoring and Forecasting. *Am J Epidemiol* 1995; 142: 1059-68.
11. Bruno G, LaPorte RE, Merletti F, Biggeri A, McCarty D, Pagano G. National diabetes programs. Application of capture-recapture to count diabetes? *Diabetes Care* 1994; 17: 548-56.
12. Tilling K, Sterne JA, Wolfe CD. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol* 2001; 30: 1351-9.
13. Pezzotti P, Piovesan C, Michieletto F, Zanella F, Rezza G, Gallo G. Estimating the cumulative number of human immunodeficiency virus diagnoses by cross-linking from four different sources. *Int J Epidemiol* 2003; 32: 778-83.
14. Van Hest NA, Smit F, Verhave JP. Underreporting of malaria incidence in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-7.

15. Iglesias Gozalo MJ, Rabanaque Hernandez MJ, Gomez Lopez LI. [Tuberculosis in the Zaragoza province. Estimation by means of the capture-recapture method]. *Rev Clin Esp* 2002; 202: 249-54.
16. Mayoral Cortes JM, Garcia Fernandez M, Varela Santos MC, Fernandez Merino JC, Garcia Leon J, Herrera Guibert D, Martinez Navarro F. Incidence of pulmonary tuberculosis and HIV coinfection in the province of Seville, Spain, 1998. *Eur J Epidemiol* 2001; 17: 737-42.
17. Decreto Del Presidente Del Consiglio Dei Ministri. "Popolazione legale della Repubblica in base al censimento del 21 ottobre 2001". *Gazzetta Ufficiale Serie Generale* 2003; N. 81(Suppl. Ordinario n.54).
18. EuroTB (InVS/KNCV) and the national coordinators for tuberculosis surveillance in the WHO European Region. *Surveillance of tuberculosis in Europe. Report on tuberculosis cases notified in 2001*. Paris: EuroTB, 2003.
19. Fienberg S. The multiple-recapture census for closed populations and the 2k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
20. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med* 2001; 20: 3123-57.
21. Ministero della Salute. *Bollettino Epidemiologico-Dati Definitivi*. Rome: Ministero della Salute, 2001.
22. Ismail AA, Beeching NJ, Gill GV, Bellis MA. How many data sources are needed to determine diabetes prevalence by capture-recapture? *Int J Epidemiol* 2000; 29: 536-41.
23. KNVC Tuberculosis Foundation. *Index Tuberculosis 2001-2002*. The Hague: KNVC Tuberculosis Foundation, 2005.
24. Tuberculosis Section. Communicable Disease Surveillance Centre, Health Protection Agency. *Annual report on tuberculosis cases reported in 2001 in England, Wales and Northern Ireland*. London: Health Protection Agency, 2004
25. Felton CP Ford JG. Tuberculosis in the inner city. In: Reichman L, Hershfield E, eds. *Tuberculosis - A comprehensive international approach*. New York: M Dekker, 1993: pp. 483-98.
26. Zumla A, Malon P, Henderson J, Grange JM. Impact of HIV infection on tuberculosis. *Postgrad Med J* 2000; 76: 259-68.
27. Corbett EL, Watt CJ, Walker N, Maher D, Williams BG, Raviglione MC, Dye C. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* 2003; 163: 1009-21.
28. Baussano I, Cazzadori A, Scardigli A, Concia E. Clinical and demographic aspects of extrathoracic tuberculosis: experience of an Italian university hospital. *Int J Tuberc Lung Dis* 2004; 8: 486-92.
29. Centers for Disease Control and Prevention. Prevention and control of tuberculosis in facilities providing long-term care to the elderly. Recommendations of the Advisory Committee for Elimination of Tuberculosis. *MMWR Recomm Rep* 1990; 39(RR-10): 7-13.
30. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1996; 25: 474-8.

8

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 Department of Tuberculosis Control, Municipal Public Health Service Rotterdam Area, Rotterdam
2 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam the Netherlands
3 Respiratory Diseases Department, Centre for Infections, Health Protection Agency, London
4 Statistics, Modelling and Bioinformatics Department, Centre for Infections, Health Protection Agency, London United Kingdom
5 Department of Tuberculosis, Legionellosis and Imported Diseases, Institut de Veille Sanitaire, Saint-Maurice, France

Submitted

Abstract

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

Introduction

Since 1987, a rise in notifications of tuberculosis has been observed in England.[1] This increase is believed to be real, reflecting an increase in diagnoses of tuberculosis, rather than an artefact due to improved reporting.[2] Nevertheless, between 7% and 27% of cases of tuberculosis have been estimated not to be notified in the United Kingdom (UK).[3] In 1999, a revised national routine surveillance system for tuberculosis, Enhanced Tuberculosis Surveillance, 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 tuberculosis 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 reported in Enhanced Tuberculosis Surveillance 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 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 UK.[14-16] The same methodology has been applied to tuberculosis surveillance in studies in both the UK and elsewhere.[17-20]

Methods

Case definition and data sources

For the purpose of estimating the number of unobserved tuberculosis cases we defined as eligible for inclusion those active tuberculosis 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. Tuberculosis cases notified through Enhanced Tuberculosis Surveillance (Notification).
2. Cases with *Mycobacterium tuberculosis* complex isolates reported to MycobNet (Laboratory).
3. Cases admitted to National Health Service hospitals with a first or secondary hospital discharge code of tuberculosis (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 one 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 myco-

bacteriological confirmation, records three months before and 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 tuberculosis 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 tuberculosis 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 tuberculosis or MOTT infection Notification cases not known to Laboratory were linked with Treatment Outcome Monitoring data, containing data on Notification cases with a final diagnosis other than tuberculosis. At the time of this study Treatment Outcome Monitoring 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 tuberculosis identified a considerable proportion of remaining false-positives among unlinked Hospital cases after examining individual patients' medical files.[17,19] That 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 tuberculosis episode, rank number of tuberculosis diagnosis (14 possible positions) and 37 different ICD-10 tuberculosis diagnosis codes) and the incidence of Hospital records linked with Notification and/or Laboratory to estimate the number of true tuberculosis cases among unlinked records, under the assumption that all linked Hospital cases are true tuberculosis cases and unlinked Hospital cases are a mixture of true and false-positive tuberculosis cases. The best-fitting logistic regression model calculates for every Hospital case the predicted Bernoulli parameter p (reflecting the probability of being a true tuberculosis patient) from the covariates. Linked and unlinked Hospital cases have characteristic frequency distributions of values p as "signatures". After standardisation we used these signature curves to separate the mixture of unlinked Hospital cases, assuming the subpopulation of true tuberculosis cases has a similar signature curve to linked Hospital cases and the false-positive tuberculosis cases have a different signature curve (population mixture model available from the authors). The corrected annual number of true tuberculosis cases only known to Hospital was calculated using the formula:

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

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 tuberculosis cases by logistic regression respectively, $Prop_{true}$ the estimated annual proportion of true tuberculosis cases by logistic regression and $Prop_{MOTT}$ the projected annual proportion of MOTT infection cases.

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

Capture-recapture analysis

The annual and total number of unobserved tuberculosis 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 tuberculosis 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] Estimated source-specific coverage rates were defined as the number of tuberculosis cases in each data source divided by the estimated number of tuberculosis cases by capture-recapture analysis, expressed as a percentage.

Results

Table 8.1 shows the 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". The proportion of excluded records is small for all three tuberculosis data sources and consistent over the years examined.

Table 8.1 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%)

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 or more by the record-linkage software (from a maximum of 4000 points) were accepted as true links.

Table 8.2 shows the number, proportion and distribution of tuberculosis 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 Treatment Outcome Monitoring and Notification data sources for the year 2001 identified 4.1% of cases only known to Notification and 4.1% of cases known to Notification and Hospital with a final diagnosis of not tuberculosis or MOTT infection. Record-linkage between Hospital records and the MOTT database for the year 2003 identified 3.8% of Hospital cases as having MOTT infection. The population mixture model gave a range of the proportion of true tuberculosis cases only known 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 only known to Notification or Laboratory was reducing and the number of Notification cases linked to Laboratory or Laboratory and Hospital was increasing. Of all 28 678 tuberculosis cases included in this study, 2990 (10.4%) were identified in the Laboratory data source with a positive culture for *M. tuberculosis* but not notified.

Table 8.3 shows the annual and overall observed number of tuberculosis cases after record-linkage and correction for false-positive records. The overall observed

Table 8.2 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 8.3 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 hospitalised tuberculosis cases in England between 1999 and 2002.

	Observed tuberculosis cases*	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)

* Case-ascertaiment; † UI: uncertainty interval

source-specific coverage rates of notified, culture-confirmed and hospitalised tuberculosis 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 8.4 shows the annual and overall estimated number of unobserved and total tuberculosis cases after capture-recapture analysis. For all estimates the saturated log-linear model was preferred based on the Akaike Information Criterion.[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 hospitalised tuberculosis cases were 56.2%, 36.2% and 27.7% respectively. Overall estimated under-notification was 43.8%. The number of unobserved tuberculosis cases reduced 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 approximated 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 over-estimate of the number of unobserved cases of 1.5% (model available from the authors).

Discussion

Main findings

This study shows that record-linkage of tuberculosis data sources and cross-validation with additional tuberculosis-related datasets improves data accuracy as well as case-ascertainment. For large tuberculosis data sources sophisticated record-linkage software is required and a population mixture model to estimate the proportion of false-positive tuberculosis cases among unlinked hospital cases. Since the introduction of Enhanced Tuberculosis Surveillance the annual observed completeness of notification has increased. Still 10.4% of the observed tuberculosis cases in this study were laboratory-confirmed but not notified. The overall observed under-notification of 15.9% is consistent with prior reports. The 43.8% overall under-notification estimated by a saturated log-linear capture-recapture model is highly inconsistent with prior reports and the validity needs further examination.[3,17]

Under-notification

Increasing completeness of Notification could be influenced by improved data accuracy and record-linkage over the years. An overall observed under-notification of 15.9% suggests that in England approximately 1100 tuberculosis patients may not be not notified annually of which the majority (2990/4534) is culture-confirmed, representing 10.4% of all tuberculosis cases. This reflects the most serious public health aspect of under-notification as culture-confirmed tuberculosis cases are assumed true cases and are potentially infectious. Failure to notify laboratory-confirmed cases jeopardises control measures, including contact-tracing. Capture-recapture studies in Italy and the Netherlands show proportions of not notified culture-confirmed tuberculosis cases of

Table 8.4 Annual and overall estimated number of unobserved and total tuberculosis cases by saturated log-linear capture-recapture model in England between 1999 and 2002 (after using a proportion of 28% of true tuberculosis cases known to Hospital only in the corrections for false-positive cases).

Year	Estimated unobserved tuberculosis cases by the saturated log-linear capture-recapture model (95%ACI*)	Estimated total tuberculosis 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

Table 8.5 Annual and overall estimated number of unobserved and total tuberculosis cases by structural source model and truncated Poisson mixture model in England between 1999 and 2002 (after using a proportion of 28% of true tuberculosis cases known to Hospital only in the corrections for false-positive cases).

Year	Estimated unobserved tuberculosis cases by the structural source model (95%ACI*)	Estimated total tuberculosis cases by the structural source model (95%ACI)	Estimated unobserved tuberculosis cases by the truncated Poisson mixture model (95%ACI)	Estimated total tuberculosis 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

5.5% and 4.9% respectively.[19,20] The proportion of not notified culture-confirmed tuberculosis 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 tuberculosis 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 3000 points or more, and only 5.1% with such a score were not linked. This indicates that in only a minority of candidate-links an error of classification could 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 tuberculosis 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 tuberculosis data sources results in observed and estimated numbers of tuberculosis cases being too high. Lack of specificity of data sources used in capture-recapture studies as a limitation to the validity of this method is previously described.[22,23] Not all tuberculosis 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 tuberculosis occurs which will also reduce 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 tuberculosis incidence in the UK and elsewhere, indicating a plausible logistic regression model but expressing concern about the contribution of unscrutinised Hospital data sources to accurate estimates of tuberculosis 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 tuberculosis 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 can be identified and incorporated in the log-linear model. The three-way

interaction however, 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 tuberculosis as the opportunities for notification, culture-confirmation or hospitalisation are largely determined within a short period of time. However, this violation could result in overestimation of the number of patients.

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

Also likely is violation of the homogeneity assumption: age, location of disease and infectiousness, among others, can cause different probabilities of being observed in a tuberculosis 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 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 examined the data again 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 tuberculosis in the population. The annual and overall estimated number of unobserved and total tuberculosis cases is shown in Table 8.5 but the structural source model did not fit well. The number of unobserved tuberculosis cases is very high in 1999 but then reduces 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 tuberculosis cases to have extrapulmonary tuberculosis. Local under-notification of non-respiratory tuberculosis of 47% has been reported in the UK.[8] This possibly reflects health service organisation in the UK where extrapulmonary cases are less

likely to be managed by clinicians familiar with notification of infectious diseases. Apart from under-estimating the burden of tuberculosis, the implications for public health are limited as extrapulmonary tuberculosis 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 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 best fitting 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 given that the estimates of the truncated models could be preferable. Table 8.5 shows the annual and overall estimated numbers of unobserved and total tuberculosis cases. The estimated numbers of unobserved tuberculosis cases were low compared to the structural source model, especially in 1999. From 2000 onwards the estimates reduce every year. According to Zelterman's model estimated completeness of Notification was 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.

Hook and Regal state that "In no sense is there any proof or re-assurance 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 Enhanced Tuberculosis Surveillance, improves accuracy of surveillance data as well as completeness of case-ascertainment of tuberculosis. Hospital-derived data added a limited number of possible true tuberculosis patients. Since the introduction of Enhanced Tuberculosis Surveillance the annual observed completeness of notification has increased. This is most likely due to improvements in case-reporting combined with improved data collection and record-linkage. This study

shows that observed under-notification of tuberculosis 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 prior reports. Overall under-notification estimated by a saturated log-linear capture-recapture model is highly inconsistent with prior reports and could be an over-estimate 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 Enhanced Tuberculosis Surveillance, 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 tuberculosis cases. According to Zelterman's truncated model, in England and Wales the estimated completeness of the notification and laboratory records combined was 78.2%, 74.1%, 81.0% and 83.8% for the years 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.[9,20] Approaching and encouraging the clinicians treating the culture-positive tuberculosis cases not linked to Notification to notify these patients, considering the unlinked MycobNet cases as "pre-notifications", would increase the completeness of the notifications register.

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.

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? *J Public Health Med* 2002: 24: 136-37.
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? *Commun Dis Public Health* 1998; 1: 219-20.
5. Sheldon CD, King K, Cock H, Wilkinson P, Barnes NC. Notification of tuberculosis: how many cases are never reported. *Thorax* 1992; 47: 1015-8.
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. *Comm Dis Rep* 1995; 5: R137-140.
8. Mukerjee AK: Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data source *Commun Dis Public Health* 1999; 2: 143-4.
9. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
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: MIT-Press, 1975.

12. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-63.
13. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation II: Applications in human diseases. *Am J Epidemiol* 1995; 142: 1059-68.
14. Devine MJ, Bellis MA, Tocque K, Syed Q. Whooping cough surveillance in the north west of England. *Commun Dis Public Health* 1998; 1: 121-5.
15. Crowcroft NS, Andrews N, Rooney C, Brisson M, Miller E. Deaths from pertussis are underestimated in England. *Arch Dis Child* 2002; 86: 336-338.
16. Breen E, Ghebrehewet S, Regan M, Thomson AP. How complete and accurate is meningococcal disease notification? *Commun Dis Public Health* 2004; 7: 334-8.
17. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-3.
18. Cailhol J, Che D, Jarlier V, Decludt B, Robert J. Incidence of tuberculous meningitis in France, 2000: a capture-recapture analysis. *Int Tuberc Lung Dis* 2005; 9: 803-8.
19. Baussano I, Bugiani M, Gregori D, van Hest R, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung Dis* 2006; 10: 415-21.
20. Van Hest NA, Smit F, Baars HW, De Vries G, De Haas P, Westenend PJ, Nagelkerke NJ, Richardus JH. Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis? (Submitted).
21. Van Hest NA, Smit F, Verhave JP. Underreporting of malaria incidence in The Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-7.
22. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1999; 25: 474-8.
23. Borgdorff MW, Glynn JR, Vynnycky E. Using capture-recapture methods to study recent transmission of tuberculosis. *Int J Epidemiol* 2004; 33: 905-6.
24. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *Int J Epidemiol* 1994; 23: 1322-3.
25. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995; 6: 42-8.
26. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *J Clin Epidemiol* 1999; 52: 909-14.
27. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* 2000; 152: 771-9.
28. Jarvis SN, Lowe PL, Avery A, Levene S, Cormack RM. Children are not goldfish-mark-recapture techniques and their application to injury data. *Inj Prev* 2000; 6: 46-50.
29. Tilling K. Capture-recapture methods-useful or misleading? *Int J Epidemiol* 2001; 30: 12-4.
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. *Stat Med* 1998; 17: 69-74.
31. Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; 79: 543-53.
32. Hook EB, Regal RR. Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *Am J Epidemiol* 1982; 116: 168-76.
33. Smit F, Reinking D, Reijerse M. Estimating the number of people eligible for health service use. Evaluat Prog Plan 2002; 25: 101-5.
34. Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J Statistic Plan Inf* 1988; 18: 225-37.
35. Smit F, Toet J, Van der Heijden PJ. 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 1997: EMCDDA, pp 47-66.
36. Bohning D, Suppawattanabodee B, Kusolvitkul, W, Viwatwongkasem C. Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *Eur J Epidemiol* 2004; 19: 1075-83.
37. Hay G, Smit F. Estimating the number of hard drug users from needle-exchange data. *Addiction Res Theory* 2003; 11: 235-243.
38. Van Hest NA, Grant A, Smit F, Story A, Richardus JH. Estimating infectious disease incidence: validity of capture-recapture analysis and truncated models for incomplete count data. *Epidemiol Infect* 2007; Published on-line: 11 March 2007; doi:10.1017/ S0950268807008254.
39. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997;46:515-20

9

Estimating the coverage of tuberculosis screening among drug users and homeless persons with truncated models

N.A.H. VAN HEST[1,2], G. DE VRIES[1,2], F. SMIT[3,4] A.D. GRANT[5] and J.H. RICHARDUS[1,2]

1 Department of Infectious Disease Control, Municipal Public Health Service Rotterda-Rijnmond, Rotterdam
2 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam
3 Trimbos Institute of Mental Health and Addiction, Utrecht
4 Department of Clinical Psychology, Vrije Universiteit Amsterdam, Amsterdam
the Netherlands
5. Statistics, Modelling and Bioinformatics Department, Centre for Infections, Health Protection Agency, London
United Kingdom

Epidemiol Infect (in press)

Abstract

Truncated models are indirect methods to estimate the size of a hidden population which, in contrast to the capture-recapture method, can be used on a single information source. We estimated the coverage of a tuberculosis screening programme among illicit drug users and homeless persons with a mobile digital X-ray unit between 1 January 2003 and 31 December 2005 in Rotterdam, the Netherlands, using truncated models. The screening programme reached approximately two-third of the estimated target population at least once yearly. The intended coverage (at least two chest X-rays per person per year) was approximately 23%. We conclude that simple truncated models can be used relatively easily on available single source routine data to estimate the size of a population of illicit drug users and homeless persons. We assume the most likely overall bias in this study to be overestimation and therefore the coverage of the targeted mobile tuberculosis screening programme would be higher.

Introduction

The epidemiological pattern of tuberculosis in low incidence countries is changing, with an increasing number of tuberculosis patients living in urban areas.[1-3] This is due to overrepresentation of immigrants from countries with a high incidence of tuberculosis in large cities and to urban risk groups for tuberculosis such as illicit drug users and homeless persons.[4-6] Conventional tuberculosis control methods such as contact-tracing and preventive treatment are inadequate among marginalised care-avoiders.[5,7,8] As an alternative radiological screening programmes for illicit drug users and homeless persons have been recommended in European cities.[9-12]

Tuberculosis re-emerged among illicit drug users and homeless persons in Rotterdam (population approximately 600 000) in 2001, after periodic radiological screening was discontinued in 1996. In response, a periodic radiological screening programme was re-introduced in May 2002, using a mobile digital X-ray unit (MDXU) and visiting day and night shelters and hostels for homeless persons, methadone dispensing centres and safe drug consumption rooms for opiate users, as well as the street prostitution zone in Rotterdam. The programme aimed to screen clients of these facilities and services bi-annual.[5,13]

For priority setting, service planning and resource allocation it is necessary to know the number of persons in a targeted group. This number can also be used to assess the coverage of an intervention.[14] Often direct (enumeration) techniques are not feasible to estimate the size of hidden populations and indirect techniques have to be used. One such indirect technique, capture-recapture analysis,[15-17] has been used to estimate the size of hidden populations, including illicit drug users[18,19] and homeless persons.[20,21] However, capture-recapture analysis preferably needs at least three linked data sources, which are not always available for hidden populations. As an alternative, truncated models are described in the literature.[17,22,23] Contrary to conventional capture-recapture analysis, truncated models use frequency data from a single source of information. These models have been applied to estimate the size of hidden populations such as criminals,[24,25] illegal residents,[26] and illicit drug users and homeless persons.[14,27-30]

The objective of this study is to estimate the coverage of a mobile tuberculosis screening programme among illicit drug users and homeless persons in Rotterdam, using simple truncated models.

Methods

Ethics committee approval was not required for this study.

Study design, participants and study years

Participants in this descriptive study are individuals using the services of shelters and hostels for homeless persons, methadone dispensing centres or safe drug consumption rooms for opiate users, or working in the street prostitution zone in Rotterdam, with at least one chest X-ray taken in the MDXU of the mobile tuberculosis screening programme between 1 January 2003 and 31 December 2005. Because 2002 was an

incomplete year of screening and not all facilities were visited twice by the MDXU these data were excluded. A proportion of individuals in the target group use multiple facilities and their chest X-ray can be taken at different locations, sometimes more than twice yearly. Chest X-rays were read by public health tuberculosis physicians on location or within a few working days at the Public Health Service.

Data collection and validation

Data on participants of the MDXU screening programme, such as name, date of birth, sex, date of chest X-ray and chest X-ray result, are routinely entered into the electronic Client Information System of the Tuberculosis Control Section of the Municipal Public Health Service Rotterdam-Rijnmond, using a unique personal identification number. To avoid misclassification of individuals due to clerical errors such as misspelling of names or typing errors, all names and dates of birth of the participants were double-checked in the Client Information System during data entry. Since 2005 the Universal Mobile Telecommunications System (UMTS) provides wireless connection between the MDXU and the Client Information System facilitating checking personal data of participants on location. The number of individuals participating in the tuberculosis screening programme and the frequency of their visits per year and for the total study period were extracted from the Client Information System.

Truncated models

The number of illicit drug users and homeless persons in the target group for the mobile tuberculosis screening programme, and hence the coverage of the programme, was estimated through simple truncated models. Although their results are expected to be similar as two examples we used Zelterman's truncated Poisson mixture model and Chao's truncated heterogeneity model, which can be applied to frequency counts of observations of individuals in a single register.[31-33] They aim to estimate the number of unobserved persons in the (truncated) zero-frequency class based upon information of the lower observed frequency classes, assuming a specific truncated distribution of the observed data, e.g. Poisson, binomial or a mixture.[17,31-34] Observed frequency distributions may not be strictly Poisson and to relax this assumption Zelterman and Chao based their models on a Poisson mixture distribution. This allows greater flexibility and applicability on real life data because the models explicitly cater for departures from the strict Poisson assumption. Zelterman's Poisson mixture model of the estimated total population size, est(N), is given by

$$\text{est}(N) = \text{obs}(N)/[1 - \exp(-2f_2/f_1)]$$

and Chao's heterogeneity model by

$$\text{est}(N) = \text{obs}(N) + (f_1)^2/2f_2$$

where f_1 denotes the number of persons falling in the first frequency class, f_2 denotes the number of persons falling in the second frequency class, obs(N) denotes the number of all observed individuals and exp is the exponential.

The simple truncated models do not need statistical packages and have performed well when compared to log-linear capture-recapture analysis.[35] They

supposedly perform well even when data are sparse. Frequency data are less sensitive to privacy regulations. The truncated models of Zelterman and Chao were previously used to estimate the number of problematic illicit drug users in Rotterdam and detailed conceptual aspects of these models have been described.[14,28,30] An overview of a range of truncated models given elsewhere.[22] The underlying assumptions and limitations of truncated models will be discussed later.

Coverage

The coverage is defined as the number of individuals screened at least once per year (obs(*N*) or the annual case-ascertainment) divided by the estimated annual number of illicit drug users and homeless persons in the target group for periodic tuberculosis screening (est(*N*)), expressed as a percentage (obs(*N*)/(est(*N*))x100). This definition is different from the use of the word coverage by Chao in her heterogeneity model article,[31] which is related to the proportion of times that the confidence interval includes the true number of cases in a simulation study, or another well-known publication of Chao, in which it is related to a measure to quantify the source overlap information.[36]

Results

Between 1 January 2003 and 31 December 2005 a total of 7075 chest X-rays were made of 3034 individuals. Table 9.1 shows the total number of screened individuals per frequency class and number of chest X-rays taken. Nearly half of the individuals screened (45.6%) entered the programme only once.

Table 9.2 shows the annual number of screened individuals, people not previously screened and number of X-rays taken, per frequency class and in total. The annual number of individuals screened gradually decreased over the years. The annual number of people not previously screened strongly decreased but in 2004 and 2005 still a considerable number of these persons enter the programme.

Table 9.1 Total number of screened individuals per frequency class and number of chest X-rays taken in the mobile radiological tuberculosis screening programme among illicit drug users and homeless persons in 2003 - 2005 in Rotterdam.

Frequency class	Number of individuals	Percentage	Number of chest X-rays
1x	1384	45.6%	1384
2x	585	19.3%	1170
3x	397	13.1%	1191
4x	267	8.8%	1068
5x	218	7.2%	1090
> 6x	183	6.0%	1172
Total	3034	100%	7075

Table 9.2 Annual number of individuals screened, people not previously screened and number of X-rays taken, per frequency class and in total in the mobile radiological tuberculosis screening programme among illicit drug users and homeless persons in 2003-2005 in Rotterdam.

	2003			2004*			2005*		
Frequency class	Number of individuals screened	People not previously screened (%)	Number of chest X-rays	Number of individuals screened	People not previously screened (%)	Number of chest X-rays	Number of individuals screened	People not previously screened (%)	Number of chest X-rays
1x	1162	1162 (100%)	1162	1058	594 (56%)	1058	997	405 (41%)	997
2x	555	555 (100%)	1110	597	144 (24%)	1194	489	56 (11%)	978
≥ 3x	107	107 (100%)	333	57	8 (14%)	179	21	3 (14%)	64
total	1824	1824 (100%)	2605	1712	746 (44%)	2431	1507	464 (31%)	2039

*corrected for screening of a large shelter in January 2005 planned for December 2004.

Table 9.3 Annual number of observed and estimated individuals and coverage of the mobile radiological tuberculosis screening programme among illicit drug users and homeless persons in 2003 - 2005 in Rotterdam.

	Zelterman's Poisson mixture model[35]				Chao's heterogeneity model[34,36]			
Year	Obs*(N)**	Est*(N)*†	CI‡ (95%)	% covered	Obs*(N)*	Est*(N)*	CI (95%)	% covered
2003	1824	2964	(2803-3152)	(62%)	1824	3040	(2868-3241)	(60%)
2004	1712	2531	(2411-2671)	(68%)	1712	2649	(2512-2810)	(65%)
2005	1507	2411	(2274-2572)	(63%)	1507	2523	(2369-2705)	(60%)

Obs(N)*: Number of individuals observed; †Est*(N)*: Number of individuals estimated; ‡CI: Confidence Interval.

The annual number of individuals in the first frequency class (seen once), second frequency class (seen twice) and total number of individuals screened respectively represent f_1, f_2 and obs*(N)* in the formula of the truncated models.

Table 9.3 shows the annual observed and estimated number of illicit drug users and homeless persons in the target group for periodic tuberculosis screening for the two truncated models, as well as the estimated coverage of the mobile tuberculosis screening programme. The estimates of Chao's model are slightly higher but in the same range as Zelterman's model. The radiological mobile targeted tuberculosis screening programme reaches approximately 63% of the estimated target population at least once per year. The intended coverage of the screening programme (at least two chest X-rays per person per year) is approximately 22%, 25% and 21% in 2003, 2004 and 2005 respectively.

Discussion

Main findings

This study demonstrates that truncated models can be used relatively easily on available single source routine data to estimate the size of a hidden population of illicit drug users and homeless persons. Our results show that a radiological mobile targeted tuberculosis screening programme among illicit drug users and homeless persons in Rotterdam reaches approximately two-third of the estimated target population at least once per year. Between 21% and 25% of the estimated target population meets the objective of the programme and has two or more chest X-rays taken per year.

Limitations

As with capture-recapture analysis, the validity of the estimates of truncated models depends on the possible violation of the underlying assumptions. These assumptions are perfect identification (i.e. no misclassification of the number of visits of one client), a closed population (i.e. no in-migration or out-migration in the time period studied), ideally but not necessarily a homogeneous population (i.e. no subgroups with markedly different probabilities to be observed and re-observed) and a constant probability of being observed (i.e. there should be no individual behavioural response and the probability of being re-observed should not be influenced by the experience of a previous visit) and, as explained earlier in the Methods, a specific truncated distribution of the observed data.[14,30]

Perfect identification assumption

In this programme individuals were assigned unique identification numbers in the Client Information System and personal identifiers were double-checked upon data entry to avoid misclassification. The staff of the facilities visited assisted the programme by providing a list of names and dates of birth of clients eligible for screening. Most clients had personal identification cards which were checked at screening. Social workers from the services furthermore assisted on the day of screening which also reduced the possibility to misclassify individuals. Violation of the perfect record-linkage assumption is therefore considered minimal.

Closed population assumption

To reduce bias as a result of violation of the closed population assumption we divided the study in one-year periods. The MDXU visits each location for one day twice a year. This limits the opportunity for passers-by and short-term clients to be observed. Table 9.1 and Table 9.2 however show that every year a substantial number of people not previously screened enter the programme. These can be individuals belonging to the target group but not yet captured by the screening programme, individuals not belonging to the target group or individuals that recently joined the target group. Influx of the last two categories will result in annual estimates of the target population of long-term illicit drug users and homeless persons being too high and hence the estimate of the screening programme coverage being too low.

Homogeneity assumption

Some problematic illicit drug users and homeless persons, such as cocaine users or persistent rough sleepers, will never be reached. Their probability to attend the tuberculosis screening programme is zero because they never utilise the facilities and services. This group is not included in the truncated model estimate.[28]

We cannot exclude individuals entering the screening programme, e.g. among individuals entering the programme only once, that do not belong to the group of long-term illicit drug user and homeless persons. In a previous conventional log-linear capture-recapture estimation of the number of clients of a methadone maintenance programme it was demonstrated that differences in capture-probabilities of the population of interest, problematic drug users, and the sampled population, also including non-problematic drug users, could considerably overestimate the size of the population of interest.[19]

We cannot exclude heterogeneity among individuals belonging to the target group entering the screening programme but the opportunities to participate in the screening (opting-out strategy) or not to participate (not attending the facility or service on the day of screening) are assumed to be largely similar for the majority. The truncated models are arguably more robust to violation of the homogeneity assumption because they are partly based upon the lower frequency classes, assumed to have more resemblance to the zero frequency class. The relative insensitivity to violation of the homogeneity assumption of Zelterman's and Chao's model is also supported mathematically and through simulation studies.[14,22,32] However, in the presence of heterogeneity they can underestimate the population.

An alternative approach to estimate a heterogeneous population would be to use a population mixture model. Such a model (for the data in Table 9.1) regards the eligible population for each visit as a mixture of "local clients at the facilities", having six opportunities to be observed and "roaming clients" from other facilities, visiting more places than their own facility and can be captured at other facilities by the MDXU as well. They have possibly more than a total of six opportunities to be observed. For each visit the capture of the local clients could be modelled as Binomial (6, p1) and the capture of roaming clients as Poisson (lambda), where lambda is probably less than 6 times p1. However, in our population of homeless persons and illicit drug users a clear distinction

between local and roaming clients is arbitrary as many clients use multiple services, e.g. methadone dispensing centres due to their addiction and day en night care facilities due to their homelessness, and their need for specific services may change over time. Furthermore, we have not considered such a population mixture model or E-M algorithm because, although more accurate, their complexity disagrees with the appealing ease of use of the simple truncated models. For the purpose of our study more exact but complex to calculate estimates were subsidiary to the simplicity of a method which should be close enough. As described for capture-recapture analysis simple truncated models are useful under certain circumstances, e.g. when the likely direction of the bias caused by violation of the underlying assumptions can be predicted and plausible lower and upper boundaries of the prevalence or incidence of a disease or the coverage of a community health care intervention can be estimated.[17,37,38]

Constant (re)observation probability assumption

For the majority of the individuals in the target group of the mobile tuberculosis screening programme the facilities and services where screening took place are providing important needs, namely methadone and shelter. These needs are probably constant over time and create a considerable probability of attending the services. Frequent users have the highest risk of tuberculosis but are also most likely to be screened. Although incentives, such as chocolate bars and soft drinks, were given to participants at some locations, it is unlikely that this creates an important positive behavioural response to participate again. This also applies to clients with radiographic abnormalities inconsistent with tuberculosis as they are referred to a chest-physician in one of the general hospitals in Rotterdam where further analysis and follow-up is performed. The opting-out strategy and (strong) persuasion by the staff of the social and medical services to participate prevents a negative behavioural response. The pressure particular institutions put on their clients to participate in the screening programme is considered relatively constant on each screening day. The coverage of the screening programme will never be perfect as each year a proportion of the target group will temporarily have a low or zero probability to attend, e.g. due to admission in a rehabilitation clinic or prison sentence. Finally, elsewhere is has been explained that the probability of being observed does not have to be constant as long as a capture or non-capture does not influence a possible change in probability.[26]

Poisson distribution of the observed data assumption

Zelterman and Chao based their model on a Poisson mixture distribution, catering for departures from the strict Poisson assumption. We have examined whether the Zelterman model used tolerates the departures from the Poisson distribution observed in our data. We have performed negative Binomial regression, with number of times screened as the covariate and number of individuals as the outcome, on the Table 9.1 data (counting > 6 as 6). The variance of the data is larger than that of a Poisson distribution. This overdispersion is statistically significant ($P = 0.11$), but small ($alpha = 0.024$), and so does not invalidate the use of Zelterman's estimator.[32] Therefore it seems reasonable to use this simple model in the context of our study as explained earlier.

A further limitation is that persons in the target group could have indicated on the day of screening that recently a chest X-ray was taken in the MDXU, a general hospital, upon detention in prison or at the Tuberculosis Control Section upon referral, exempting them from the screening exercise. This information, together with improved experience, better co-ordination and UMTS access over the years, would prevent some clients from being recorded twice or more than twice yearly in the screening programme, as is reflected in Table 9.2, leading to overestimation, but we assume this effect to be limited.

Cross-validation of the estimates of the target group

The number of problematic illicit drug users in Rotterdam, already including many homeless persons, was most recently estimated in 2003 with two-source capture-recapture analysis, using a similar case-definition, which observed and estimated 1910 and 2856 clients respectively.[39] These numbers are similar to our results in 2003.

Alternative simple truncated models

Although we used truncated Poisson mixture models, an alternative is to use a truncated binomial model such as est(N) = obs(N) + $(f_1)^2/4f_2$. This model, close to Chao's model, estimates a lower number of 2432, 2181 and 2015 illicit drug users and homeless persons in 2003, 2004 and 2005 respectively, resulting in a slightly higher estimated coverage of the screening programme.

Conclusion

Although the limitations of the single-source truncated models should be appreciated and bias cannot be excluded, alternative methods for estimating the number of illicit drug users and homeless persons have their own restrictions. Conventional two-source and three-source capture-recapture analysis have similar underlying assumptions and hence limitations, and for hidden populations sufficient adequate registers for record-linkage may not be available. Compared to alternative estimators the ease of use of the truncated models is appealing. We could extract, check and prepare the required data from an existing routine dataset in two days and calculate the point estimates on a pocket calculator. We assume the most likely overall bias in this study to be overestimation and therefore the coverage of the targeted mobile tuberculosis screening programme among problematic illicit drug users and homeless persons in Rotterdam would be higher than the 63% one chest X-ray per year and 21-25% for at least two chest X-rays per year, especially among those with the highest risk.

Acknowledgements

We thank Monica Straal, software application manager of the Tuberculosis Control Section for assistance in preparing the final data file.

References

1. Fujiwara PI, Frieden TR. Tuberculosis epidemiology and control in the inner city. In: Rom W, Garay S, eds. *Tuberculosis*. New York: Little, Brown & Co, 1996: pp 99-112.
2. Moore-Gillon JC. Tuberculosis and poverty in the developed world. In: Davis PD, ed. *Clinical tuberculosis*. London: Chapman and Hall, 1998: pp 383-396.
3. Hayward AC, Darton T, Van-Tam JN, Watson JM, Coker R, Schwoebel V. Epidemiology and control of tuberculosis in Western European cities. *Int J Tuberc Lung Dis* 2003; 7: 751-7.
4. Valin N, Antoun F, Chouaid C, Renard M, Dautzenberg B, Lalande V, Avache B, Morin P, Sougakoff W, Thiolet JM, Truffot-Pernot C, Jarlier V, Decludt B. Outbreak of tuberculosis in a migrants' shelter, Paris, France, 2002. *Int J Tuberc Lung Dis* 2005; 9: 528-33.
5. De Vries G, Van Hest R. From contact investigation to tuberculosis screening of drug addicts and homeless persons in Rotterdam. *Eur J Public Health* 2006; 16: 133-6
6. Story A, Van Hest R, Hayward A. Tuberculosis and social exclusion. *BMJ* 2006; 333: 57-8.
7. Barnes PF, Yang Z, Preston-Martin S, Pogoda JM, Jones BE, Otaya M, Eisenach KD, Knowles L, Harvey S, Cave MD. Patterns of tuberculosis transmission in Central Los Angeles. *JAMA* 1997; 278: 1159-63.
8. Bock NN, Metzger BS, Tapia JR, Blumberg HM. A tuberculin screening programme and isoniazid preventive therapy programme in an inner-city population. *Am J Respir Crit Care Med* 1999; 159: 295-300.
9. Kumar D, Citron KM, Leese J, Watson JM. Tuberculosis among the homeless at a temporary shelter in London: report of a chest x ray screening programme. *J Epidemiol Community Health* 1995; 49: 629-33.
10. Southern A, Premaratne N, English M, Balazs J, O'Sullivan D. Tuberculosis among homeless people in London: an effective model of screening and treatment. *Int J Tuberc Lung Dis* 1999; 3: 1001-8.
11. Solsona J, Cayla JA, Nadal J, Bedia M, Mata C, Brau J, Maldonado J, Mila C, Alcaide J, Altet N, Galdos-Tanguis H. Screening for tuberculosis upon admission to shelters and free-meal services. *Eur J Epidemiol* 2001; 17: 123-8.
12. De Vries G, Van Hest NA, Šebek MM. Active tuberculosis screening with mobile digital X-ray units among drug addicts and homeless people in Rotterdam. In: *Abstracts of the 3rd Congress of European Region International Union against Tuberculosis and Lung Disease*. Moscow, Russia; 2004 June 22-26; Abstract 15. Moscow, Russia: Russian Respiratory Society, 2004.
13. De Vries G, Van Hest R, Richardus JH. Impact of mobile radiographic screening on tuberculosis among drug users and homeless persons. *Am J Respir Crit Care Med* 2007; Published on-line 5 April 2007: doi:10.1164/ rccm.200612-1877OC
14. Smit F, Reinking D, Reijerse M. Estimating the number of people eligible for health service use. *Eval Prog Plan* 2002; 25: 101-5.
15. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
16. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation II: Applications in human diseases. *Am J Epidemiol* 1995; 142: 1059-68.
17. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-63.
18. Hay G, McKeganey N. Estimating the prevalence of drug misuse in Dundee, Scotland: an application of capture-recapture methods. *J Epidemiol Community Health* 1996; 50: 469-72.
19. Buster MC, Van Brussel GH, Van den Brink W. Estimating the number of opiate users in Amsterdam by capture-recapture: the importance of case definition. *Eur J Epidemiol* 2001; 17: 935-42.
20. Fisher N, Turner SE, Pugh R, Taylor C. Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis. *BMJ* 1994; 308: 27-30.
21. Gurgel RQ, Da Fonseca JD, Neyra-Castaneda D, Gill GV, Cuevas LE. Capture-recapture to estimate the number of street children in a city in Brazil. *Arch Dis Child* 2004; 89: 222-4.
22. Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; 79: 543-53.
23. Van Hest NA, Grant AD, Smit F, Story A, Richardus JH.. Estimating infectious disease incidence: validity of capture-recapture analysis and truncated models for incomplete count data. *Epidemiol Infect* 2007; Published on-line: 11 March 2007; doi:10.1017/S0950268807008254.
24. Rossmo DK, Routledge R. Estimating the size of criminal populations. *J Quant Criminol* 1990; 6: 293-314.
25. Van der Heijden PG, Cruyff MJ, Van Houwelingen, H. Estimating the size of a criminal population from police registrations using the truncated Poisson regression model. *Stat Modelling* 2003; 3: 305-22.
26. Van der Heijden PG, Bustami R, Cruyff MJ, Engbersen G, Van Houwelingen HC Point and interval estimation of the truncated Poisson regression model. *Stat Modelling* 2003; 3: 305-22.
27. Hser YI. Population estimation of illicit drug users in Los Angeles County. *J Drug Issues* 1993; 23: 323-34.

28. Smit F, Toet J, Van der Heijden PJ. 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.
29. Bohning D, Suppawattanabodee B, Kusolvitkul, W, Viwatwongkasem C. Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *Eur J Epidemiol* 2004; 19: 1075-83.
30. Hay G, Smit F. Estimating the number of hard drug users from needle-exchange data. *Addiction Res Theory* 2003; 11: 235-43.
31. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; 43: 783-91.
32. Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J Statistic Plan Inf* 1988; 18: 225-37.
33. Chao A. Estimating animal abundance with capture frequency data. *J Wildl Manage* 1988; 52: 295-300.
34. Chao A. Estimating population size for sparse data in capture-recapture experiments. *Biometrics* 1989; 45: 427-38.
35. Hook EB, Regal RR. Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. Am J Epidemiol 1982; 116: 168-76.
36. Chao A, Tsay PK, Lin S-H, Shau W-Y, Chao D-Y. The applications of capture-recapture models to epidemiological data. *Statist Med* 2001; 20: 3123-57.
37. Hook EB, Regal RR. Capture-recapture methods. *Lancet* 1992; 339: 742.
38. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. *Am J Epidemiol* 1992; 135: 1060-7.
39. Biesma S, Snippe J, Bieleman B. [*Illicit drug users registered. Characteristics, population size and mobility of problematic illicit drug users in Rotterdam*]. Groningen: Intraval, 2004.

10

Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data

N.A.H. VAN HEST[1,2], A.D. GRANT[5], F. SMIT[3,4], A. STORY[6] and J.H. RICHARDUS[1,2]

1 Department of Infectious Disease Control, Municipal Public Health Service Rotterdam-Rijnmond, Rotterdam
2 Department of Public Health, Erasmus MC, University Medical Centre Rotterdam, Rotterdam
3 Trimbos Institute of Mental Health and Addiction, Utrecht
4 Department of Clinical Psychology, Vrije Universiteit Amsterdam, Amsterdam
the Netherlands
5 Statistics, Modelling and Bioinformatics Department, Centre for Infections, Health Protection Agency, London
6 Respiratory Diseases Department, Centre for Infections, Health Protection Agency, London
United Kingdom

Epidemiol Infect 2007; Published on-line: 11 March 2007; doi:10.1017/S0950268807008254

Abstract

Capture-recapture analysis has been used to evaluate infectious disease surveillance. Violation of the underlying assumptions can jeopardize the validity of the capture-recapture estimates and a tool is needed for cross-validation. We re-examined nineteen datasets of log-linear model capture-recapture studies on infectious disease incidence using three truncated models for incomplete count data as alternative population estimators. The truncated models yield comparable estimates to independent log-linear capture-recapture models and to parsimonious log-linear models when the number of patients is limited or the ratio between patients registered once and twice is between 0.5 and 1.5. Compared to saturated log-linear models the truncated models produce considerably lower and often more plausible estimates. We conclude that for estimating infectious disease incidence independent and parsimonious three-source log-linear capture–recapture models are preferable but truncated models can be used as a heuristic tool to identify possible failure in log-linear models, especially when saturated log-linear models are selected.

Introduction

Surveillance of infectious diseases is an essential part of public health. Mandatory notification is one of the mechanisms to carry out such surveillance but under-notification has been widely reported. For meaningful interpretation of the number of patients with infectious diseases the completeness of notification should be estimated. This can be done through a statistical technique called capture-recapture analysis. Based on certain assumptions, capture-recapture methods use information on the overlap of linked disease registers to estimate the number of patients unknown to all registers and thus the estimated total number of patients.[1] Completeness of notification can then be assessed relative to the estimated total number of patients. In biomedical sciences capture-recapture analysis is frequently used for estimating the number of accidents and injuries[2] and patients with mostly chronic diseases such as congenital deformities,[3] insulin-dependent diabetes mellitus,[4] cancer,[5] neurological conditions[6] or rheumatological diseases.[7] Less frequently it has been applied for evaluating infectious disease surveillance, especially when record-linkage is based on more than two registers.

The validity of capture-recapture estimates depends on possible violations of the underlying assumptions: cases can be uniquely identified (i.e. registers have a perfect positive predictive value), 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 (re)capture probabilities). In two-source capture-recapture methods one must also assume independence between registers (i.e. the probability of being observed in one register is not affected by being (or not being) observed in the other registers). In the three-source capture-recapture approach pair-wise dependencies, i.e. dependencies between two registers, can be identified and accounted for in a log-linear model.[1,8-11] The three-way (highest-order) interaction however, i.e. dependency between all three registers, cannot be incorporated in the model and its absence must be assumed.

In epidemiological studies violation to some degree of most of the underlying capture-recapture assumptions is unavoidable. This and other limitations of capture-recapture analysis are described elsewhere in more detail.[10,12-19] Infectious diseases carry an elevated risk that some capture-recapture analysis assumptions are violated. Especially absence of dependence between the available registers, including three-way interaction, and heterogeneity among the patients cannot be excluded and should be expected. Consequently, the validity of two-source and three-source capture-recapture studies requires critical scrutiny.

Sometimes it becomes evident that a capture-recapture model breaks down and produces erratic results. While performing three-source log-linear model capture-recapture studies on the completeness of notification of tuberculosis in the Netherlands[20] and England we were confronted with unexpected and unrealistic estimates of tuberculosis incidence, despite using well-described procedures for finding the best log-linear model.[21] In this context, solely relying on three-source capture-recapture analysis without any cross-validation seems to be inappropriate. We suggest that three-source capture-recapture analyses should be complemented by alternative methods to arrive at,

and cross-validate, estimates of population size. Alternative models related to capture-recapture analysis have been described and offer the opportunity to cross-validate outcomes. The aim of this study is to re-examine the data of published and current three-source log-linear model capture-recapture studies on infectious disease incidence with various truncated models for incomplete count data and describe the apparent agreement or discrepancy of the estimates.

Methods

Data sources

Data sources used were 19 datasets in 16 published or current three-source log-linear model capture-recapture studies on infectious disease incidence known to us.

Truncated population estimators

The data sources were re-examined with three alternative population estimators: a truncated binomial model, a truncated Poisson mixture model (Zelterman)[22] and a truncated Poisson heterogeneity model (Chao).[23,24] Out of the many possible methods we have chosen this combination of truncated models because they have been described as an alternative to capture-recapture methods,[10,25] can be used on the same data that is needed for the three-source log-linear model and are easy to apply.[26,27]

In epidemiology, truncated estimators are usually applied to frequency counts of observations of individuals in a single data source.[28] They aim to estimate the number of unobserved persons (falling in the zero-frequency class) based upon information on the number of times a person has been observed. Technically, one assumes a specific truncated distribution of the observed data, e.g. Poisson or binomial, and then extrapolates from the observed series to the unobserved number of people never seen.[10] Observed frequency distributions may not be strictly Poisson and to relax this assumption Zelterman based his model on a Poisson mixture distribution, allegedly allowing greater flexibility and applicability on real-life data.[28] Conceptual aspects of the Zelterman and Chao models have been discussed in some detail elsewhere.[27,29-31] The simple truncated estimators do not need statistical packages. In the social sciences truncated models have been employed to estimate the size of hidden populations such as criminals,[26,32] illegal residents[33] and illicit drug users and homeless persons.[27-29,34] To our knowledge, truncated estimators have not been used before to estimate the number of infectious disease patients.

As with capture-recapture analysis, the validity of the estimates of truncated models depends on the possible violation of the underlying assumptions. These assumptions are similar to the capture-recapture assumptions described earlier but in addition equiprobability (i.e. equal ascertainment probabilities of all registers) should be assumed when using multiple sources.[10] Some truncated models are arguably more robust to population heterogeneity because they are partly based upon the lower frequency classes, and the people rarely seen are assumed to have a greater resemblance with the people never seen. This relative insensitivity to violation of the homogeneity assumption

of some truncated estimators is supported mathematically and through simulation studies but they can occasionally underestimate the true population size in the presence of heterogeneity.[22,29]

Frequency counts

It is possible to extract frequency counts for the truncated models from multiple-source capture-recapture data, allowing us to use the reported data from the log-linear studies for the truncated models. The ratio between the number of patients registered once (f_1) and registered twice (f_2) plays an important role in the truncated models. When "1" represents being known to a register and "0" represents being unknown to a register, and three linked registers are used, frequency count f_1 is the sum of the cells 100, 010 and 001 in the 2 x 2 x 2 contingency table and frequency count f_2 corresponds to the sum of the cells 110, 101 and 011. Similarly, patients observed in all three registers, f_3, are denoted as 111. For all 19 datasets the number of patients in these seven cells are shown later. We use the f_1/f_2 ratio to examine a possible relationship between this ratio and the performance of truncated models vis-à-vis the log-linear models.

Results

Table 10.1 shows the various three-source log-linear model capture-recapture studies of infectious disease incidence and completeness of notification with the number of patients observed and their frequency counts, the objective of the study, the data sources used and the selected log-linear model. The studies involved eight infectious diseases and were performed at the local, regional or national level. One study collected data over a 4-months period, the other studies over 1- to 5-year periods. The observed number of patients varied from 33 to 28 678 persons. Notification, laboratory and hospital registers were the most conventional data sources used. The distribution of the patients over three linked registers in the various three-source capture-recapture studies of infectious diseases is shown in Table 10.2.

The log-linear and truncated model estimates with their respective confidence and prediction intervals are shown in Table 10.3, as well as the f_1/f_2 ratio among the observed patients and the coefficient of variation of the data source probabilities (see Discussion). The capture-recapture studies varied in estimated number of patients from 46 to 42 969. A second truncated Poisson estimator, Chao's bias-corrected homogeneity model, $\text{est}(N) = \text{obs}(N) + [(f_1^2-f_1)/(2(f_2-1))]$, gave similar estimates as Chao's heterogeneity model.[35] A second truncated binomial estimator, $\text{est}(N) = \text{obs}(N)/[1-(1/(1+f_2/f_1))^3]$, gave similar estimates as the truncated binomial model used (data not shown).

f_1/f_2 ratio

On the basis of the f_1/f_2 ratio the studies can be divided in four categories:

a. $f_1/f_2 < 0.5$ (dataset 7). In this study all estimates were similar but the number of observed patients was small.

b. $0.5 < f_1/f_2 < 1.5$ (datasets 1, 2, 6, 8-10, 13a, 13b, 14, 16a, 16b). In these studies the truncated binomial model and Zelterman's model gave similar results as the

Table 10.1 Overview of the various three-source capture-recapture studies of infectious diseases since 1997

Study number	Disease and number of patients observed	Objective	Data-sources	Selected capture-recapture model and interactions
1	Legionnaires' disease Obs(N) = 256; f_1 = 126; f_2 = 116; f_3 = 14	To estimate the level of underreporting of Legionnaires' disease and to evaluate the feasibility of a laboratory-based reporting system in France in 1995	1. National notification system 2. Reference laboratory database 3. Hospital laboratory survey	Independent model
2	HIV/AIDS Obs(N) = 173; f_1 = 65; f_2 = 75; f_3 = 33	To estimate the completeness of the prison AIDS register in Spain in 2000	1. Prison register of AIDS patients 2. Prison register of tuberculosis patients 3. Prison register of hospital admissions	Independent model
3a (1-4 yrs)	Pertussis Obs(N) = 435; f_1 = 375; f_2 = 56; f_3 = 4 Obs(N) = 420; f_1 = 376; f_2 = 42; f_3 = 2	To estimate undernotification of whooping cough in the north west of England, 1994-1996	1. Notification database from office for national statistics 2. Hospital admission data 3. Public health laboratory reports	Parsimonious model with one two-way interaction (notification * laboratory)
3b (>5 yrs)				
4	Salmonella infection Obs(N) = 608; f_1 = 520; f_2 = 68; f_3 = 20	To assess the number of foodborne Salmonella outbreaks in France in 1995	1. Mandatory public health notification 2. Mandatory veterinary notification 3. National Salmonella reference centre	Parsimonious model with one two-way interaction (public health notification * veterinary notification)
5	Pertussis Obs(N) = 33; f_1 = 19; f_2 = 12; f_3 = 2	To improve estimates of deaths from pertussis in England and to identify reasons for under ascertainment, 1994-1999	1. Hospital episode statistics 2. Enhanced laboratory pertussis surveillance 3. Office for national statistics mortality data	Parsimonious model with one two-way interaction (hospital statistics * national statistics)
6	Meningococcal meningitis Obs(N) = 53; f_1 = 9; f_2 = 14; f_3 = 30	To evaluate the exhaustiveness of three information sources on meningococcal disease in Tenerife, Spain, 1999-2001	1. Mandatory notifiable disease surveillance system 2. Laboratory survey 3. Hospital information database registry	Parsimonious model with one two-way interaction (hospital * laboratory)

Study number	Disease and number of patients observed	Objective	Data-sources	Selected capture-recapture model and interactions
7	Meningococcal meningitis Obs(N) = *81;* f_1 = 2; f_2 = 30; f_3 = 49	To assess the completeness of meningococcal disease in South Cheshire, UK, 1999-2001	1. Notification database 2. Laboratory reports 3. Hospital discharge codes database	Parsimonious model with one two-way interaction (hospital * laboratory)
8	Tuberculosis Obs(N) = 657; f_1 = 219; f_2 = 285; f_3 = 153	To assess completeness of the tuberculosis systems and estimation of under-reporting in the Piedmont Region of Italy in 2001	1. Physician notification system 2. Reference laboratory database 3. Hospital admission statistics	Parsimonious model with one two-way interaction (laboratory * notification)
9	Meningitis, bacterial Obs(N) = 199; f_1 = 64; f_2 = 59; f_3 = 76	To estimate the incidence of bacterial meningitis and to assess the quality of the surveillance systems in the Lazio Region of Italy, 1995-1996.	1. Mandatory notifiable disease surveillance system 2. Voluntary hospital laboratory-based surveillance system 3. Hospital discharge code registry	Parsimonious model with two two-way interactions (notification * hospital and notification * laboratory)
10	Malaria Obs(N) = 667; f_1 = 284; f_2 = 258; f_3 = 123	To estimate the completeness of notification of malaria by physicians and laboratories in the Netherlands in 1996	1. Passive national notification register 2. Active laboratory survey 3. National hospital admission registration	Parsimonious model with two two-way interactions (notification * hospital and laboratory * hospital)
11	Legionnaires' disease Obs(N) = 715; f_1 = 386; f_2 = 234; f_3 = 95	To evaluate improvements made to the mandatory notification system for Legionnaires' disease in France in 1998	1. National notification system 2. Reference laboratory database 3. Hospital laboratory survey	Parsimonious model with two two-way interactions (notification * reference laboratory and notification * hospital laboratory)
12	Hepatitis A Obs(N) = 271; f_1 =187; f_2 = 56; f_3 = 28	To estimate the number of individuals infected with hepatitis A during an outbreak in Taiwan in 1995.	1. Laboratory serological test records 2. Hospital reports 3. Epidemiological questionnaires	Saturated model with all interactions (including highest order interaction)
13a	Tuberculosis Obs(N) = 1499; f_1 = 472; f_2 = 639; f_3 = 388	To describe systematic examination and case-verification, record-linkage, capture-recapture analysis and assessment of the completeness of three tuberculosis registers in the Netherlands in 1998	1. Physician notification system 2. Reference laboratory database 3. Hospital admission statistics	Saturated model with all interactions (including highest order interaction)

Study number	Disease and number of patients observed	Objective	Data-sources	Selected capture-recapture model and interactions
13b	Obs(*N*) = 1441; f_1 = 376; f_2 = 677; f_3 = 388	To describe systematic examination and case-verification, record-linkage, capture-recapture analysis and assessment of the completeness of three tuberculosis registers in the Netherlands in 1998	1. Physician notification system 2. Reference laboratory database 3. Hospital admission statistics	Parsimonious model with two two-way interactions (laboratory * hospital and notification * laboratory)
14	Tuberculosis Obs(*N*) = 28 678; f_1 = 11 799; f_2 = 10 804; f_3 = 6075	To describe case-verification, record-linkage, capture-recapture analysis and assessment of completeness of three tuberculosis registers in England, 1999-2002	1. Notification database 2. Laboratory reports database 3. Hospital discharge codes database	Saturated model with all interactions (including highest order interaction)
15	Legionnaires' disease Obs(*N*) = 780; f_1 = 418; f_2 = 207; f_3 = 155	To assess Legionella incidence and completeness of notification in the Netherlands, 2000-2001	1. Passive national notification register 2. Active laboratory survey 3. National hospital admission registration	Saturated model with all interactions (including highest order interaction), later log-linear model with covariates
16a	Meningococcal disease Obs(*N*) = 4599; f_1 = 1054; f_2 = 1311; f_3 = 2234	Assessment of completeness of three data sources for meningococcal disease after correction for false-positive diagnoses in the Netherlands, 1993-1999	1. Notification register 2. Hospital episode statistics 3. Reference laboratory for bacterial meningitis records	Saturated model with all interactions (including highest order interaction)
16b	Obs(*N*) = 4506; f_1 = 961; f_2 = 1311; f_3 = 2234			Saturated model with all interactions (including highest order interaction) after correction for estimated number of non-laboratory confirmed false-positive cases

Obs(*N*): number of observed patients; f_1: number of patients registered once f_2: number of patients registered twice; f_3: number of patients registered by all three sources; HIV: Human Immunodeficiency Virus; AIDS: Acquired Immunodeficiency Syndrome

Study numbers:

1. Infuso A, Hubert B, Etienne J. Underreporting of Legionnaires' disease in France: the case for more active surveillance. *Euro Surveill* 1998; 3: 48-50.
2. Devine MJ, Bellis MA, Tocque K, Syed Q. Whooping cough surveillance in the north west of England. *Commun Dis Public Health* 1998; 1: 121-5.
3. Acin E, Gomez P, Hernando P, Corella I. Incidence of AIDS cases in Spanish penal facilities through the capture-recapture method, 2000. *Euro Surveill* 2003; 8: 176-81.

4. Gallay A, Vaillant V, Bouvet P, Grimont P, Desenclos JC. How many foodborne outbreaks of Salmonella infection occurred in France in 1995? Application of the capture-recapture method to three surveillance systems. *Am J Epidemiol* 2000; 152: 171-7.
5. Crowcroft NS, Andrews N, Rooney C, Brisson M, Miller E. Deaths from pertussis are underestimated in England. *Arch Dis Child* 2002; 86: 336-8.
6. Izquierdo Carreno A, Matute Cruz P, Martinez Navarro F. [The use of the capture- recapture method in evaluating the epidemiological meningococcal disease monitoring system in Tenerife, Spain (1999-2000)]. *Rev Esp Salud Publica* 2003; 77: 701-11
7. Breen E, Ghebrehewet S, Regan M, Thomson AP. How complete and accurate is meningococcal disease notification? *Commun Dis Public Health* 2004; 7: 334-8.
8. Baussano I, Bugiani M, Gregori D, Van Hest R, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung Dis* 2006; 10: 415-21.
9. Faustini A, Fano V, Sangalli M, Ferro S, Celesti L, Contegiacomo P, Renzini V, Perucci CA. Estimating incidence of bacterial meningitis with capture-recapture method, Lazio Region, Italy. *Eur J Epidemiol* 2000; 16: 843-8.
10. Van Hest NA, Smit F, Verhave JP. Improving malaria notification in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-77.
11. Nardone A, Decludt B, Jarraud S, Etienne J, Hubert B, Infuso A, Gallay A, Desenclos JC. Repeat capture-recapture studies as part of the evaluation of the surveillance of Legionnaires' disease in France. *Epidemiol Infect* 2003; 131: 647-54.
12. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med* 2001; 20: 3123-57.
13. Van Hest NA, Smit F, Baars HW, De Vries G, De Haas PE, Westenened PJ, Nagelkerke NJ, Richardus JH. Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiol Infect* 2006; Published on-line: 7 December 2006; doi:10.1017/S0950268806007540
14. Van Hest NA, Story A, Grant AD, Antoine D, Crofts JP, Watson JM. Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999 -2002. Unpublished observation.
15. Van Hest NA, Hoebe CJ, Den Boer JW, Vermunt JK, IJzerman EP„ Boersma WG, Richardus JH.. Incidence and completeness of notification of Legionnaires' disease in the Netherlands: covariate capture-recapture analysis acknowledging geographical differences. Accepted for publication Epidemiol Infect.
16. De Greeff SC, Spanjaard L, Dankert J, Hoebe CJ, Nagelkerke N, De Melker HE. Underreporting of meningococcal disease incidence in the Netherlands: Results from a capture-recapture analysis based on three registration sources with correction for false-positive diagnoses. *Eur J Epidemiol* 2006; 21: 315-21.

Table 10.2 Total number of patients observed and their distribution over three registers in various three-source capture-recapture studies of infectious diseases since 1997.

Study number	Disease (Number of patients observed)	Only register 1 (100)	Only register 2 (010)	Only register 3 (001)	Register 1 and register 2 (110)	Register 1 and register 3 (101)	Register 2 and register 3 (011)	All registers (111)
1	Legionnaires' disease (256)	7	73	46	6	10	100	14
2	HIV/AIDS (173)	26	17	22	17	29	29	33
3a	Pertussis (1-4 yrs) (435)	24	285	66	19	4	33	4
3b	Pertussis (>5 yrs) (420)	17	308	51	20	1	21	2
4	Salmonella infection (608)	45	24	451	10	39	19	20
5	Pertussis (33)	12	1	6	2	6	4	2
6	Meningococcal meningitis (53)	5	2	2	4	7	3	30
7	Meningococcal meningitis (81)	1	1	0	14	15	1	49
8	Tuberculosis (657)	125	64	30	183	96	6	153
9	Meningitis, bacterial (199)	5	52	7	7	6	46	76
10	Malaria (667)	54	41	189	37	94	127	123
11	Legionnaires' disease (715)	132	93	161	77	52	105	95
12	Hepatitis A (271)	69	55	63	21	17	18	28
13a	Tuberculosis (1499)	78	93	301	30	510	99	388
13b	Tuberculosis (1441)	40	35	301	30	548	99	388
14	Tuberculosis (28 678)	7777	2478	1544	6503	3789	512	6075
15	Legionnaires' disease (780)	56	30	332	31	131	45	155
16a	Meningococcal disease (4599)	189	253	612	189	314	808	2234
16b	Meningococcal disease (4506)	172	250	536	189	314	808	2234

HIV/AIDS: Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome

Table 10.3 Comparison of the various log-linear model, truncated Poisson mixture model, truncated Poisson heterogeneity model and truncated binomial model estimates of three-source capture-recapture studies of infectious diseases since 1997, with the ratio between the number of patients registered once and twice and coefficient of variation of the data source probabilities.

Study number	Log-linear model (95% CI)	Truncated binomial model est(N) = obs(N) + (f_1)^2/3f_2 (PI)	Truncated Poisson mixture model[22] est(N)= obs(N)/[1–exp(-2f_2 /f_1)] (PI)	Poisson heterogeneity model[23,24] est(N) = obs(N) + (f_1)^2/2f_2 (PI)	f_1/f_2	Coefficient of variation
1	291 (276-308)	302 (281-322)	304 (271-346)	323 (299-362)	1.09	0.52
2	190 (181-203)	192 (172-212)	192 (170-222)	201 (188-226)	0.87	0.07
3a	895 (735-1055)	1272 (1022-1522)	1685 (1363-2206)	1691 (1335-2187)	6.70	0.76
3b	1153 (785-1521)	1542 (1159-1925)	2098 (1640-2910)	2103 (1590-2841)	8.95	0.90
4	1065 (913-1217)	1933 (1576-2291)	2642 (2133-3469)	2596 (2088-3278)	7.65	0.86
5	46 (31-71)	43 (36-50)	46 (43-76)	48 (38-78)	1.58	0.33
6	54	55 (42-68)	55 (47-68)	56 (54-67)	0.64	0.07
7	81	81 (63-99)	81 (71-94)	81 (95% PI inestimable)	0.07	0.09
8	704 (688-728)	713 (672-754)	710 (666-759)	741 (715-776)	0.77	0.27
9	236 (206-306)	222 (200-244)	236 (210-270)	234 (217-262)	1.08	0.26
10	774 (740-821)	769 (733-805)	794 (739-857)	823 (781-877)	1.10	0.26
11	1124 (973-1275)	927 (893-962)	1018 (936-1115)	1033 (959-1125)	1.65	0.06
12	1314 (685-2899)	479 (392-586)	601 (494-769)	583 (480-737)	3.30	0.04
13a	2053 (1871-2443)	1615 (1552-1679)	1606 (1541-1677)	1673 (1635-1721)	0.74	0.29
13b	1547 (1513-1600)	1510 (1433-1592)	1481 (1424-1544)	1545 (1519-1580)	0.56	0.33
14	42 969 (41 360-44 783)	32 973 (32 732-33 215)	34 149 (33 773-34 532)	35 121 (34 822-35 434)	1.09	0.30
15	1253 (1019-1715)	1061 (1017-1106)	1241 (1136-1367)	1202 (1108-1323)	2.02	0.39
16a	5962 (5581-6343)	4881 (4738-5029)	5016 (4908-5129)	5023 (4959-5098)	0.80	0.12
16b	5577	4741 (4548-4940)	4821 (4719-4927)	4858 (4802-4924)	0.73	0.12

CI: Confidence Interval; est(N): number of estimated patients; obs(N): number of observed patients; f_1: number of patients registered once; f_2: number of patients registered twice; PI: Probability Interval; exp: exponential

independent (without interactions) or parsimonious log-linear model while Chao's model estimates were slightly higher. When a saturated log-linear model (with all two-way interactions) was selected the truncated estimates were considerably lower than the log-linear model estimates.

c. $1.5 < f_1/f_2 < 3.5$ (datasets 5, 11, 12, 15). In the first study the results of all truncated models were similar to the parsimonious log-linear model estimate but the number of observed patients was small. In the second study the estimates of Zelterman's and Chao's truncated models were lower but within the 95% confidence interval (CI) of the parsimonious log-linear model estimate while the truncated binomial model estimate was considerably lower. In the third study all truncated model estimates were considerably lower than the saturated log-linear model estimate, the truncated binomial estimate again being lowest. In the fourth study all truncated model estimates were lower than the saturated log-linear model estimate but fell within the broad 95% CI, the truncated binomial model estimates again lowest.

d. $f_1/f_2 > 3.5$: datasets (3a, 3b, 4). In all studies the truncated model estimates were considerably higher than the parsimonious log-linear model estimates, especially the Zelterman and Chao models.

Selected log-linear model

On the basis of the selected log-linear model the studies can be divided in three categories:

a. Independent log-linear model (datasets 1, 2). In these studies the truncated models produce similar estimates as the log-linear model.

b. Parsimonious log-linear model (datasets 3-11). In the 11 studies with a parsimonious log-linear model selected three observations can be made:
   - In the three studies with $f_1 >> f_2$ (datasets 3, 4) the truncated binomial model estimates a higher number of patients than the log-linear model while the truncated Poisson and Poisson mixture models estimate a considerably higher number of patients.
   - In the three studies (datasets 5-7) with a small number of observed patients the estimates of the log-linear model and truncated Poisson, Poisson mixture and binomial models are comparable.
   - In the studies with the f_1/f_2 ratio between 0.5 and 1.5 (datasets 8-10, 13b) the truncated model estimates are similar to the log-linear model but the Chao models can be relatively higher and in one study the truncated Poisson mixture estimate was relatively low.

c. Saturated log-linear model (datasets 12-16, apart from 13b). In all but one of the studies with a saturated model selected (datasets 12, 13a, 14, 16) the truncated models gave considerable lower and mutually comparable estimates.

Discussion

Main findings

In three-source log-linear model capture-recapture studies of infectious disease incidence with an independent log-linear model selected, truncated models yield comparable estimates. The truncated models also give similar results when parsimonious log-linear models are selected and the number of patients is limited or the f_1/f_2 ratio is between 0.5 and 1.5. When $f_1>>f_2$ truncated models give considerable higher estimates than parsimonious log-linear models. Compared to saturated log-linear models the truncated models produce considerably lower and often more plausible estimates.

Capture-recapture analysis and chronic diseases

For human diseases capture-recapture analysis has predominantly been applied to estimate the prevalence, incidence or completeness of registers of specific groups of diseases, often diseases with a chronic character as mentioned earlier. Apparently the characteristics of most of these diseases, their patients and their registers best fulfil criteria for feasibility of capture-recapture studies as well as validity of the underlying assumptions. Perhaps with the exemption of some neurological and rheumatological conditions, the case-definition is probably unambiguous and uniform over the various registers. Arguably, for these categories of diseases sufficient registers are available and possible relationships between these registers, e.g. clinical registers, laboratory registers, health insurance registers or patient support and advocacy group registers, be they positive or negative, could be avoided by source selection or source merging or accounted for in a log-linear model, thus limiting violation of the independent registers assumption. The permanent character of most of these conditions can reduce violation of the closed population assumption.

Capture-recapture analysis and infectious diseases

For infectious diseases the number of available registers for record-linkage, usually notification-, laboratory- or hospital-based registers, is often limited and (strong) positive interaction between these registers should be expected as a result of the characteristics of infectious disease diagnosis and treatment and public health regulations. Infectious disease control and surveillance is often organized around close collaboration between clinicians, microbiologists and public health professionals, such as infectious disease and tuberculosis physicians and nurses. Only two of the 19 datasets studied selected the independent log-linear model and 11 datasets selected parsimonious log-linear models incorporating one or two pair-wise dependencies. However, six datasets selected the saturated log-linear model, i.e. including all two-way interactions and assuming absence of the three-way interaction.[16,36] Our studies of tuberculosis incidence in England and, before correction for suggested imperfect record-linkage and remaining false-positive hospital cases, in the Netherlands both selected a saturated model, resulting in unexpectedly and unrealistically high estimates of the number of tuberculosis patients. The two previous three-source log-linear model capture-recapture studies of tuberculosis incidence resulted in a parsimonious model and both produced plausible estimates within the range of prior expectations.[37,38] According to Hook and Regal, if the saturated model

is selected by any criterion the investigator should be particularly cautious about using the associated outcome.[10] At the time of our studies on tuberculosis incidence all but one of the published three-source log-linear capture-recapture studies of infectious incidence used independent or parsimonious log-linear models (studies 1–11). The one published study selecting a saturated log-linear model (study 12) gave a much higher estimate (n = 1314) of the number of hepatitis A patients in an outbreak in Taiwan than later established by serology results (n = 545).[19] Recently a three-source log-linear model capture-recapture study of meningococcal disease incidence also selected a saturated log-linear model and resulted in relatively high estimates (study 16).[39] Perhaps confidence in the validity of capture-recapture results may reflect publication bias in favour of apparently successful capture-recapture studies.[40] The unexpectedly high estimates of the saturated log-linear model capture-recapture studies do not result from violation of the "absent three-way interaction" assumption. In the case of infectious disease registers, existing three-way interaction is almost certainly positive, causing a capture-recapture estimate biased downwards.[39] The reason for the high estimates must, therefore, be violation of (a combination of) the other underlying assumptions. After correction for possible false-positive records and possible imperfect record-linkage the capture-recapture studies on tuberculosis and meningococcal disease in the Netherlands (studies 13 and 16) produced much lower and lower estimates respectively. Compared to an initial saturated log-linear model, a covariate log-linear capture-recapture model, reducing violation of the homogeneity assumption, also resulted in a much lower estimate of 886 (95%CI 827-1022) Legionnaires' disease patients in the Netherlands (study 15).

Truncated estimators and infectious diseases

Infectious disease studies where an independent log-linear model was selected produce estimates very similar with the truncated models, which can be partly explained by the independent register assumption underlying the truncated models when applied to three registers. That truncated estimators perform well when data are sparse is demonstrated in studies 5, 6 and 7 as the estimates of the log-linear and the various truncated models are similar. The truncated models also give similar results as the log-linear models when $0.5 < f_1/f_2 < 1.5$ but give considerably higher estimates when $f_1 >> f_2$. In the case of saturated log-linear models (studies 12-16), with unexpectedly high estimates of infectious disease incidence, the lower truncated model estimates are more plausible but are they also preferable? We have two arguments to support the view they might be:

1. In study 12 the saturated log-linear model estimated 1314 patients with hepatitis A infection in an outbreak in Taiwan while the truncated models estimate between 500 and 600 patients. The National Quarantine Service of Taiwan, on the basis of serology tests, later concluded that the true number of infected persons was about 545, making this one of the few capture-recapture datasets where later a true number of patients was established.[19]

2. A saturated log-linear model in dataset 13a gave an implausible estimate of 2053 (95%CI 1871-2443) tuberculosis patients in the Netherlands in 1998, while truncated models estimated between 1600 and 1675 patients. The implausible estimate caused the investigators to have a critical look at the data again and make further corrections

for probable imperfect record-linkage and possible remaining false-positive records in the hospital register. The parsimonious log-linear model of dataset 13b fitted the adjusted data well and gave an estimate of 1547 (95%CI 1513-1600) tuberculosis patients and corresponding truncated model estimates. The initial truncated model estimates came relatively close to the final log-linear model estimate.

The equiprobability and number of data sources assumptions

The truncated binomial model assumes that all sources have the same probability of capturing a case. In addition the truncated Poisson model assumes an infinite number of sources, although in our data the number of sources was limited to three. On this argument the truncated binomial model for three data sources is a more realistic alternative estimator. However, any departure from equiprobability results in an estimation error, which analytically is overestimation (see Appendix). Realistic estimates of this error can be obtained from the data. In Table 3 the last column shows the coefficients of variation, a measure of variability in the coverages of the three data sources for each study. This is calculated as the standard deviation divided by the mean from the three quantities N_1 (number of cases known on source 1), N_2 (number of cases known on source 2) and N_3 (number of cases known on source 3). We demonstrate the possible effect of violation of the equiprobability assumption by studies 4 and 11. For study 4, which has a high coefficient of variation (0.86), if the sources were truly independent, the number of unobserved cases would be 702, calculated by fitting the log-linear model with main effects only. Our truncated binomial estimator gives 1325 cases, nearly twice as large. For study 11, with a low coefficient of variation (0.06), independence implies that there are 155 unobserved cases, while the truncated binomial estimate is 212, an overestimation by about 30%. Studies 3 and 4 indicate that the high f_1/f_2 ratios result from violation of the equiprobability assumption, producing overestimates by the truncated models.

Two-source validation

Any three-source study can be used to test two-source estimation by treating one source as though it were a complete list of cases and extract a complete 2 x 2 table. We demonstrate this for two studies, numbers 4 and 11, which we chose above for their coefficients of variation and took register 3 as the complete set. Validation was by comparing the Petersen estimator ($N_{10}\ N_{01}/N_{11}$) [1] and the truncated binomial estimator, which for two lists is $(f_1)^2/4f_2$, on the 2 x 2 table with the known "unlisted" number. For study 4 there were 451 "unlisted" cases, i.e. on neither of registers 1 and 2. The Petersen estimator is 37 and the truncated binomial estimator 42. The two estimators are similar because registers 1 and 2 have approximately equal coverage but both are far short of the true figure (Zelterman and Chao models estimates are 79 and 84 respectively). For study 11 there were 161 "unlisted" cases and the two estimators were 57 and 64. Again the estimators agree but are short of the true figure. Now the Zelterman and Chao model estimates are 107 and 130, respectively, and perform slightly better. However, we had some hesitation in extracting 2 x 2 tables from three-source capture-recapture data, more specifically from capture-recapture studies on infectious disease incidence. As explained earlier, (positive) interdependencies between the three conventional registers used for

such studies should be expected. Extracting 2 x 2 tables ignores possible conditional dependence confounding the results thus obtained. For studies 4 and 11 the log-linear models included one respectively two interaction terms for pair-wise dependencies, which may explain the underestimation in the Petersen and truncated estimators. We therefore also validated the two studies with independent log-linear models (studies 1 and 2). We took register 2 as the complete set for study 1 and register 3 as the complete set for study 2. For study 1 there were 73 "unlisted" cases. The Petersen estimator, 43, is a little low, but the truncated binomial estimator, at 201, is too high (Zelterman and Chao models estimates are 397 and 401, respectively). The discrepant (over)estimate by the truncated models can be explained by the different coverages of registers 1 and 3, i.e. violation of the equiprobability assumption. In study 2 the coefficient of variation was low and the coverage of registers 1 and 2 similar. For study 2 there were 22 "unlisted" cases. The Petersen estimator and the truncated binomial estimator are both 25 and similar to the known "unlisted" number, explained by almost absent violation of both the independent sources and equiprobability assumptions. The Zelterman and Chao models estimates are 43 and 51 respectively and the discrepancy with the truncated binomial model estimate can be explained by violation of the "infinite number of sources" assumption.

Alternative models

As an alternative to log-linear capture-recapture models a structural source model has been proposed.[36] Whereas log-linear models only partly identify and incorporate dependencies between registers, the structural source model models potential interdependencies of the registers and heterogeneity of the population, partly based on prior knowledge, and estimates the probabilities of conditions that produce these interactions between the registers. However, the published data of the capture-recapture studies were insufficient to re-examine these studies with a structural source model.

Conclusion

We have indicated conditions where estimates of infectious disease incidence from log-linear models are similar or dissimilar to alternative truncated models for incomplete count data. Our results suggest that for estimating infectious disease incidence and completeness of notification independent and parsimonious three-source log-linear capture–recapture models are preferable. When saturated models are selected as best-fitting model and the estimates are unexpectedly high and seem implausible, first, 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 estimated number of infectious disease patients. Second, in case of such discrepancy between the log-linear and the truncated model estimates, the data should be re-examined for possible violation of the underlying capture-recapture assumptions, such as imperfect record-linkage, false-positive records or heterogeneity, corrected and the capture-recapture analysis repeated on the corrected data. When after repeated capture-recapture analysis the discrepancy between the log-linear and the truncated model estimates remains or no violation of the underlying assumptions can be identified, the investigator should be cautious about using the associated outcome.[10] Using truncated model estimates as an early alert could prevent flawed capture-recapture

estimates finding their way into the scientific literature. The role of the f_1/f_2 ratio in the agreement or disagreement between three-source log-linear capture-recapture and truncated model estimates for the number of infectious disease patients, especially when a parsimonious log-linear model is selected, should be subject of further mathematical or statistical studies.

Appendix

Equations for the truncated population estimators

Truncated binomial model: est(N) = obs(N) + $(f_1)^2/3f_2$

Truncated Poisson mixture model: est(N) = obs(N)/[1–exp(-$2f_2/f_1$)]

Truncated Poisson heterogeneity model: est(N) = obs(N) + $(f_1)^2/2f_2$

Equiprobability

If the truncated binomial model is true, i.e. if the sources are independent and equiprobable with probability of capturing any case = p, our estimator $(f_1)^2/3f_2$ is correct in the sense that the expected number of unlisted cases is given by

$$\mathbf{E}f_0 = Nq^3 = \frac{(\mathbf{E}f_1)^2}{3\mathbf{E}f_2}. \tag{1}$$

If we introduce a small departure from equiprobability so that the list probabilities are $(p - h, p, p + h)$ instead of (p, p, p), the estimation error can be defined as

$$g(h,p) = \frac{(\mathbf{E}f_1)^2}{3\mathbf{E}f_2} - \mathbf{E}f_0. \tag{2}$$

Differentiating with respect to h, we find that

$$g(0,p) = \frac{\partial g}{\partial h}(0,p) = 0; \quad \frac{\partial^2 g}{\partial h^2}(0,p) = \frac{2N(1-p)}{3p^2}, \tag{3}$$

so that we overestimate, at least for small h. The same happens if we consider an asymmetrical departure, $(p - h, p, p)$. In that case,

$$g(0,p) = \frac{\partial g}{\partial h}(0,p) = 0; \quad \frac{\partial^2 g}{\partial h^2}(0,p) = \frac{2N(1-p)}{9p^2}, \tag{4}$$

and there is again an overestimate.

References

1. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation I: History and theoretical development. *Am J Epidemiol* 1995; 142: 1047-58.
2. LaPorte RE, Dearwater SR, Chang YF, Songer TJ, Aaron DJ, Anderson RL, Olsen T. Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring. *Am J Epidemiol* 1995; 15; 142: 1069-77.
3. Orton H, Richard R, Miller L. Using active medical record review and capture-recapture methods to investigate the prevalence of Down Syndrome among live-born infants in Colorado. *Teratology* 2001; 64: S14-9.
4. EURODIAB ACE Study Group. Variation and trends in incidence of childhood diabetes in Europe. *Lancet* 2000; 355: 873-6.
5. McClish D, Penbertthy L. Using Medicare data to estimate the number of cases missed by a cancer registry: a 3-source capture-recapture model. *Med Care* 2004; 42: 1111-6.
6. Tilling K, Sterne JA, Wolfe CD. Estimation of the incidence of stroke using a capture-recapture model including covariates. *Int J Epidemiol* 2001; 30: 1351-9.
7. Mahr A, Guillevin L, Poisonnet M, Avme S. Prevalences of polyarteritis nodosa, microscopic polyangiitis, Wegener's granulomatosis, and Churg-Strauss syndrome in a French urban multiethnic population in 2000: a capture-recapture estimate. *Arthritis Rheum* 2004; 51: 92-9.
8. Fienberg SE. The multiple-recapture census for closed populations and the 2k incomplete contingency table. *Biometrika* 1972; 59: 591-603.
9. Bishop YM, Fienberg SE, Holland PW. *Discrete Multivariate Analysis*. Cambridge: MIT-Press, 1975.
10. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-63.
11. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation II: Applications in human diseases. *Am J Epidemiol* 1995; 142: 1059-68.
12. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. *Int J Epidemiol* 1994; 23: 1322-3.
13. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology* 1995; 6: 42-8.
14. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. *J Clin Epidemiol* 1999; 52: 909-14.
15. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. *Int J Epidemiol* 1999; 25: 474-8.
16. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* 2000; 152: 771-9.
17. Jarvis SN, Lowe PL, Avery A, Levene S, Cormack RM. Children are not goldfish-mark/recapture techniques and their application to injury data. *Inj Prev* 2000; 6: 46-50.
18. Tilling K. Capture-recapture methods-useful or misleading? *Int J Epidemiol* 2001; 30: 12-4.
19. Chao A, Tsay PK, Lin SH, Shau WY, Chao DY. The applications of capture-recapture models to epidemiological data. *Stat Med* 2001; 20: 3123-57.
20. Van Hest NA, Smit F, Baars HW, De Vries G, De Haas PE, Westenened PJ, Nagelkerke NJ, Richardus JH. Completeness of notification of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiol Infect* 2006*;* Published online: 7 December 2006. doi:10.1017/S0950268806007540
21. Van Hest NA, Smit F, Verhave JP. Improving malaria notification in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-7.
22. Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J Stat Plann Inf* 1988; 18: 225-37.
23. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; 43: 783-91.
24. Chao A. Estimating animal abundance with capture frequency data. *J Wildl Manage* 1988; 52: 295-300.
25. Hook EB, Regal RR. Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *Am J Epidemiol* 1982; 116: 168-76.
26. Van der Heijden PG, Cruyff MJ, Van Houwelingen, H. Estimating the size of a criminal population from police registrations using the truncated Poisson regression model. *Stat Neerl* 2003; 57: 289-304.
27. 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.

28. Bohning D, Suppawattanabodee B, Kusolvitkul, W, Viwatwongkasem C. Estimating the number of drug users in Bangkok 2001: a capture-recapture approach using repeated entries in one list. *Eur J Epidemiol* 2004; 19: 1075-83.
29. Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; 79: 543-53.
30. Smit F, Reinking D, Reijerse M. Estimating the number of people eligible for health service use. *Eval Prog Plan* 2002; 25: 101-5.
31. Hay G, Smit F. Estimating the number of hard drug users from needle-exchange data. *Addiction Res Theory* 2003; 11: 235-43.
32. Rossmo DK, Routledge R. Estimating the size of criminal populations. *J Quant Criminol* 1990; 6: 293-314.
33. Bustami R., Van der Heijden P, Van Houwelingen H, Engbersen G. Point and interval estimation of the population size using the truncated Poisson regression model. In: Klein B, Korsholm L, eds. *New trends in statistical modelling. Proceedings of the 16th international workshop on statistical modelling.* Odense: University of Southern Denmark, 2001: pp 87-94.
34. Hser YI. Population estimation of illicit drug users in Los Angeles County. *J Drug Issues* 1993; 23: 323-34.
35. Chao A. Estimating population size for sparse data in capture-recapture experiments. *Biometrics* 1989; 45: 427-38.
36. Regal RR, Hook EB. Marginal versus conditional versus structural source models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Stat Med* 1998;17: 69-74.
37. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture-recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-3.
38. Baussano I, Bugiani M, Gregori D, Van Hest R, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung Dis* 2006; 10: 415-21.
39. De Greeff SC, Spanjaard L, Dankert J, Hoebe CJ, Nagelkerke N, De Melker HE. Underreporting of meningococcal disease incidence in the Netherlands: Results from a capture-recapture analysis based on three registration sources with correction for false-positive diagnoses. *Eur J Epidemiol* 2006; 21: 315-21.
40. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; 46: 515-20.

11

General discussion

The aim of this thesis is to investigate the feasibility and validity of capture-recapture methods in surveillance of tuberculosis and other infectious diseases. This chapter provides answers to the three research questions of this thesis (section 11.1), briefly discusses some findings of this thesis in the context of surveillance of tuberculosis and other infectious diseases (section 11.2) and lists the conclusions and recommendations (section 11.3).

11.1 Answering the research questions

Question 1

How do the characteristics of various infectious diseases and their registers in the Netherlands influence the feasibility and validity of capture-recapture analysis?

Different characteristics of disease, patients and registrations can influence the feasibility and validity of capture-recapture analysis as a method to estimate infectious disease incidence and completeness of registration. In this thesis we performed three three-source log-linear capture-recapture studies on infectious diseases in the Netherlands: malaria, Legionnaires' disease and tuberculosis. First the feasibility of these studies will be discussed. After that the validity of the results of these studies will be addressed by discussing possible violation of the underlying capture-recapture assumptions; the closed population assumption, the perfect record-linkage assumption, the perfect positive predictive value assumption, the absence of specific interdependencies assumption and the homogeneous population assumption.

Feasibility

The capture-recapture study on tuberculosis incidence had the best feasibility. The Netherlands is a country with a well-organised system of tuberculosis control and seven existing tuberculosis or tuberculosis-related national registers or datasets were available for record-linkage or cross-validation. For the study on malaria incidence insufficient data sources were available for three-source log-linear capture-recapture analysis and, in the absence of a national malaria reference laboratory (as exists for tuberculosis), a survey among microbiology laboratories was necessary to create the third dataset. The capture-recapture study on Legionnaires' disease incidence even needed three surveys. Similar to the malaria study, in the absence of a national Legionella reference laboratory, an additional (third) dataset had to be created through a laboratory survey. Due to insufficient identifiers for reliable record-linkage in the notification register and the absence of a specific International Code for Diseases (ICD-9 code) in the hospital episode statistics, two further surveys were performed: one collecting additional data, such as date of birth, of notified Legionnaires' disease patients from the Public Health Services and a second one requesting chest-physicians to voluntarily report hospitalised Legionnaires' disease patients. Additional surveys cost time and money. They risk poor response (the survey among the chest-physicians yielded no additional Legionnaires' disease patients indicating that unlike laboratories, with good response rates in the two surveys mentioned, clinicians may not be a useful source of information for capture-recapture

studies). Obtaining additional data sources through surveys contravenes one of the alleged advantages of capture-recapture analysis for (infectious) disease surveillance, namely of being cheap, quick and simple by using existing data sources.[1]

Absence of a specific code for Legionnaires' disease in ICD-9 had another negative influence on the feasibility of this study. The hospital episode statistics used a proxy-code for Legionnaires' disease, ICD-9 code 482.8, inviting the presence of an unknown number of false-positive patients without Legionnaires' disease.

These experiences underpin the need for thorough examination of availability, quality and specificity of infectious disease registrations as part of the preparation of possible capture-recapture studies.

Validity

Closed population assumption

Violation of the closed population assumption is intuitively highest for malaria as malaria infection occurs almost by definition abroad while Legionnaires' disease and tuberculosis are both partly endemic diseases. However, for all three infectious diseases, translated into the capture-recapture probabilities for an individual patient, violation of the closed population assumption is presumably limited as the opportunities for notification, laboratory-confirmation or hospitalisation are determined within a relatively short period of time. All three diseases were category B disease at the time of the study, i.e. to be notified by the diagnosing physician within 24 hours. For malaria both thick and thin smear microscopy and serological antigen tests are processed fast. For Legionnaires' disease, with the increased use of the urinary antigen test, laboratory results are also rapidly available.[2] For tuberculosis sputum collection or bronchoscopy can be done quickly and smear-microscopy results are readily available. The culture can take weeks to become positive but this does not influence the capture-probability in the laboratory register. Laboratory pre-notification of positive test results to public health physicians contributes to rapid notification. Laboratory pre-notification of positive tuberculosis and Legionnaires' disease results to the Public Health Services is arguably more likely than for malaria because tuberculosis is transmitted from human to human and Legionnaires' disease can be caused by local contaminated water sources and urgent preventive interventions may be necessary. When Legionnaires' disease, a disease with a relatively high case-fatality rate, is suspected on clinical and/or epidemiological grounds, patients are likely to be immediately hospitalised, as suggested by the highest ascertained register-specific coverage for the hospital register (85%) of the three infectious diseases studied. For malaria the ascertained hospitalisation rate is lower (49.3%), indicating that mild presentations of benign forms of malaria are treated on an out-patient basis, but *Plasmodium falciparum* malaria patients are likely to be immediately hospitalised. The probability for hospitalisation for tuberculosis (40.7%) is less compared to Legionnaires' disease and malaria because in the Netherlands non-infectious tuberculosis patients are predominantly treated as out-patients and even infectious cases are preferably isolated at home. But for most tuberculosis patients the decision on hospitalisation will be taken rapidly. Limited violation of the closed population assumption could result in some

overestimation of the number of malaria, Legionnaires' disease and tuberculosis patients in the Netherlands.

Perfect record-linkage assumption

For all three infectious diseases studied, unique identification numbers, such as a social security number, used in all registers, allowing optimal record-linkage, did not exist. In all studies record-linkage was manual, through almost similar procedures described in chapters 4 to 6, using patients' identifiers such as date of birth and postcode. The malaria and tuberculosis studies were performed before 1999 when in the Netherlands a new Infectious Diseases Act came into force,[3] recording only year of birth instead of date of birth, effectively ruling out reliable record-linkage between the notification and other registers. Therefore still sufficient demographical, geographical and microbiological identifiers could be collected at source for adequate record-linkage. In addition, for tuberculosis record-linkage was relaxed and near-links were double-checked. Despite these efforts still indications for imperfect record-linkage exist. First, the notification register and the Netherlands Tuberculosis Register (NTR) should overlap completely but we found only 91.1% overlap after record-linkage. Second, some culture-positive tuberculosis patients could not be found in the notification register but were linked to the NTR. Although before cross-validation misclassification of tuberculosis patients was considered to be minimal, this study demonstrated that imperfect record-linkage can exceed expectations, with considerable impact on the capture-recapture estimates. Initially unidentified indications for misclassification could not be investigated for malaria and Legionnaires' disease.

The capture-recapture study on the incidence of Legionnaires' disease was performed after the new Infectious Diseases Act came into force. Therefore information on the date of birth of the patients could no longer be collected from the national notification register at source but had to be obtained from the local Public Health Services processing the notifications, creating more opportunities for clerical errors and remaining incomplete information on personal identifiers in some records, jeopardising reliable record-linkage. This may have caused more misclassification of patients over the registers compared to the malaria and tuberculosis capture-recapture studies.

The impact of changes in the infectious disease legislation in the Netherlands on the validity of capture-recapture estimates was demonstrated in a follow-up capture-recapture study on malaria incidence in the Netherlands between 1995 and 2003.[4] After 1999 the malaria incidence estimates more than doubled, almost certainly reflecting overestimation as a result of imperfect record-linkage. Inaccurate record-linkage, i.e. incorrectly establishing the recapture, can substantially alter the observed and unobserved fractions.[5]

Depending on the number of missed links and mislinks, violation of the perfect record-linkage assumption can result in overestimation or underestimation of the number of malaria, Legionnaires' disease and tuberculosis patients in the Netherlands.

Perfect positive predictive value assumption

The positive predictive value of the notification and laboratory registers for malaria is considered to be high. Malaria is a specific disease, often requiring a history of recent travelling in tropical areas, and unlikely to be diagnosed without confirmation or strong suspicion. The laboratory plays a crucial role in the diagnosis through thick and thin smear microscopy and serological antigen tests, all with a high specificity. Malaria has a specific ICD-9 code and although a number of patients could have been admitted to hospital for observation after developing fever following a tropical journey without a final diagnosis of malaria, compared to Legionnaires' disease and tuberculosis, the positive predictive value of the malaria hospital episode register is also expected to be high.

For the notification of Legionnaires' disease the criteria require a clinical picture compatible with pneumonia and a confirmed or probable microbiology laboratory diagnosis, according to the European Working Group for Legionella Infections (EWGLI) definition. We demonstrate after record-linkage that these criteria apparently are not applied uniformly over all registers. More than malaria, the registers could use different, less specific, case-definitions, resulting in a proportion of false-positive cases in these registers, e.g. as the result of the absence of pneumonia, the low positive predictive value of the single Legionella antibody titre test or the absence of a specific Legionnaires' disease code in ICD-9.[6,7]

For tuberculosis the number of false-positive cases is assumed to be zero in the reference laboratory register and limited in the notification register. The latter is due to a good organisation of tuberculosis surveillance in the Netherlands and identification of false-positive cases with an infection caused by non-tuberculous mycobacteria or another diagnosis than non-tuberculous mycobacteriosis or tuberculosis through our cross-validation. However, foreign reports indicate considerable contamination of tuberculosis hospital registers with false-positive cases.[8,9] The results of our study support these observations as 62.4% of the unlinked hospital cases could not be verified through cross-validation, compared to 7.6% of the unlinked notified cases. These possibly remaining false-positive cases likely contribute to considerable bias in the capture-recapture estimate. Cross-validation and identification of assumed false-positive cases could not be performed for malaria and Legionnaires' disease. Violation of the perfect positive predictive value assumption results in overestimation of the number of malaria, Legionnaires' disease and tuberculosis patients in the Netherlands.

Absence of specific interdependencies assumption

For all three infectious diseases studied, co-operation between the conventional registers used (notification, laboratory and hospital) is expected, resulting in positive dependence and underestimation of the number of cases in two-source capture-recapture models. Therefore three-source log-linear capture-recapture analysis, incorporating possible pair-wise dependencies, was selected in the study design to reduce bias.

For malaria, significant interaction between the laboratory and notification registers is not identified and could indicate absent pre-notification of laboratory results to the Public Health Services at the time of this study. The explicable interactions of notification by

clinicians to the Public Health Services and laboratories likely to actively approach clinicians in case of falciparum malaria, a potentially short-term fatal illness requiring immediate treatment, are identified in the data and incorporated in a relatively parsimonious log-linear capture-recapture model. In this model the absence of three-way interaction, i.e. interaction between all three registers, has to be assumed but, when present, three-way interaction bias is arguably less than in a saturated capture-recapture model, incorporating all two-way interactions.

For Legionnaires' disease internal validity analysis and stratified capture-recapture analysis indicated a significant interaction between the notification and the laboratory registers. This interaction is expected as the notification criteria specifically include laboratory diagnosis and at the time of this study many laboratories probably pre-notified positive Legionella test results to public health physicians. Log-linear capture-recapture analysis for Legionnaires' disease incidence initially selected the saturated model as the best-fitting model. However, capture-recapture models with a better fit do not necessarily produce a more reliable estimate.[10] Three-way interaction cannot be incorporated in the saturated model and, when present, could render the estimates less valid. However, violation of the absent three-way interaction assumption does not seem to explain possible bias in the capture-recapture estimate for Legionnaires' disease incidence. This estimate is considered to be high while dependence between the registers is expected to be positive, resulting in underestimation, and the unbiased estimate would be even higher.[11-13]

The most significant (positive) interactions between the registers are expected for tuberculosis, not only because tuberculosis is an infectious disease with human to human transmission but also as a result of the well-organised system of tuberculosis control and surveillance in the Netherlands. Initially a saturated log-linear capture-recapture model, with the best goodness-of-fit, was selected but the estimate seemed implausibly high. For similar reasons as described for Legionnaires' disease, violation of the absent three-way interaction assumption cannot explain bias in this estimate. After corrections for possible violations of the perfect record-linkage and perfect positive predictive value assumptions, a relatively parsimonious log-linear capture-recapture model fitted the data best and gave a considerably lower and more plausible estimate. The two significant two-way interactions in this model are between the notification and laboratory registers and between the notification and hospital registers. The first interaction can be partially explained by laboratory pre-notification and partially by tuberculosis control physicians, who are processing the notifications, diagnosing approximately one-third of all tuberculosis patients in the Netherlands, almost exclusively patients with pulmonary tuberculosis and a high bacteriological confirmation rate. The fact that in the Netherlands two-third of all tuberculosis cases are treated by a limited group of hospital-based clinicians, often familiar with the notification procedures, and referral of patients by tuberculosis control physicians for further clinical evaluation or isolation could explain the interaction between the notification and hospital registers. Perhaps the fact that most cases of extrapulmonary tuberculosis, less often culture-confirmed, are diagnosed in a hospital explains why in our capture-recapture model the interaction between the laboratory and hospital registers is not significant.

Homogeneous population assumption

Possible heterogeneity of the patient population, i.e. the presence of categorical covariates associated with the probability of capture in a register, causing bias in the capture-recapture estimate, cannot be excluded in the three infectious diseases studies and was examined in three different ways. For the malaria study we performed a stratified capture-recapture analysis by *Plasmodium* species which showed limited variety in capture-recapture probabilities. However, we cannot exclude the possible presence of other (but unmeasured) sources of heterogeneity,

For the tuberculosis data, after conventional log-linear capture-recapture analysis, we applied alternative truncated population estimation models, arguably more robust in the presence of heterogeneity, and these models gave identical results.

Because regional differences in the incidence rate of Legionnaires' disease were described in the Netherlands and abroad,[6,14] and observed in our data after record-linkage, alternative to conventional log-linear capture-recapture analysis a log-linear covariate capture-recapture model was specified, with region as covariate, to reduce bias due to geographical heterogeneity. The better goodness-of-fit and narrow confidence interval suggest a more valid estimate with less statistical uncertainty compared to the outcome of the conventional saturated model. Unfortunately, the data-quality of the hospital episode register prevented meaningful inclusion of other covariates in the model, possibly causing bias, such as the method of laboratory diagnosis, described as a relevant covariate elsewhere.[6] At the time of the malaria and tuberculosis studies we were not familiar with covariate capture-recapture techniques and did not explore this methodology to investigate the impact of possible heterogeneity.

Question 2

How do the characteristics of tuberculosis surveillance systems in different countries influence the feasibility and validity of capture-recapture analysis?

Different characteristics of tuberculosis surveillance systems can influence the feasibility and validity of capture-recapture analysis. In this thesis we describe three three-source log-linear capture-recapture studies on tuberculosis incidence and completeness of notification in three different European countries, the Netherlands, the Piedmont region of Italy and England, at a different scale, both administrative and regarding the number of patients involved, and with different tuberculosis surveillance systems. The feasibility and validity of these studies will be discussed by country, after a short introduction to tuberculosis surveillance in each country, followed by a comparison of the results of the three tuberculosis capture-recapture studies.

The Netherlands

In the Netherlands the annual national tuberculosis incidence has been decreasing during the last decade. This can be partially explained by the nation-wide system of Public Health Tuberculosis Clinics, with public health tuberculosis physicians and nurses, performing diagnostic, curative, preventive and screening activities, parallel to and in good co-

operation with the chest physicians in the hospitals. Furthermore, there is a strong non-governmental tuberculosis control advocacy organisation, the KNCV Tuberculosis Foundation. A comprehensive and reliable system of tuberculosis surveillance was assumed and the level of under-notification of tuberculosis was previously estimated at 8%.[15] This influenced the feasibility of the capture-recapture study as at least seven existing tuberculosis or tuberculosis-related data sources were available at the national level for record-linkage and cross-validation of the three conventional capture-recapture sources (notification, laboratory and hospital). None of these data sources were routinely linked. The size of the study still allowed for manual record-linkage and double-checking of data but this was a time-consuming process.

The validity of the final estimate is considered high as many aspects of possible violation of the assumptions underlying capture-recapture analysis could be investigated in detail and, when identified, corrected, as explained in the answer to research question 1.

The Piedmont region of Italy

Tuberculosis incidence in Italy remained relatively stable over the past decade. Although Italy operates a system of Public Health Tuberculosis Clinics the overall level of organisation of tuberculosis control and surveillance is assumed to be less compared to the Netherlands, e.g. due to the size of the country, probable regional differences and the absence of a strong non-governmental national tuberculosis control advocacy organisation. Overall under-notification of tuberculosis in Italy was previously estimated at 12% but could be as high as 37%-54% in some areas.[16-18] The feasibility of the capture-recapture study is positively influenced by the regional level and absence of major legal obstructions. Four, later three, existing tuberculosis data sources were available. Because of the overlap of the notification and treatment outcome monitoring registers these two registers were merged into one 'physician notification system'. None of the registers were routinely linked. Legal and privacy regulations in Italy made it possible to collect information on the (known) HIV status of tuberculosis patients. Despite the smaller scale of the study record-linkage was not performed completely manually but feasibility was promoted by initial computerised deterministic record-linkage followed by manual review of the near-links. The regional set-up also promoted the feasibility of investigating hospital charts for false-positive cases by hand.

The validity of the final estimate is considered to be high as violation of the assumptions underlying capture-recapture analysis is assumed to be limited. For similar reasons as in the Netherlands, explained in the answer to research question 1, violation of the closed population assumption is considered minor for tuberculosis. Computerised deterministic record-linkage followed by manual review of near-links limits violation of the perfect record-linkage assumption. Laboratory data in Piedmont are routinely checked for false-positive records and manual examination of hospital charts could exclude a considerable number of false-positive cases in this register, both exercises limiting violation of the perfect positive predictive value assumption. A parsimonious log-linear model was selected, incorporating one two-way interaction, reducing violation of the independent registers assumption. Possible relevant population heterogeneity was examined during log-linear modelling (Table 7.2) and excluded on the basis of statistical

arguments. In addition, stratified capture-recapture analysis was performed, mostly giving similar estimates of the total number of tuberculosis patients compared to the unstratified capture-recapture estimate. However, Table 7.3 reflects some possible heterogeneity and violation of the homogeneity assumption cannot be completely excluded.

England

Since 1987 a rise in notifications of tuberculosis has been observed in England and Wales, reflecting an increase in diagnoses of tuberculosis rather than an artefact due to improved reporting.[19] The quality of tuberculosis control and surveillance in the United Kingdom has been questioned.[20] The United Kingdom has no Public Health Tuberculosis Clinics and responsibilities for diagnosis, treatment, prevention and surveillance are divided between chest physicians in the hospitals, consultants in communicable disease control in the Primary Care Trusts and tuberculosis nurses and social workers. There is no strong non-governmental national tuberculosis control advocacy organisation. Tuberculosis under-notification in the United Kingdom is estimated between 7% and 27%.[21] In 1999, a revised national routine surveillance system for tuberculosis, Enhanced Tuberculosis Surveillance (ETS), was introduced to improve the completeness of notification as well as the information on notified cases. Three data sources were available for record-linkage and two tuberculosis-related datasets for cross-validation although these two did not cover the study period and deductions had to be made. Notifications through ETS are routinely linked to *Mycobacterium tuberculosis* isolate records from the reference laboratories in the United Kingdom Mycobacterial Network (MycobNet), an advantage for the feasibility of the capture-recapture study. A further advantage in terms of feasibility was that appropriate deterministic-probabilistic record-linkage software was already developed by the Centre for Infections and could, with some modifications, be used for record-linkage of the hospital episode records. On average 7000 tuberculosis patients were annually ascertained in England during the four years studied which had a disadvantageous impact on the feasibility of the capture-recapture study compared to the much smaller scale of the studies performed in the Netherlands and the Piedmont region of Italy.

The validity of the unexpectedly high and apparently implausible final estimate is debatable, as a result of possible violation of the assumptions underlying capture-recapture analysis. For similar reasons as explained for the Netherlands and the Piedmont region of Italy, violation of the closed population assumption is considered limited for tuberculosis. Despite the large number of tuberculosis patients involved, preventing detailed manual review of the outcomes of the computerised record-linkage procedure, violation of the perfect record-linkage procedure may be limited. We found that 94.9% of the linked hospital tuberculosis cases had a likelihood of association score of 3000 points or more and only 5.1% with such a score were not linked to the ETS registers, suggesting a reliable cut-off point in the record-linkage procedure, with possibly balanced misclassification. However, the study in the Netherlands shows that even limited violation of the perfect record-linkage assumption can have a considerable impact on the capture-recapture estimates. Also vulnerable to violation is the perfect positive predictive value assumption as the proportion of false-positive records among the unlinked hospitalised cases had to be estimated with a logistic regression population mixture model. Estimation

of the proportion of false-positive records through complex mathematical procedures for infectious disease capture-recapture analysis has been described previously.[22] The estimated proportions of false-positive records in the hospital register or among unlinked hospitalised cases were similar to those found in a previous local capture-recapture study in England[8] and in the studies in the Netherlands and the Piedmont region of Italy. The cross-validation of non-culture-confirmed cases also limited violation of the perfect positive predictive value assumption but it cannot be excluded. The log-linear capture-recapture model with the best goodness-of-fit was the saturated model. As discussed in the answer to research question 1, violation of the absent (positive) three-way interaction assumption is unlikely to explain the implausible high estimate.[22] According to Hook and Regal, if the saturated model is selected by any criterion the investigator should be particularly cautious about using the associated outcome.[11] In the Netherlands the initial saturated capture-recapture model also gave unexpected and implausible high estimates of the number of unobserved tuberculosis patients. All three-source log-linear capture-recapture studies on tuberculosis and other infectious diseases incidence presented in Table 1.1 and 3.1 used independent or parsimonious log-linear models, apart from the capture-recapture study on meningitis incidence in the Netherlands by De Greeff et al. The only other published capture-recapture study using a saturated model was Chao's study on hepatitis A in Taiwan, discussed in chapter 10. Perhaps the validity of capture-recapture results reflects publication bias in favour of successful capture-recapture studies rather than the inherent strength of the methodology.[23] Stratification by relevant covariates associated with the probability of capture, to identify possible violation of the homogeneous population assumption, was not feasible. A truncated Poisson mixture model, arguably more robust in the presence of heterogeneity, estimated a considerably lower annual and total number of tuberculosis patients, with a smaller confidence interval. We can neither prove nor exclude that the estimates of the alternative truncated model portray a more accurate estimate of the true number of tuberculosis patients in England.

Comparison of the results of the three tuberculosis capture-recapture studies

Table 11.1 shows the completeness of three tuberculosis registers in the Netherlands, the Piedmont region in Italy and England, ascertained after record-linkage and estimated after capture-recapture analysis. To calculate the capture-recapture estimates in the Netherlands SPSS statistical software was used, in Italy S-PLUS software with the CARE library[24] and in England Stata software, but these different tools should produce similar results.

Notification:
After record-linkage initially the ascertained completeness of the notification registers is similar it but becomes higher in the Netherlands after the correction for possibly imperfect record-linkage and possibly imperfect positive predictive value of the hospital register, probably reflecting the good organisation of tuberculosis surveillance.

After capture-recapture analysis, the estimated completeness of the notification register is highest in the Netherlands and slightly lower in Italy. The estimated completeness of the notification register in England is highly inconsistent with the two other estimates.

Table 11.1 Completeness of thee tuberculosis registers in the Netherlands, Italy and England, ascertained after record-linkage and estimated after capture-recapture analysis

		Notification (%)	Laboratory (%)	Hospital (%)
Ascertained completeness	The Netherlands	86.6*	67.1	40.7
	Piedmont region, Italy	84.1	43.3	61.8
	England	84.1	54.3	41.6
Estimated completeness	The Netherlands	86.4	65.0	35.7
	Piedmont region, Italy	79.1	40.5	57.7
	England	56.2	36.2	27.7

* 89.9% for culture-confirmed and verified tuberculosis patients and 92.7% after correction for possibly imperfect record-linkage and possibly imperfect positive predictive value of the hospital register

Laboratory:
The ascertained completeness of the laboratory register in the Netherlands is high compared to the two other studies, indicating efforts to establish bacteriological confirmation of the diagnosis. The lowest proportion of bacteriologically confirmed tuberculosis patients is found in Italy, suggesting that fewer attempts are made to confirm the diagnosis and more patients are treated on empirical grounds.

After capture-recapture analysis, the estimated completeness of the laboratory register in the Netherlands and Italy do not change much but it strongly decreases in England due to the high estimated total number of tuberculosis patients.

Hospital:
In the Netherlands and England, the ascertained completeness of the hospital register is low, likely reflecting common policies of preferably treating tuberculosis patients as out-patients, including isolation at home for infectious patients. The high proportion of hospitalised tuberculosis patients in Italy suggests a policy of (initial) clinical analysis, diagnosis, treatment or isolation.

After capture-recapture analysis, the estimated completeness of the hospital register in the Netherlands and Italy slightly decrease but this effect is most profound in England due to the high estimated total number of tuberculosis patients.

Question 3:

What is the feasibility and validity of truncated population estimation models in infectious disease surveillance?

For priority setting, service planning and resource allocation it is necessary to know the number of persons in a targeted group. This number can also be used to assess the coverage of an intervention. Often direct (enumeration) techniques are not feasible to estimate the size of hidden populations. Instead, indirect techniques such as capture-recapture analysis have to be used. Paradoxically, for hidden populations often the preferred three linked registers, allowing for log-linear capture-recapture analysis in order to reduce bias in the estimates, are not available. As an alternative, truncated models, related to capture-recapture analysis, and applicable to frequency counts of observations of individuals in a single source of information, are described in the literature. Two such truncated models are Zelterman's Poisson mixture model and Chao's heterogeneity model.[24-26] These models aim to estimate the number of unobserved persons in the (truncated) zero-frequency class based upon information from the lower observed frequency classes, assuming a specific truncated distribution of the observed data, e.g. Poisson in Chao's model. Observed frequency distributions may not be strictly Poisson and to relax this assumption Zelterman based his model on a Poisson mixture distribution, allegedly allowing greater flexibility and applicability on real life data. The validity of truncated model estimates depends on the possible violation of the underlying assumptions, similar to capture-recapture analysis as described earlier, although the independent registers assumption is replaced by the constant (re)observation probability assumption when using a single data source. In addition, equiprobability (i.e. equal ascertainment probabilities of all registers) should be assumed when using multiple sources. This violation could be as much, possibly even more, as for capture-recapture analysis.[11]

We estimated the coverage of a tuberculosis control intervention, a targeted mobile tuberculosis screening programme among illicit drug users and homeless persons in Rotterdam. Application of truncated models was feasible because this screening programme uses a single register. Although capture-recapture techniques for estimating the size of a population from a single register have been described occasionally,[27-29] for feasibility one prefers the simplest technique with almost similar assumptions. We could extract, check and prepare the required data from the existing routine dataset in two days and calculate the point estimates on a pocket calculator. Violation of the perfect record-linkage assumption is considered minimal because of good computerised and visual identification of the clients in the screening programme. However, the closed population assumption is violated because every year a substantial number of people not previously screened enter the programme, resulting in under-estimation of the coverage. We cannot exclude heterogeneity among individuals belonging to the target group of the screening programme but this could cause limited bias in the model estimates. Truncated models are arguably more robust to violation of the homogeneity assumption since they are partly based upon the lower frequency classes, assumed to have more resemblance to the zero frequency class. The constant (re)observation probability assumption will be violated as well to some extent but this effect could be limited due to the nature and organisation of

the screening programme. Often cross-validation of the population estimates from truncated models is not possible. However, we could compare our estimates with an independent assessment of the number of problematic illicit drug users in Rotterdam in 2003 established through two-source capture-recapture analysis that used a similar case-definition of the target group. These capture-recapture estimates were comparable to those of the truncated models. In the context of its advantages and limitations, we conclude that the use of truncated population estimation models is a feasible and valid method for estimating the coverage of a public health intervention programme among hidden populations.

A more detailed study of the validity of truncated population estimation models in infectious disease surveillance compared the performance of some truncated population estimation models with three-source log-linear capture-recapture analysis using data from published and current capture-recapture studies on infectious disease incidence (chapter 10). This comparative research was triggered by the results of the studies described in chapter 6 and chapter 8, indicating that conventional three-source log-linear capture-recapture models sometimes break down and produce unexpected and implausible results. Solely relying on three-source capture-recapture analysis seemed inappropriate and we perceived that a tool is needed to cross-validate capture-recapture estimates for infectious disease incidence. Ideally, this would be a model robust to violation of all capture-recapture assumptions but such models do not exist. We used truncated models because they performed well when compared to log-linear capture-recapture analysis earlier.[30] The comparative research was feasible as truncated models are easy to apply and can be used on the data of three-source capture-recapture studies (but not vice-versa). A limitation of our approach is that the number of possible frequency classes is restricted to three, violating the "infinite number of sources" assumption for truncated Poisson models. The truncated models are at least as vulnerable to violation of the perfect positive predictive value, perfect record-linkage, closed population and independent registers assumptions as capture-recapture models. It has been argued that truncated models are more robust to heterogeneity among the patients than capture-recapture studies.[25,31] In contrast, the equiprobability assumption of the truncated models is almost certainly violated when using multiple-source data from capture-recapture studies. Therefore we introduced the coefficient of variation, a measure of variability in the coverages of the three data sources for capture-recapture studies, to estimate the error from the data. We conclude that, in the context of validity, for estimating infectious disease incidence and completeness of notification independent and parsimonious three-source log-linear capture-recapture models are preferable compared to the truncated models examined. When saturated models are selected as best-fitting 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. Possible failure in the saturated log-linear model or unidentified violation of the underlying assumptions should be suspected when the truncated models produce a (considerably) lower estimated number of infectious disease patients.

11.2 Some findings of this thesis in the context of surveillance of tuberculosis and other infectious diseases

Hospital episode statistics often are not a reliable source for record-linkage or capture-recapture analysis for infectious disease surveillance for several reasons. Firstly, it may be difficult to examine the hospital episode statistics dataset for multiple entries of one patient, reflecting transfers between wards counted as separate disease episodes during one uninterrupted stay in the hospital, and cleaning is time-consuming. It is also possible that day-care visits or out-patient visits erroneously appear as admissions in the hospital episode statistics. Secondly, the disease codes assigned to hospital episode statistics records can reflect the differential diagnoses upon admission, e.g. an observation for a presumed malaria episode or a diagnostic procedure for a specific disease, such as a bronchoscopy because of radiological abnormalities compatible with tuberculosis among other lung diseases, without subsequent confirmation of the diagnosis. Finally, the specificity can be reduced when the absence of specific disease codes causes a proportion of false-positive records, as reflected in chapter 5. For example, in the Netherlands still the ICD-9 codes are used for hospital episode statistics instead of the more recent, comprehensive and internationally used ICD-10 codes. But even in the presence of detailed and specific disease codes, as exist for tuberculosis, the proportion of false-positive records can be high, from 27% among all patients in a local tuberculosis hospital register in Liverpool, United Kingdom,[8] to possibly 62% among the unlinked patients in a national tuberculosis hospital register in the Netherlands (chapter 6) and certainly 80% among the unlinked patients in a regional tuberculosis hospital register in the Piedmont region, Italy (chapter 7). The logistic regression population mixture model described in chapter 8 estimated the proportion of false-positive records among the unlinked patients in a national tuberculosis hospital register in England to be 72%. Possible false-positive cases in hospital episode registers could explain why most of the published capture-recapture studies on tuberculosis are local or regional, involving a relatively small number of patients, as shown in Table 1.1. This allows for the hospital charts to be scrutinised for false-positive cases manually, although this is time-consuming and comes closer to counting than estimating patients. Only when the number of patients was too small for a local study, e.g. tuberculous meningitis patients, capture-recapture analysis was performed at the national level.[32]

The capture-recapture studies for malaria and tuberculosis in chapters 4, 6, 7 and 8 show that in addition to the linked notification and laboratory registers a limited number of patients were identified through the hospital register, possibly including a substantial number of false-positive cases. Only the capture-recapture study on the incidence of Legionnaires' disease described in chapter 5 found the majority of the cases through the hospital register. These five chapters of this thesis have shown that capture-recapture analysis, as a method to estimate infectious disease incidence and completeness of registration, is not the cheap, quick, simple and reliable method as once advocated. 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, 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. An example of an infectious disease surveillance system that routinely links notification data with laboratory data is the ETS in England and Wales. This type of record-linkage should ideally be web-based for a timely reflection of trends. An example of a web-based notification system is OSIRIS in the Netherlands but this system is not linked to laboratory reports.[13] For improving quality, completeness and timeliness of infectious disease surveillance, a web-based infectious disease surveillance system that routinely links notification data with laboratory data could essentially fulfil the qualities once attributed to capture-recapture analysis.

11.3 Conclusions and recommendations

Conclusions

- Infectious disease incidence capture-recapture analysis requires adequate knowledge of disease, patients and registrations.
- In capture-recapture analysis small variations in the quality of data and record-linkage can lead to highly variable outcomes. Therefore previous successful infectious disease capture-recapture studies cannot be repeated uncritically.
- Hospital episode statistics often contain many false-positive records, which, when not identified, lead to biased capture-recapture estimates.
- When categorical covariates associated with the probability of capture in an infectious disease register are present, covariate capture-recapture analysis can reduce bias as a result of heterogeneity.
- In the absence of a gold standard, truncated models can be used as a heuristic tool to identify possible failure in log-linear models, especially when saturated models are selected.

Recommendations

- Infectious diseases capture-recapture studies should be performed with a multi-disciplinary team including public health physicians, clinicians, statisticians and data managers.
- For more reliable record-linkage of notifiable infectious disease registers the Dutch Infectious Disease Act 1999 should be amended and provide recording the date of birth of the patients instead of the year of birth.

- To improve timeliness and completeness of infectious disease surveillance, web-based record-linkage between notifications and positive laboratory results should be implemented.
- The value of truncated models as an alternative to or for cross-validation of capture-recapture analysis for estimating infectious disease incidence should be further explored.

11.4 References

1. LaPorte RE. Assessing the human condition: capture-recapture techniques. *BMJ* 1994; 308: 5-6.
2. Joseph CA. Legionnaires'disease in Europe 2000-2002. *Epidemiol Infect* 2004; 132: 417-24.
3. Infectious Disease Act 7 July 1998. *Statute Book 394*. The Hague: Netherlands Government Printing Press, 1998.
4. Klein S, Bosman A. Completeness of malaria notification in the Netherlands 1995-2003 assessed by capture-recapture method. *Euro Surveill* 2005; 10: 244-6.
5. Jarvis SN, Lowe PL, Avery A, Levene S, Cormack RM. Children are not goldfish-mark/recapture techniques and their application to injury data. *Inj Prev* 2000; 6: 46-50.
6. Nardone A, Decludt B, Jarraud S, Etienne J, Hubert B, Infuso A, Gallay A, Desenclos JC. Repeat capture-recapture studies as part of the evaluation of the surveillance of Legionnaires' disease in France. *Epidemiol Infect* 2003; 131: 647-54.
7. Braun JJ, de Graaff CS, de Goey, Zwinderman AH, Petit PL. [Community-acquired pneumonia: pathogens and course in patients admitted to a general hospital]. *Ned Tijdschr Geneeskd* 2004; 148: 836-40.
8. Tocque K, Bellis MA, Beeching NJ, Davies PD. Capture-recapture as a method of determining the completeness of tuberculosis notifications. *Commun Dis Public Health* 2001; 4: 141-3.
9. Baussano I, Bugiani M, Gregori D, Van Hest R, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung Dis* 2006; 10: 415-21.
10. Davies AG, Cormack RM, Richardson AM. Estimation of injecting drug users in the City of Edinburgh, Scotland, and number infected with human immunodeficiency virus. *Int J Epidemiol* 1999; 28: 117-21.
11. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. *Epidemiol Rev* 1995; 17: 243-64.
12. Regal RR, Hook EB. Marginal versus conditional versus structural source models: a rationale for an alternative to log-linear methods for capture-recapture estimates. *Stat Med* 1998; 17: 69-74.
13. Hook EB, Regal RR. Accuracy of alternatives to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. *Am J Epidemiol* 2000; 152: 771-9.
14. Den Boer JW, Friesema IH, Hooi JD. [Reported cases of Legionnaires' disease in the Netherlands, 1987-2000]. *Ned Tijdschr Geneeskd* 2002; 46: 315-20.
15. Van Loenhout-Rooijackers JH, Leufkens HG, Hekster YA, Kalisvaart NA. Pyrazinamide use as a method to estimate underreporting of tuberculosis. *Int J Tuberc Lung Dis* 2001; 5: 1156-60.
16. World Health Organization. Global Tuberculosis Control: Surveillance, Planning, Financing. *WHO Report 2003*. Geneva: WHO, 2003.
17. Moro ML, Malfait P, Salamina G, D'Amato S. [Tuberculosis in Italy: available data and open questions]. *Epidemiol Prev* 1999; 23: 27-36.
18. Buiatti E, Acciai S, Ragni P, Tortoli E, Barbieri A, Cravedi B, Santini MG. [The quantification of tuberculous disease in an Italian area and the estimation of underreporting by means of record linkage]. *Epidemiol Prev* 1998; 22: 237-41.
19. Rose AM, Gatto AJ, Watson JM. Recent increases in tuberculosis notifications in England and Wales – real or artefact? *J Public Health Med* 2002: 24: 136-7.
20. Evans MR. Is tuberculosis taken seriously in the United Kingdom? *BMJ* 1995; 311: 1483-5.
21. Pillaye J, Clarke A. An evaluation of completeness of tuberculosis notification in the United Kingdom. *BMC Public Health* 2003; 1: 31.
22. De Greeff SC, Spanjaard L, Dankert J, Hoebe CJ, Nagelkerke N, De Melker HE. Underreporting of meningococcal disease incidence in the Netherlands: Results from a capture-recapture analysis based on three registration sources with correction for false-positive diagnoses. *Eur J Epidemiol* 2006; 21: 315-21.
23. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. *Statistician* 1997; 46: 515-20.

24. Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 1987; 43: 783-91.
25. Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. *J Statistic Plan Inf* 1988; 18: 225-37.
26. Chao A. Estimating animal abundance with capture frequency data. *J Wildl Manage* 1988; 52: 295-300.
27. Laska EM, Meisner M. A plant-capture method for estimating the size of a population from a single sample. *Biometrics* 1993; 49: 209-20.
28. Laska E, Lin S, Meisner M. Estimating the size of a population from a single sample: methodology and practical issues. *J Clin Epidemiol* 1997; 50: 1143-54.
29. Brugal MT, Domingo-Salvany A, Diaz de Quijano E, Torralba L. Prevalence of problematic cocaine consumption in a city in southern Europe, using capture-recapture with a single list. *J Urban Health* 2004; 81: 416-27.
30. Hook EB, Regal RR. Validity of Bernouilli census, log-linear and truncated binomial models for correcting for underestimates in prevalence studies. *Am J Epidemiol* 1982; 116: 168-76.
31. Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* 1992; 79: 543-53.
32. Cailhol J, Che D, Jarlier V, Decludt B, Robert J. Incidence of tuberculous meningitis in France, 2000: a capture-recapture analysis. *Int Tuberc Lung Dis* 2005; 9: 803-8.
33. Ward M, Brandsema P, Van Straten E, Bosman A. Electronic reporting improves timeliness and completeness of infectious disease notification, The Netherlands, 2003. *Euro Surveill* 2005; 10: 27-30.

Summary

Surveillance is an essential part of infectious disease control. A concern of any surveillance system is the quality of the data collected, including the degree of ascertainment of affected individuals. A conventional surveillance system is notification, but it may contain false-positive cases and is often incomplete for true-positive cases. Important for the assessment of the quality and completeness of infectious disease registers is record-linkage, i.e. comparing patient data across multiple registers. Completeness of notification can then be assessed through capture-recapture analysis, a technique originally developed for studies of animal abundance.

After a brief introduction to capture-recapture analysis **Chapter 1** describes aspects of tuberculosis under-notification, methods of estimating tuberculosis incidence and the application and limitations of capture-recapture methods in tuberculosis surveillance, followed by a summary of published capture-recapture studies in this field. This chapter then presents the research questions in this thesis: 1) How do the characteristics of various infectious diseases and their registers in the Netherlands influence the feasibility and validity of capture-recapture analysis; 2) How do the characteristics of tuberculosis surveillance systems in different countries influence the feasibility and validity of capture-recapture analysis and 3) What is the feasibility and validity of truncated population estimation models in infectious disease surveillance? **Chapter 2** describes the methodology of capture-recapture analysis, addresses the underlying assumptions and gives the mathematical framework. Alternative truncated population estimation models, related to capture-recapture analysis, are briefly mentioned. The chapter continues to describe the application and limitations of capture-recapture analysis in epidemiological studies and gives a stepwise overview of relevant issues to be addressed while planning, applying, presenting and evaluating capture-recapture techniques. In addition to a previous overview of published capture-recapture studies until 1997, **Chapter 3** presents a synopsis of capture-recapture studies on infectious diseases published between 1997 and 2006.

In the context of the first research question of this thesis **Chapter 4** describes a capture-recapture analysis after record-linkage of three malaria registrations to estimate the completeness of notification of malaria by physicians and laboratories in 1996. As for all studies estimating completeness of notification in this thesis, three conventional infectious disease registers were used: Notifications, Laboratory results and Hospital admissions. A parsimonious capture-recapture model, reducing bias due to interdependence between registers, estimated the total number of malaria patients at 774 (95% confidence interval (CI) 740-821) and the completeness of notification at 69.1% and 40.2% for the laboratories and physicians respectively. We conclude that laboratory-based notification can considerably increase the number of officially reported malaria cases in the Netherlands. In order to estimate the incidence and completeness of notification of Legionnaires' disease in 2000 and 2001, **Chapter 5** describes record-linkage and capture-recapture analysis of the three conventional registers. A saturated log-linear capture-recapture model estimated 1253 Legionnaires' disease patients (95%CI

1019-1715). To reduce possible bias due to heterogeneity among the patients, a covariate capture-recapture model was specified, i.e. a capture-recapture model including categorical covariates associated with the probability of capture in a register, because we expected and observed regional differences in the incidence rate of Legionnaires' disease. The covariate model including "region" as covariate estimated 886 Legionnaires' disease patients (95%CI 827-1022), resulting in an estimated completeness of notification of 42.1%. The notified, ascertained and estimated average annual incidence rates of Legionnaires' disease were 1.2, 2.4 and 2.8 per 100 000 inhabitants respectively but higher in the southern region of the Netherlands. We conclude that covariate capture-recapture analysis, acknowledging regional differences of Legionnaires' disease incidence, appears to reduce bias in the estimated national incidence rate in the Netherlands. In **Chapter 6** we describe a systematic process of record-linkage, cross-validation, case-ascertainment and capture-recapture analysis to assess the quality of tuberculosis registers and to estimate the completeness of notification of incident tuberculosis cases in 1998. A saturated log-linear capture-recapture model initially estimated an unexpectedly high number of 2053 (95%CI 1871-2443) tuberculosis cases, resulting in an estimated completeness of notification of 63.2%. After adjustment for possible imperfect record-linkage and remaining false-positive hospital cases a more parsimonious and better fitting capture-recapture model estimated 1547 (95%CI 1513-1600) tuberculosis cases, resulting in a completeness of notification of 86.4%. Truncated population estimators gave similar results. In this chapter we demonstrate the possible impact of violation of the perfect record-linkage and perfect positive predictive value assumptions on capture-recapture estimates.

In the context of the second research question of this thesis, after chapter 6 describing capture-recapture analysis to estimate completeness of notification of tuberculosis during one year at the national level in the Netherlands, in **Chapter 7** we estimate tuberculosis incidence and completeness of tuberculosis registration systems during one year at the regional level in the Piedmont Region of Italy. A parsimonious capture-recapture model estimated 704 (95%CI 688-728) tuberculosis patients, resulting in an estimated completeness of notification of 79.1%. The overall estimated tuberculosis incidence rate in the Piedmont Region is 16.7 cases per 100 000 population but varies between different subsets of the population. We conclude that when multiple recording systems are available, record-linkage can improve case-detection and capture-recapture analysis can be used to assess tuberculosis incidence and the completeness of notification, contributing to a more accurate surveillance of local tuberculosis epidemiology. In **Chapter 8** we estimated the completeness of tuberculosis notification in England at the national level for four years to assess the performance of the Enhanced Tuberculosis Surveillance (ETS) system, introduced in 1999. Due to the scale of this study (28 678 observed patients), for record-linkage of the hospitalised tuberculosis cases sophisticated record-linkage computer software was required and the proportion of false-positive cases among the unlinked hospital-derived tuberculosis records was estimated through a logistic regression population mixture model. According to a saturated capture-recapture model the estimated completeness of notification is 56.2%, highly inconsistent with prior estimates. A truncated population estimator, Zelterman's truncated Poisson-mixture model, estimated the completeness of notification at 79.5%. We conclude that record-

linkage of notification and laboratory registers, as performed in ETS, improves the accuracy of surveillance data as well as the completeness of case-ascertainment of tuberculosis. The validity of capture-recapture analysis, especially when the saturated model is selected, and truncated population estimation models, in the context of infectious disease surveillance, should be further examined.

In the context of the third research question of this thesis in **Chapter 9** we estimate the coverage of a periodic radiological mobile tuberculosis screening programme among illicit drug users and homeless persons in Rotterdam, using truncated population estimation models. We extracted the total and annual number and frequency counts of chest X-rays taken in this screening programme from a single data source. According to the two truncated models used, the tuberculosis screening programme reached approximately two-third of the estimated target population at least once per year and the coverage of the intended aim, at least two chest X-rays per person per year, was estimated at approximately 23%. We conclude that truncated models can be used relatively easily on available single source routine data to estimate the coverage of tuberculosis screening among illicit drug users and homeless persons. In **Chapter 10** we re-examine nineteen datasets of published and current three-source log-linear model capture-recapture studies on infectious disease incidence with three truncated models for incomplete count data: a binomial model, a Poisson mixture model and a Poisson heterogeneity model. Specific attention was given to the ratio between the number of clients registered once (f_1) and twice (f_2) and the kind of log-linear model selected. We discuss the (dis)agreement between the various estimates and possible violation of the underlying assumptions, especially the equiprobability assumption. We conclude that for estimating infectious disease incidence independent and parsimonious three-source log-linear capture-recapture models are preferable but truncated models can be used as a heuristic tool to identify possible failure in the log-linear model, especially when saturated models produce unexpectedly high and implausible estimates.

The General Discussion in **Chapter 11** reviews the research questions and the results of the studies in this thesis. It discusses aspects of the feasibility and validity of three-source log-linear capture-recapture analysis and related truncated models for estimating the incidence of tuberculosis and other infectious diseases. The conclusions and recommendations that follow from the research in this thesis are formulated and are described below.

Conclusions

- Infectious disease incidence capture-recapture analysis requires adequate knowledge of disease, patients and registrations.
- In capture-recapture analysis small variations in the quality of data and record-linkage can lead to highly variable outcomes.
- Hospital episode statistics often contain many false-positive records, leading to biased capture-recapture estimates.

- When categorical covariates associated with the probability of capture in an infectious disease register are present, covariate capture-recapture analysis can reduce bias as a result of heterogeneity.
- Truncated models can be used as a heuristic tool to identify possible failure in log-linear models, especially when saturated models are selected.

Recommendations

- Infectious diseases capture-recapture studies should be performed with a multi-disciplinary team.
- For more reliable record-linkage of notifiable infectious disease registers the Dutch Infectious Disease Act 1999 should be amended and provide recording the date of birth of the patients.
- To improve infectious disease surveillance, web-based record-linkage between notifications and positive laboratory results should be implemented.
- The value of truncated models for cross-validation of capture-recapture analysis for estimating infectious disease incidence should be further explored.

Samenvatting

Surveillance is een essentieel onderdeel van infectieziektebestrijding. Voor iedere vorm van surveillance is de kwaliteit van de verzamelde gegevens, waaronder de volledigheid van de rapportage van patiënten, van belang. Een gangbare vorm van surveillance zijn de aangiften maar dit systeem kan foutpositieve meldingen bevatten terwijl vaak niet alle patiënten die de aandoening wél hebben worden gerapporteerd. Het koppelen van verschillende bestanden, dat wil zeggen het vergelijken van de patiëntengegevens binnen verschillende registers die betrekking hebben op een bepaalde infectieziekte, is een belangrijke methode om inzicht te krijgen in de kwaliteit en de volledigheid van infectieziekteregistraties. De volledigheid van het aantal aangiften kan dan worden geschat met behulp van de vangst-hervangst methode, een techniek die oorspronkelijk werd ontwikkeld voor het schatten van de omvang van dierenpopulaties.

Na een korte inleiding over vangst-hervangst analyse bespreekt **Hoofdstuk 1** aspecten van de onderrapportage van tuberculose, verschillende methoden om de incidentie van tuberculose te schatten en de toepassing en de beperkingen van de vangst-hervangst methode op het gebied van tuberculose surveillance, gevolgd door een overzicht en samenvatting van gepubliceerde vangst-hervangst onderzoeken op dit gebied. Dit hoofdstuk noemt vervolgens de onderzoeksvragen die in dit proefschrift worden behandeld: 1) Hoe beïnvloeden de kenmerken van verschillende infectieziekten en hun registraties in Nederland de uitvoerbaarheid en betrouwbaarheid van de vangst-hervangst analyse; 2) Hoe beïnvloeden de kenmerken van verschillende surveillance systemen voor tuberculose in verschillende landen de uitvoerbaarheid en betrouwbaarheid van de vangst-hervangst analyse en 3) Hoe is de uitvoerbaarheid en betrouwbaarheid van "truncated" modellen voor populatieschattingen voor infectieziekten surveillance? **Hoofdstuk 2** bespreekt de methodologie van vangst-hervangst analyse, beschrijft de onderliggende aannames en geeft de wiskundige structuur. Alternatieve "truncated" modellen voor populatieschattingen, verwant aan de vangst-hervangst methode, worden kort benoemd. Het hoofdstuk bespreekt vervolgens de toepassing en de beperkingen van vangst-hervangst analyse voor epidemiologisch onderzoek en geeft een gestructureerd overzicht van relevante aandachtspunten bij het voorbereiden, toepassen, presenteren en evalueren van vangst-hervangst onderzoeken. In aanvulling op een eerder overzicht van vangst-hervangst onderzoeken gepubliceerd voor 1997 geeft **Hoofdstuk 3** een overzicht en samenvatting van de vangst-hervangst studies op het gebied van infectieziekten, gepubliceerd tussen 1997 en 2006.

In het kader van de eerste onderzoeksvraag van dit proefschrift beschrijft **Hoofdstuk 4** een vangst-hervangst analyse na koppeling van drie malaria registers om de volledigheid van het aantal aangiften door artsen en laboratoria in 1996 te schatten. Zoals voor alle onderzoeken in dit proefschrift die de volledigheid van het aantal aangiften schatten werd gebruikt gemaakt van drie gangbare infectieziekteregistraties, namelijk het aangifteregister, een laboratoriumuitslagenregister en het ziekenhuisopnameregister. Een spaarzaam vangst-hervangst model, dat verstoring van de uitkomst door onderlinge

afhankelijkheid van registers vermindert, schatte het totale aantal malaria patiënten op 774 (95% betrouwbaarheidsinterval (BI) 740-821) en de volledigheid van het aantal aangiften op 69.1% en 40.2% voor respectievelijk de laboratoria en de artsen. Wij concluderen dat het aantal officieel aangegeven malariapatiënten in Nederland aanzienlijk kan toenemen na invoering van melding door het laboratorium. Om de incidentie en de volledigheid van het aantal aangiften van *Legionella* pneumonie (veteranenziekte) in 2000 en 2001 te schatten, beschrijft **Hoofdstuk 5** de koppeling en de vangst-hervangst analyse van de gebruikelijke drie registers. Een verzadigd log-lineair vangst-hervangst model schatte 1253 (95%BI 1019-1715) patiënten met de veteranenziekte. Het vermoeden bestond dat deze schatting verstoord werd door verwachte en geïdentificeerde regionale verschillen in de incidentie ratio van de veteranenziekte. Om deze verstoring te beperken werd tevens een covariaat vangst-hervangst model gespecificeerd, dat wil zeggen een vangst-hervangst model dat één of meerdere categorische covariaten bevat die invloed hebben op de kans om in een bepaald bestand geregistreerd te worden, in dit geval de covariaat "regio". Het resultaat was een schatting van 886 (95%BI 827-1022) patiënten met de veteranenziekte, ofwel een geschatte volledigheid van het aantal aangiften van 42.1%. De incidentie ratio's op grond van het aantal aangegeven, het aantal geïdentificeerde en het aantal geschatte patiënten met de veteranenziekte bedragen respectievelijk 1.2, 2.4 and 2.8 per 100 000 inwoners, maar zijn hoger in de zuidelijke regio van Nederland. Wij concluderen dat covariate vangst-hervangst analyse, rekening houdend met de regionale verschillen in incidentie van de veteranenziekte, de verstoring in de schatting van de landelijke incidentie ratio in Nederland lijkt te beperken. In **Hoofdstuk 6** beschrijven we een systematisch proces van koppeling van bestanden, kruisvalidatie, identificatie van patiënten en vangst-hervangst analyse om inzicht te krijgen in de kwaliteit van tuberculoseregisters en om de volledigheid van het aantal aangiften van tuberculose in 1998 te schatten. Een verzadigd log-lineair vangst-hervangst model schatte aanvankelijk het totaal aantal tuberculose-patiënten onverwacht hoog op 2053 (95%BI 1871-2443), ofwel een geschatte volledigheid van het aantal aangiften van 63.2%. Na de correctie voor een mogelijk niet helemaal juiste koppeling van de bestanden en het mogelijk nog aanwezig zijn van enkele foutpositieve dossiers in het ziekenhuisbestand, werd een meer spaarzaam en beter passend vangst-hervangst model geselecteerd, dat het aantal tuberculosepatiënten op 1547 (95%BI 1513-1600) schatte, ofwel een geschatte volledigheid van het aantal aangiften van 86.4%. "Truncated" modellen voor populatieschattingen gaven vergelijkbare resultaten. In dit hoofdstuk laten we de mogelijke invloed zien van schending van de aannames van perfecte koppeling en perfecte positief voorspellende waarde van registers op de uitkomsten van vangst-hervangst analyse.

Nadat hoofdstuk 6 een vangst-hervangst analyse heeft beschreven om de volledigheid van het aantal aangiften van tuberculose te schatten op landelijk niveau gedurende één jaar in Nederland, schatten wij in **Hoofdstuk 7,** in het kader van de tweede onderzoeksvraag van dit proefschrift, de incidentie van tuberculose en de volledigheid van het aantal aangiften op regionaal niveau gedurende één jaar in de regio Piemonte in Italië. Een spaarzaam vangst-hervangst model schatte 704 (95%BI 688-728) tuberculosepatiënten, resulterend in een geschatte volledigheid van het aantal aangiften van 79.1%. De totale geschatte regionale tuberculose incidentie ratio in Piemonte is 16.7 patiënten per 100 000 inwoners (95% BI 16.3-17.3) maar deze varieert voor verschillende

subgroepen binnen de bevolking. Wij concluderen dat wanneer meerdere tuberculoseregistraties beschikbaar zijn, het koppelen van deze bestanden het aantal vastgestelde patiënten met tuberculose kan verhogen en dat de vangst-hervangst analyse gebruikt kan worden om de incidentie van tuberculose en de volledigheid van het aantal aangiften te schatten, hetgeen bijdraagt aan een meer nauwkeurige surveillance van de locale epidemiologie van tuberculose. In **Hoofdstuk 8** hebben we de volledigheid van het aantal aangiften van tuberculose geschat in Engeland op landelijk niveau gedurende vier jaar om de resultaten van het Enhanced Tuberculosis Surveillance (ETS) systeem te beoordelen dat in 1999 werd ingevoerd. Vanwege de omvang van dit onderzoek, 28 678 vastgestelde patiënten, moest gebruik gemaakt worden van hoogwaardige computer software om de tuberculosepatiënten uit het ziekenhuisopnameregister te koppelen aan de andere bestanden en het percentage foutpositieve dossiers onder de niet gekoppelde meldingen in het ziekenhuisopnameregister moest geschat worden met behulp van een logistisch regressie mengpopulatie model. Een verzadigd log-lineair vangst-hervangst model schatte de volledigheid van het aantal aangiften op 56.2%, hetgeen in sterke tegenspraak is met eerdere schattingen. Een "truncated" model voor populatie-schattingen, Zelterman's truncated Poisson-mixture model, schatte de volledigheid van het aantal aangiften op 79.5%. Wij concluderen dat de koppeling van de aangifte en laboratoriumregisters, zoals verricht in het ETS systeem, de nauwkeurigheid van de surveillance gegevens en de volledigheid van het aantal vastgestelde tuberculosegevallen verbetert. De betrouwbaarheid van de vangst-hervangst analyse, zeker als het verzadigde model wordt geselecteerd, alsmede de "truncated" modellen voor populatieschattingen, in de context van infectieziekten surveillance, dient verder onderzocht dient te worden.

In het kader van de derde onderzoeksvraag van dit proefschrift, hebben we in **Hoofdstuk 9** de dekkingsgraad geschat van een mobiel periodiek röntgenologisch screeningsprogramma voor tuberculose onder harddrugsverslaafden en dak- en thuislozen in Rotterdam, met behulp van "truncated" modellen voor populatieschattingen. Wij extraheerden het totale en jaarlijkse aantal thoraxröntgenfoto's gemaakt in dit screeningsprogramma, evenals de frequentieverdeling, uit één gegevensbestand. Volgens de twee gebruikte "truncated" modellen bereikte het screeningsprogramma voor tuberculose ongeveer tweederde van de geschatte doelgroep tenminste één maal per jaar en werd de dekkingsgraad van het beoogde doel, tenminste twee thoraxröntgenfoto's per persoon per jaar, geschat op circa 23%. We concluderen dat "truncated" modellen relatief eenvoudig kunnen worden toegepast op routinematig verzamelde gegevens in een enkele informatiebron voor het schatten van de dekkingsgraad van een mobiel screenings-programma voor tuberculose onder harddrugsverslaafden en dak- en thuislozen. In **Hoofdstuk 10** onderzoeken we 19 gegevensbestanden van gepubliceerde en lopende driebrons log-lineaire vangst-hervangst studies op het gebied van infectieziekte-incidentie opnieuw met drie "truncated" modellen die kunnen worden toegepast op onvolledige tellingen: een binomiaal model, een Poisson-mixture model en een Poisson-heterogeneity model. Specifieke aandacht werd gegeven aan de verhouding tussen het aantal cliënten dat één keer was geobserveerd (f_1) en het aantal cliënten dat twee keer was geobserveerd (f_2) en aan het soort log-lineaire model dat werd geselecteerd. We concluderen dat voor het schatten van infectieziekte-incidentie de voorkeur uitgaat naar onafhankelijke en spaarzame driebrons log-lineaire vangst-hervangst modellen maar dat "truncated"

modellen gebruikt kunnen worden als heuristisch instrument om het mogelijke falen van een log-lineaire model te onderkennen, in het bijzonder wanneer verzadigde modellen onverwachte en onwaarschijnlijk hoge schattingen geven.

De Discussie in **Hoofdstuk 11** bespreekt de onderzoeksvragen en de bevindingen van de onderzoeken opgenomen in dit proefschrift. Zij bediscussieert verschillende aspecten van de uitvoerbaarheid en de betrouwbaarheid van driebrons log-lineaire vangst-hervangst analyse en gerelateerde "truncated" modellen voor het schatten van de incidentie van tuberculose en andere infectieziekten. Tenslotte worden de conclusies en aanbevelingen van dit proefschrift geformuleerd, zoals hieronder nogmaals beschreven.

Conclusies

- Vangst-hervangst analyse van infectieziekte-incidentie vereist geëigende kennis van ziekte, patiënten en registraties.
- Bij vangst-hervangst analyse kunnen kleine variaties in de kwaliteit van de gegevens en de koppeling van de gegevensbestanden leiden tot grote verschillen in de omvangschattingen.
- Ziekenhuisopnameregisters bevatten vaak veel foutpositieve data die de betrouwbaarheid van een vangst-hervangst schatting verstoren.
- Wanneer categorische co-variabelen aanwezig zijn die de kans kunnen beïnvloeden op het wel of niet bekend zijn in een infectieziekteregister, lijkt covariate vangst-hervangst analyse de verstoring van de resultaten ten gevolge van heterogeniciteit te beperken.
- "Truncated" modellen kunnen dienen als heuristisch instrument om het mogelijke falen van een log-lineair model te onderkennen, in het bijzonder wanneer verzadigde modellen zijn geselecteerd.

Aanbevelingen

- Vangst-hervangst studies op het gebied van infectieziekten dienen te worden uitgevoerd door een multidisciplinaire groep onderzoekers.
- Voor een meer betrouwbare koppeling van infectieziekteregisters dient de Infectieziektewet uit 1999 te worden aangepast en het registreren van de volledige geboortedatum weer verplicht gesteld.
- Om de kwaliteit van infectieziekten surveillance te verbeteren dient het koppelen van de aangiften en positieve laboratoriumgegevens via Internet te worden gerealiseerd.
- De waarde van zogenaamde "truncated" modellen voor de validatie van vangst-hervangst schattingen betreffende de incidentie van infectieziekten dient verder onderzocht te worden.

Acknowledgements

When I returned to the Netherlands after working in Ghana for five years and studying in London, United Kingdom, for one year I realised that I was one of the few hard core members of the Nijmegen Medical School soul band The Booze Brothers who was not having or completing a PhD. As one of the front men this was intolerable of course and I gradually started contemplating on an opportunity.

I thank my supervisor professor Dik Habbema for giving me the opportunity to obtain a PhD and his, often one-line, advices that ultimately changed the structure of the thesis, and my supervisor Jan Hendrik Richardus for coaching me through the final years, especially at moments when my battery was low and needed recharging.

I thank the professors Gouke Bonsel, Martien Borgdorff and Bert Hofman and professors Arno Hoes and Henk Hoogsteden for their willingness to be a member of the small and large doctoral degree committee respectively.

When in statistics an outcome can be caused by coincidence it means that it is not significant. However, this thesis demonstrates that coincidence can be very relevant. It started when, by coincidence, I was asked by Jan Peter Verhave of the Department of Parasitology of the University Medical Centre in Nijmegen to complete a study on under-notification of malaria in the Netherlands. I thank Jan Peter for his enthusiasm during this study that resulted in my first publication and substantiated the plan to write a full PhD thesis on capture-recapture analysis. Due to initial methodological problems in the malaria study, by coincidence, a member of an Internet malaria discussion group suggested I should try capture-recapture analysis, a technique completely unknown to me. Again by coincidence I was referred to Filip Smit at the Trimbos Institute in Utrecht who had experience in performing capture-recapture studies in the field of social sciences and he altruistically offered to help me. This was the start of an on-off co-operation over five years, that resulted in four chapters in this thesis, for which I am deeply grateful. In the mean time I had met Christian Hoebe during our training in Public Health and we got involved in our mutual PhD plans. I thank him for all his support, especially while performing the capture-recapture study on Legionnaires' disease incidence in the Netherlands. At a tuberculosis course in Liverpool, United Kingdom, by coincidence I met Iacopo Baussano from the University of Turin in Italy. Iacopo needed a subject for an MSc thesis and in Liverpool the idea was born for a capture-recapture study on tuberculosis incidence in the Piedmont Region in Italy. I thank Iacopo for all the work he and his Italian colleagues performed in Turin and for a long-term friendship that resulted from this research. A few months later, and again by coincidence, on the World Tuberculosis Conference in Paris, France, I met Alistair Story from the Tuberculosis Section of the Respiratory Diseases Department at the Centre for Infections of the Health Protection Agency in London. It must have been some kind of fine ale pheromones that identified the two of us as tuberculosis control mavericks. It was the start of an intense co-operation that resulted in the realisation of the targeted mobile digital X-ray tuberculosis screening project for homeless persons, illicit drug users and

prisoners in London, a huge capture-recapture study on tuberculosis incidence in England, an editorial in the BMJ, a few "quick ones" and last orders and numerous of the weirdest moments of my life. I would like to thank Alistair and Melissa for all the roasted potatoes, carrots, parsnips, Brussels sprouts and other un-Continental foodstuff from the traditional English cuisine, the regular Indian and Thai lamb curries and the many weeks of hospitality in Berkhamsted over the past five years.

I thank Gerard de Vries and Hennie Baars, colleague tuberculosis control physicians in the Tuberculosis Control Section of the Municipal Public Health Service Rotterdam-Rijnmond, for their help in preparing some of the studies in this thesis and other international and national publications.

I thank Dick van Soolingen, Petra de Haas, Kristin Kreemer, Tridia van der Laan, Anne-Marie van der Brandt and Mimount Enaimie from the National Mycobacteria Reference Unit at the National Institute of Public Health and the Environment in Bilthoven for their hospitality and assistance when performing the record-linkage and cross-validation of the Dutch tuberculosis registers and the always pleasant atmosphere.

I thank John Watson, Delphine Antoine, David Quinn, Charlotte Anderson, Clare French and Jonathan Crofts from the Tuberculosis Section at the Centre for Infections of the Health Protection Agency in London for their assistance, advice and hospitality while performing the capture-recapture study on tuberculosis incidence in England and various good discussions in the Holly Bush public house. A special thanks goes to Andrew Grant from the Statistics, Modelling and Bioinformatics Department at the Centre for Infections for his indispensable statistical support and interest in the capture-recapture methodology which has also initiated ideas for other papers.

I thank Nico Kalisvaart of the KNCV Tuberculosis Foundation in the Hague for performing the record-linkage with the Netherlands Tuberculosis Register for me, essential to the solution of the Dutch tuberculosis capture-recapture problem.

I thank professor Nico Nagelkerke of the United Arab Emirates University and professor Jeroen Vermunt of the University of Tilburg for providing me with the solution to the capture-recapture problems in the studies on tuberculosis and Legionnaires' disease incidence in the Netherlands respectively.

I thank all the other co-authors of the chapters in this thesis for their contribution and I hope it may lead to further collaboration in the future.

I thank Len Munnik for making the beautiful and fitting drawing on the cover and Selma Nieukoop and Diane Seinstra for helping with the lay-out.

I thank the Department of Infectious Disease Control of the Municipal Public Health Service Rotterdam-Rijnmond for granting me some time and funds to perform the work for this PhD thesis. I would like to thank all my colleagues in the Tuberculosis Control Section for their patience with me in times when my mood obviously reflected the progress of some of the studies.

Finally I thank Gerard de Vries and Bert Mulder for their willingness to be my ushers and their help in preparing all the ceremonial and unceremonious activities.

Curriculum vitae

Rob van Hest was born on October 17, 1958 in Tilburg, the Netherlands. After passing the Atheneum B exam at the Cobbenhagen College in Tilburg in 1977 he studied psychology for one year at the University of Tilburg. In 1978 he started his medical studies at the University of Nijmegen. During the junior house officer rotations he spent 5 months as a student doctor at St. Anthony's Hospital in Dzodze, Volta Region, Ghana, for the elective "Medical care in developing countries". After obtaining his medical degree in 1987 he initially worked for the Ambulance Service West-Betuwe in Tiel. From 1988 until 1991, in the context of the tropical doctor training, he was a senior house officer in the St Anna Hospital in Oss (Internal Medicine and Surgery) and the IJsselland Hospital in Rotterdam (Obstetrics and Gynaecology). Afterwards he completed the Course in Tropical Medicine for Doctors at the Royal Tropical Institute in Amsterdam. Between 1991 and 1996 he worked as medical officer in-charge at the Mary-Theresa Hospital in Dodi Papase, Volta-Region, Ghana. In 1997 he obtained the Master of Science degree in Control of Infectious Diseases at the London School of Hygiene and Tropical Medicine, London, United Kingdom. In 1998 he began his training as a public health physician (tuberculosis control) at the Netherlands School of Public Health and his work as a tuberculosis control physician at the Tuberculosis Control Section of the Municipal Public Health Service Rotterdam-Rijnmond and later also at the Tuberculosis Control Section of the Municipal Health Service Zuid-Holland Zuid in Dordrecht. In 2000 he obtained the Master of Science degree in Epidemiology and registration as epidemiologist-A from the Dutch Society for Epidemiology. After completing his training in public health medicine in December 2000, he was registered as a public health physician. He continued to work as a tuberculosis control physician at the Municipal Public Health Service Rotterdam-Rijnmond where, in 2002, he officially started this PhD part-time (0.1 FTE) in collaboration with the Department of Public Health of the Erasmus MC, University Medical Centre Rotterdam, in Rotterdam. Since 2002 he is also a visiting tuberculosis control physician with the mobile digital tuberculosis screening project among homeless persons, problematic drug users, alcoholics and prisoners in London, where he is registered as a specialist in Public Health Medicine as well. Apart from this thesis he has published various peer-reviewed articles on different aspects of tuberculosis and tuberculosis control in international and Dutch journals.

Curriculum vitae

Rob van Hest werd geboren op 17 oktober 1958 te Tilburg waar hij het diploma Atheneum B behaalde aan het Cobbenhagen College. Hij studeerde een jaar psychologie aan de Universiteit van Tilburg alvorens te starten met de studie Geneeskunde aan de Universiteit van Nijmegen. Tijdens deze studie verbleef hij 5 maanden in het St. Anthony's Hospital te Dzodze, Volta Region, Ghana, in het kader van het keuze co-assistentschap "Medical care in developing countries". Na het artsexamen in 1987 werkte hij eerst bij de Ambulancedienst West-Betuwe in Tiel. Tussen 1988 en 1991 was hij arts-assistent Interne Geneeskunde en Chirurgie in het St. Anna Ziekenhuis te Oss en arts-assistent Verloskunde en Gynaecologie in het IJsselland Ziekenhuis in Rotterdam in het kader van de opleiding tot tropenarts. Vervolgens voltooide hij de Nationale Tropencursus voor Artsen en behaalde het Diploma Tropische Gezondheidszorg aan het Koninklijk Instituut voor de Tropen in Amsterdam. Tussen 1991 en 1996 werkte hij als medical officer in-charge in het Mary-Theresa Hospital in Dodi Papase, Volta-Region, Ghana. In 1997 behaalde hij de Master of Science graad in Control of Infectious Diseases aan de London School of Hygiene and Tropical Medicine te Londen, Verenigd Koninkrijk. In 1998 begon hij zijn opleiding Sociale Geneeskunde (bijzondere tak tuberculosebestrijding) aan de Netherlands School of Public Health en zijn werkzaamheden als tuberculosearts op de afdeling Tuberculosebestrijding van de Gemeentelijke Gezondheidsdienst (GGD) Rotterdam-Rijnmond, en later ook op de afdeling Tuberculosebestrijding van de GGD Zuid-Holland Zuid in Dordrecht. In 2000 behaalde hij de Master of Science graad in de Epidemiologie en werd hij ingeschreven als epidemioloog-A bij de Vereniging voor Epidemiologie. Eind 2000 werd hij geregistreerd als sociaal-geneeskundige en bleef als tuberculosearts werken bij de GGD Rotterdam-Rijnmond. Daar begon hij in 2002 officieel aan dit proefschrift, parttime (0.1 FTE), in samenwerking met de afdeling Maatschappelijke Gezondheidszorg van het Erasmus Universitair Medisch Centrum in Rotterdam. Sinds 2002 is hij ook bezoekend tuberculosearts bij het mobiele digitale screeningsprogramma op tuberculose onder dak- en thuislozen, problematische drugsgebruikers, alcoholisten en gedetineerden in Londen, waar hij eveneens geregistreerd is als sociaal-geneeskundige. Naast dit proefschrift publiceerde hij diverse artikelen over verschillende aspecten van tuberculose en tuberculosebestrijding in internationale en Nederlandse tijdschriften.

Publications

This thesis

Van Hest NA, Smit F, Verhave JP. Improving malaria notification in the Netherlands: results from a capture-recapture study. *Epidemiol Infect* 2002; 129: 371-7.

Baussano I, Bugiani M, Gregori D, **Van Hest R**, Borracino A, Raso R, Merletti F. Undetected burden of tuberculosis in a low-prevalence area. *Int J Tuberc Lung* Dis 2006; 10: 415-21.

Van Hest NA, Smit F, Baars HW, De Vries G, De Haas P, Westenend PJ, Nagelkerke N, Richardus JH. Completeness of registration of tuberculosis in the Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiol Infect* 2006; Published on-line: 7 December 2006; doi:10.1017/S0950268806007540.

Van Hest NA, Grant D, Smit F, Story A, Richardus JH. Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data. *Epidemiol Infect* 2007; Published on-line: 11 March 2007; doi:10.1017/S0950268807008254.

Van Hest NA, Hoebe CJ, Den Boer JW, Vermunt JK, IJzerman EP, Boersma WG, Richardus JH. Incidence and completeness of notification of Legionnaires' disease in the Netherlands: covariate capture-recapture analysis acknowledging geographical differences. *Epidemiol Infect* 2007 (in press)

Van Hest NA, De Vries G, Smit F, Grant AG, Richardus JH. Estimating the coverage of tuberculosis screening among drug users and homeless persons with truncated models. *Epidemiol Infect* 2007 (in press)

Other peer-reviewed indexed publications

Van Hest NA, Smit F, Verhave JP. [Considerable underreporting of malaria in the Netherlands; a capture-recapture analysis]. *Ned Tijdschr Geneeskd* 2001; 145: 175-9.

Van Hest NA, De Vries G, Van Gerven PJ, Baars HW. [Delay in the diagnosis of tuberculosis]. *Ned Tijdschr Geneeskd* 2003; 147: 1825-9.

Van Hest R, Baars H, Kik S, van Gerven P, Trompenaars M-C, Kalisvaart N, Keizer S, Borgdorff M, Mensen M, Cobelens F. Hepatotoxicity of Rifampin-Pyrazinamide and Isoniazid preventive therapy and tuberculosis treatment. *Clin Infect Dis* 2004; 39: 488-96.

Van Hest R, De Vries G, Morbano G, Pijnenburg M, Hartwig N, Baars H. Cavitating tuberculosis in an infant: case report and literature review. *Pediatr Infect Dis J* 2004; 23: 667-70.

Van Hest R, Van der Zanden A, Boeree M, Kremer K, Dessens M, Westenend P, Maraha B, Van Uffelen R, Schütte R, De Lange W. *Mycobacterium heckeshornense* infection in an immunocompetent patient and identification by 16S rDNA sequencing of culture and histopathology tissue specimen. *J Clin Microbiol* 2004; 42: 4386-9.

De Vries G, Van Altena R, Van Soolingen D, Broekmans JF, **Van Hest NA**. [An outbreak of multiresistant tuberculosis from Eastern Europe in the Netherlands]. *Ned Tijdschr Geneesk*d 2005; 35: 1921-4.

De Vries G, **Van Hest R**. From contact investigation to tuberculosis screening of drug addicts and homeless persons in Rotterdam. *Eur J Public Health* 2006; 16: 133-6.

Story A, **Van Hest R**, Hayward A. Tuberculosis and social exclusion. *BMJ* 2006; 333: 57-8.

De Vries G, **Van Hest RA**, Richardus JH. Impact of mobile radiographic screening on tuberculosis among drug users and homeless persons. *Am J Respir Crit Care Med* 2007; Published on-line 5 April 2007; doi:10.1164/rccm.200612-1877OC

Other non-indexed publications:

De Vries G, **Van Hest NAH**, Van Bergen JEAM, eds. *Partnership in International Health*. 90th Anniversary Jubilee Book Dutch Society for Tropical Medicine. Het Spinhuis, Amsterdam, 1998, 181 pp.

Verhave JP, **Van Hest NAH**. [Laboratory diagnosis of malaria in the tropics: ideal and reality]. *Analyse* 2000; 52: 261-5.

Verhave JP, **Van Hest NAH**. [Import and registration of malaria in the Netherlands]. *Pharmaceutisch Weekblad* 2000; 135: 138-41.

Van Hest NAH. [A treacherous travel companion]. *Medisch Contact* 2001; 56: 468-70.

Van Hest NAH, Van den Kerkhof JHTC. [Quantitative aspects of malaria notification in the Netherlands]. *Infectieziekten Bulletin* 2001; 12: 122-4.

Verhave JP, **Van Hest NAH**. [General practitioner and malaria]. *Huisarts en Wetenschap* 2002; 45: 22-6.

Van Hest NAH, Bosman H. [Radiology pearls I]. *Tegen de Tuberculose* 2002; 98: 24-5.

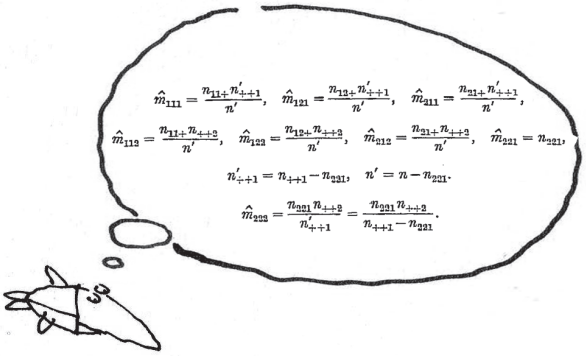
Van Hest NAH, Sonneveld J, De Lange W. [An atypical story]. *Tegen de Tuberculose 2002*; 98: 53-6.

Van Hest NAH. [Radiology pearls II]. *Tegen de Tuberculose* 2002; 98: 85.

De Vries G, **Van Hest NAH**, Šebek MMGG. [Tuberculosis among illicit drug users and homeless persons in Rotterdam]. *Infectieziekten Bulletin* 2003; 14: 362-5.

Van Hest R, Van Altena R, Arend SM, Baars H, Van Loenhout J, Hartwig N. [The diagnosis and treatment of latent tuberculosis infection in children]. *Tijdschr Kindergeneeskd* 2006; 74: 21-9.

Van Hest NAH. [Radiology pearls III]. *Tegen de Tuberculose* 2006; 102: 10-11.



$$\hat{m}_{111} = \frac{n_{11+} n'_{++1}}{n'}, \quad \hat{m}_{121} = \frac{n_{12+} n'_{++1}}{n'}, \quad \hat{m}_{211} = \frac{n_{21+} n'_{++1}}{n'},$$

$$\hat{m}_{112} = \frac{n_{11+} n_{++2}}{n'}, \quad \hat{m}_{122} = \frac{n_{12+} n_{++2}}{n'}, \quad \hat{m}_{212} = \frac{n_{21+} n_{++2}}{n'}, \quad \hat{m}_{221} = n_{221},$$

$$n'_{++1} = n_{++1} - n_{221}, \quad n' = n - n_{221}.$$

$$\hat{m}_{222} = \frac{n_{221} n_{++2}}{n'_{++1}} = \frac{n_{221} n_{++2}}{n_{++1} - n_{221}}.$$