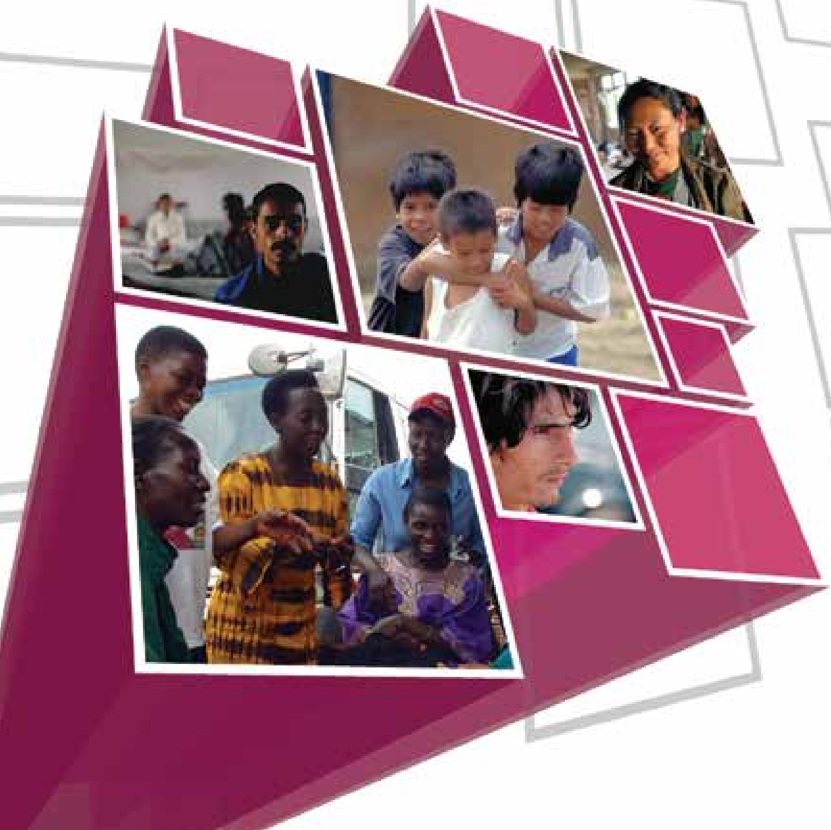
ASSESSING tuberculosis UNDER-REPORTING THROUGH inventory studies



World Health
Organization

ASSESSING tuberculosis UNDER-REPORTING THROUGH inventory studies

World Health Organization

WHO Library Cataloguing-in-Publication Data

Assessing tuberculosis under-reporting through inventory studies.

1.Tuberculosis - epidemiology. 2.Disease notification – methods. 3.Health surveys. 4.Population surveillance. I.World Health Organization.

ISBN 978 92 4 150494 2 (NLM classification: WF 205)

The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of the World Health Organization concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers' products does not imply that they are endorsed or recommended by the World Health Organization in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by the World Health Organization to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall the World Health Organization be liable for damages arising from its use.

Printed in France
WHO/HTM/TB/2012.12

Cover images
Bottom left: ©WHO/Gary Hampton; all other images: ©WHO/Davenport.

Contents

Acknowledgements

Assessing tuberculosis under-reporting through inventory studies was prepared by a core writing team of Ibrahim Abubakar,[1] Amal Bassili,[2] Ana Bierrenbach,[3] Emily Bloss,[4] Katherine Floyd,[5] Philippe Glaziou,[5] Ross Harris,[1] Knut Lonnroth,[5] Fulvia Mecatti,[6] Alexandra Pavli,[5] Charalambos Sismanidis,[5] Hazim Timimi,[5] Mukund Uplekar,[5] and Rob van Hest.[7] The group was coordinated by Philippe Glaziou with broad guidance from Katherine Floyd.

Professional affiliations:

1. Health Protection Agency, United Kingdom of Great Britain and Northern Ireland (UK)
2. World Health Organization Regional Office for the Eastern Mediterranean Region, Cairo, Egypt
3. Freelance Consultant
4. Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America (USA)
5. World Health Organization Headquarters, Geneva, Switzerland
6. University of Milan-Bicocca, Italy
7. Metropolitan Public Health Service Rotterdam-Fijnmond, the Netherlands

In addition to the core writing team, we thank Neal Alexander (Tropical Epidemiology Group, London School of Tropical Medicine and Hygiene, UK), Dankmar Böhning (Southampton Statistical Research Science Institute and Department of Mathematics, Southampton University, UK), Sian Floyd (Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, UK) and Andrew Thompson (Medicines and Healthcare products Regulatory Agency, London, UK) for their careful reviews of and input to the chapters on study design (Chapter 2) and data analysis and reporting (Chapter 5). We thank Andreas Reis (WHO) for reviewing content on ethical aspects of inventory studies. We also thank Eugene McCray and Phil Lobue (Centers for Disease Control and Prevention, USA) and Akira Shimouchi (Japanese Anti-Tuberculosis Association) for their careful reviews of the entire document.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of any of the agencies listed above. Any remaining errors remain the responsibility of the core writing team.

Funding for the development of the guide was provided by the United States Agency for International Development (USAID), including through the USAID Tuberculosis CARE I, Cooperative Agreement No. AID-OAA-A-10-000020. Funding was also provided by the Bill & Melinda

Gates Foundation.

The document was edited by Karen Ciceri (WHO), and its layout and cover designed by Rania Spatha (consultant).

The authors declared no conflict of interest, financial or personal, in preparing this publication.

Introduction

Authors: Philippe Glaziou, Alex Pavli, Emily Bloss, Mukund Uplekar, Katherine Floyd

The World Health Organization (WHO) declared tuberculosis (TB) a global public health emergency in 1993, when an estimated 7–8 million cases and 1.3–1.6 million deaths occurred each year [1]. In 2011, there were 8.3–9.0 million cases and 1.3–1.6 million deaths from TB. Despite the availability of treatment that can cure over 90% of cases, TB remains the second leading cause of death from an infectious disease worldwide after the human immunodeficiency virus (HIV), which caused an estimated 1.7 million deaths in 2011 [2].

Global targets for reducing the burden of disease caused by TB have been set for 2015. The target set within the United Nations Millennium Development Goals (MDGs) is that TB incidence – the number of new cases of TB arising each year – should be falling by 2015. The Stop TB Partnership has set two additional targets, which are to halve rates of TB prevalence and mortality (per 100 000 population) by 2015 compared with 1990. To assess whether these targets are reached, and to build a foundation for better measurement of progress in the post-2015 period, robust monitoring and evaluation of trends in the burden of TB are essential.

In 2006, WHO established a Global Task Force on TB Impact Measurement with the following mandate:

- To produce a robust, rigorous and widely endorsed assessment of whether the 2015 targets for reducing TB incidence, prevalence and mortality are achieved at global, regional and country levels;
- To regularly report on progress towards these targets in the years leading up to 2015; and
- To strengthen national capacity in monitoring and evaluation of TB control.

To fulfill its mandate, the Task Force defined three strategic areas of work for the period 2007–2015:

- **Strengthening routine surveillance.** This includes improving data quality, increasing the proportion of total TB cases and deaths accounted for by surveillance data, and enhancing the collection, management, analysis and use of data, for example through electronic systems. The ultimate goal (beyond 2015) is direct measurement of TB incidence from data on reported (notified) TB cases and direct measurement of TB mortality from national systems for registration of deaths (in which all causes of death are accurately coded using international classification standards) in all countries.

- **National surveys of the prevalence of TB disease.** These are strongly recommended in a set of 22 global focus countries in Asia and Africa that meet epidemiological and other criteria.
- **Periodic review of methods used to translate surveillance and survey data into estimates of TB incidence, prevalence and mortality.**

While the MDG target for TB is that incidence should be falling, estimation of TB incidence is a major challenge. In theory, it can be measured by following a representative sample of the population and counting the number of people that develop TB. In practice, this is not feasible because the required sample sizes are so large, the logistics too complicated and the costs too high. Estimates of TB incidence published by WHO in 2011 relied on systematic analysis of available case notification and programmatic data combined with expert opinion about (i) the percentage of estimated TB cases diagnosed but not reported (i.e. under-reporting) and (ii) the percentage of estimated cases that are not diagnosed at all [1]. Expert opinion was used because, for almost all countries, direct measurements of the levels of under-reporting and under-diagnosis are simply not available.

The lack of direct measurements of under-reporting and under-diagnosis is a problem when estimating the number of TB cases because levels of both may be considerable. This is especially true in countries in which TB is endemic, where people with symptoms of TB seek care from a wide variety of care providers working in the public, private, voluntary or corporate health sectors who may be diagnosed outside national TB control programmes (NTPs). Despite the public health importance of TB, notification (reporting) of TB cases is not mandatory in all countries; and even where notification is mandated by law, enforcement of the law may be weak. A further problem is inappropriate TB management practices among care providers that are not linked to NTPs [3,4]. Besides under-reporting, TB cases may not be diagnosed at all if there are major financial or geographical barriers to accessing health care, or where health-care staff fail to recognize the signs and symptoms of TB or take action when people with TB present at health-care facilities.

Under-reporting and under-diagnosis of TB mean that the actual burden of TB is uncertain, the number of patients who may be receiving substandard care is unknown and that funding may not be well targeted to those most in need. However, once the extent of under-reporting has been evaluated, strategies can be developed and implemented to improve national surveillance systems and, more broadly, to improve TB prevention, diagnosis and treatment services. In countries with a large private sector, one of the most effective ways to increase TB reporting and improve standards of care for patients is to engage all care providers, including public, private, voluntary and corporate providers. These public–private and public–public mix (PPM) approaches aim to align the TB management practices of non-programme care providers with national guidelines and international standards [3,4]. PPM approaches may involve incentive-based schemes, mandatory reporting of cases or reimbursement.

To help to better measure and address the problem of under-reporting of TB, a priority of the WHO's Global Task Force on TB Impact Measurement in 2011 and 2012 was to develop guidance on the design, implementation and analysis of inventory studies to measure TB under-reporting.[1] Inventory studies compare the number of TB cases meeting standard case definitions in all or in a sample of public and private health facilities with the records of TB cases notified to local and national authorities. Comparisons are made through a process called record-linkage, in which duplicate and unique records are identified. Depending on existing systems for data management, records can be linked either using existing databases or linkage may need to be preceded by special efforts (for a limited time period) to collect data on the number of cases diagnosed by all health-care providers in the country, or by all health-care providers in a random sample of well-defined geographical areas. In certain circumstances, the results from inventory studies can be combined with a type of modelling called capture–recapture analysis to estimate TB incidence. In the past 10 years, inventory studies combined with capture-recapture analysis have been implemented in the Netherlands, the UK, French Guiana, Egypt, Yemen, and Iraq [5–10]. Inventory studies can help to plan and implement PPM.

The main purpose of this guide is to describe and explain how to design, implement and analyse an inventory study to measure TB under-reporting. The guide also explains how to apply capture-recapture methods to estimate TB incidence, *emphasizing the conditions that must be fulfilled for these methods to be used.* It is structured in six chapters:

- **Chapter 1: An overview of inventory studies – what, why, when and how?** This chapter explains the main features of an inventory study, why inventory studies are important, and how their goal and objectives can be defined. The essential ingredients for an inventory study to be feasible and successful, and the enabling factors that facilitate implementation, are also discussed. The chapter concludes by highlighting the conditions that must be met for capture–recapture methods to be used.
- **Chapter 2: Study design.** This chapter describes three major study designs for an inventory study. The choice among the three alternatives depends on the objectives of the inventory study (explained in Chapter 1), the type of data and databases that are already available and the level of effort that can be made to compile the necessary data. The first two study designs require *sampling* of health-care providers and *prospective collection of data* on the number of cases diagnosed by these providers during a limited time period, using standard TB case definitions. The choice among the two prospective study designs depends on whether the study objective is (i) to quantify the level of under-reporting; or

[1] This work falls under the strategic area of work "strengthening surveillance".

(ii) to quantify the level of under-reporting and estimate TB incidence using capture–recapture methods. It is worth noting that for both study designs, the prior existence of a case-based national TB surveillance database with personal identifiers greatly enhances study feasibility. The third study design applies when the objective is to quantify under-reporting and estimate TB incidence; when a national TB surveillance database as well as other national case-based databases (e.g. a national health insurance database and a comprehensive hospital database) with unique identifiers (e.g. national identity number) are available; and when standard TB case definitions are used in these databases. This study design requires retrospective data only and no sampling of health-care providers. A fourth study design that remains at the early stages of development and that is mostly likely to be relevant in countries with high-performance surveillance systems is discussed in an Appendix.

- **Chapter 3: Preparing and implementing an inventory study.** This chapter describes how to prepare and implement an inventory study. This includes the development of a study protocol, the composition of the study team and associated job descriptions, field activities including mapping of providers, and data collection and management. The chapter also discusses ethical considerations.
- **Chapter 4: Record-linkage.** This chapter details how to link records from independent databases to identify which of the TB cases identified during the inventory study were reported (notified) to the NTP. The use of both deterministic and probabilistic matching software is discussed.
- **Chapter 5: Data analysis and reporting.** This chapter describes how to prepare and analyse the data once record-linkage is completed, including how to adjust for the effects of sampling design (discussed in Chapter 2).
- **Chapter 6: Capture–recapture modelling.** This chapter describes how to use capture–recapture analysis to estimate TB incidence, emphasizing the underlying assumptions that must be met for these methods to be valid.

Chapters 1–3 should be accessible to most if not all readers. Parts of Chapter 4 and Chapter 5, all of Chapter 6 and the appendices related to study design (in Chapter 2) are intended primarily for the epidemiologists and statisticians who will need to provide guidance and support to the design and analysis of inventory studies.

This guide is intended for use by NTP managers, researchers including epidemiologists and statisticians, and agencies that provide financial and technical support for inventory studies. Its aim is to catalyse many more inventory studies worldwide, leading to better measurement of the burden of TB disease and, in turn, better TB prevention, diagnosis and treatment services.

References

1. *Global tuberculosis report 2012*. Geneva, World Health Organization, 2012.

2. http://www.who.int/gho/en/

3. *Engaging all health care providers in TB care and control – guidance on implementing public-private mix approaches*. Geneva, World Health Organization, 2006 (WHO/HTM/TB/2006.360).

4. Uplekar M, Pathania V, Raviglione M. Private practitioners and public health: weak links in tuberculosis control. *Lancet,* 2001, 15(358):912–916.

5. van Hest NA et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture–recapture analysis? *Epidemiology and Infection,* 2007, 135(6):1021–1029.

6. van Hest R et al. Record-linkage and capture recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England in 1999-2002. *Epidemiology and Infection,* 2008, 136(12):1606–1616.

7. Bassili A et al. Estimating tuberculosis case detection rate in resource-limited countries: a capture–recapture study in Egypt. *Int J Tuberc Lung Dis,* 2010, 14(6):727–732.

8. Guernier V, Guégan JF, Deparis X. An evaluation of the actual incidence of tuberculosis in French Guiana using a capture–recapture model. *Microbes and Infection,* 2006, 8(3):721–727 (e-pub 2006 Jan 13).

9. Bassili A, Al-Hammadi A, Al-Absi A, Glaziou P, Seita A, Abubakar I, Bierrenbach AL, van Hest NA. Estimating tuberculosis burden in resource-limited countries: a capture-recapture study in Yemen. *Int J Tuberc Lung Dis,* 2013 (*in press*).

10. Huseynova S et al. Estimating tuberculosis burden and reporting in resource-limited countries: a capture–recapture study in Iraq. *Int J Tuberc Lung Dis,* 2013 (*in press*).

Chapter 1
Inventory studies: what, why, where and how?

Authors: Emily Bloss, Alex Pavli, Ibrahim Abubakar, Katherine Floyd

This chapter explains what the term "inventory study" means, why inventory studies in the context of TB are important, where they are relevant, and how to define the goal and objectives of an inventory study. It highlights the essential ingredients required for inventory studies to be feasible and successful, and the enabling factors that make them easier to implement. The chapter concludes by summarizing the conditions necessary for inventory studies to be used to estimate TB incidence using capture–recapture methods.

1.1 What is an inventory study?

In the context of TB, an inventory study is a study of the level of under-reporting of TB cases. TB inventory studies compare the number of cases meeting standard case definitions recorded in all or a sample of public and private health facilities with the records of cases notified to local and national authorities. Comparison is done through a process called record-linkage (Chapter 4). Depending on existing systems for data management, records can be linked either using existing databases or may need to be preceded by special efforts (for a limited time period) to collect data on the number of cases diagnosed by all health-care providers in the country or by all health-care providers in a random sample of well-defined geographical areas. By reviewing TB records from multiple sources, the total number of diagnosed cases can be identified (understanding that some cases will appear in more than one list) and the extent to which they have been reported to national surveillance systems can be assessed. Examples of data sources are lists in which people suspected of having TB, people who have been diagnosed with TB and people who have died from TB are recorded. Lists include registers of case notifications, laboratory records and hospital admissions, HIV records and medical prescriptions from pharmacies in the public and private sectors. Where existing data sources are not available, study registers will need to be introduced in a sample of health-care facilities [1].

Certain study designs (see Chapter 2) combine the results from an inventory study with capture-recapture modelling to estimate TB incidence. Capture-recapture methods were originally developed to study the size of wildlife populations, but have subsequently been applied in many epidemiological, demographic and surveillance studies [2–14]. In the context of TB, capture-recapture methods involve cross-matching records from at least three incomplete data sources covering the same population to identify the number of cases common to paired lists, and then using overlapping information and statistical methods to estimate the number of TB cases not identified in any of the lists [7–14].

In the past 10 years, inventory studies combined with capture-recapture analysis have been implemented in countries including the Netherlands [7], the UK [8], French Guiana [9], Egypt [10], Iraq [11], South Africa [12, 13], Yemen [14] and Pakistan.

1.2 Why are inventory studies important and where are they relevant?

Surveillance of TB disease and deaths caused by TB is essential for effective TB prevention and control. Timely, accurate, and complete recording and reporting of TB cases facilitates patient care, public health responses including detection of outbreaks, assessment of risk factors for TB, and evaluation of trends in the number and distribution of TB cases. Reliable TB surveillance systems are also needed to develop national strategies and plans, to track and report on progress in control efforts including progress towards national or global targets, to develop and seek funding for interventions, and to guide policy decisions. For monitoring of trends in the burden of disease caused by TB, the ultimate goal is that surveillance data should provide a direct measure of TB cases and deaths, from notification and vital registration systems respectively [1].

Although the basis for effective TB surveillance already exists – with data collection, management and reporting using standard WHO forms as core part of most NTP activities [15] – the number of reported cases may be considerably less than the actual number of cases. Lack of knowledge about the need or process for reporting cases, limitations in surveillance systems, poorly defined diagnostic criteria, missed diagnoses and lack of access to health care such that people with TB do not seek care [16] are important reasons why the number of reported cases may be less than the true number of TB cases. The consequence of under-reporting is uncertain estimates of important indicators such as TB incidence in many parts of the world [17]. For most countries, estimates of TB incidence published by WHO in 2011 relied on expert opinion about (i) the percentage of estimated TB cases diagnosed but not reported (i.e. under-reporting) and (ii) the percentage of estimated cases not diagnosed at all [17].

Inventory studies can help to identify (and certify) countries in which TB incidence can already be measured directly from surveillance data. In other countries, they can be used to improve estimates of TB incidence by quantifying levels of under-reporting. Measuring levels of under-reporting can also help to identify ways in which surveillance systems need to be strengthened to progress towards the ultimate goal of measuring the number of cases directly from notification data (see Box 1.1).

More broadly, inventory studies provide a solid basis for engaging all health-care providers through PPM approaches, by providing a good understanding of the number and share of diagnosed cases that are being managed by health-care providers not linked to the NTP. Although many countries have scaled up PPM programmes and have begun reporting the contribution of care providers outside NTPs to total notifications, a large proportion of non-NTP health-care providers in high TB burden countries still remain outside PPM programmes and the TB cases that they manage do not get reported. If an inventory study shows that a large proportion of diagnosed cases go unreported, urgent action is needed to ensure that cases are adequately reported and that TB care indicators, including treatment outcomes, are monitored. Study results can also provide a baseline for later evaluation of PPM and other efforts to strengthen TB surveillance. Inventory studies are thus not only a logical first step in implementing PPM but can also help to facilitate the progressive engagement of health-care providers and eventual inclusion of most if not all relevant health-care providers.

Box 1.1: Reasons to conduct (or not conduct) an inventory study

You might need to conduct an inventory study if:

- The TB surveillance system is missing a considerable proportion of the TB cases that occur each year, and estimates of TB incidence that rely on expert opinion to estimate the proportion of cases being missed are very uncertain (for instance, this can be the case in countries with a large private sector or with many general hospitals that do not routinely report TB); and/or
- A large proportion of cases are thought to be diagnosed and treated by health-care providers that are not collaborating with the national TB programme (NTP), and a better estimate of the number of these cases and of the main care providers outside the NTP are needed as a basis for strengthening collaboration with non-NTP providers through public-private mix (PPM) approaches; and/or
- The country has an excellent surveillance system and clear evidence is needed that it captures all (or virtually all) cases, such that reported (notified) cases can be used as a direct proxy for TB incidence.

You might not need to conduct an inventory study if:

- The TB surveillance system is missing a considerable proportion of the TB cases that occur each year, but uncertain estimates of TB incidence based on expert opinion are considered satisfactory; and/or
- There is a large non-NTP sector where TB cases are being diagnosed and treated, but a precise estimate of the number of these cases is not considered necessary or can wait until PPM approaches are implemented countrywide and all care providers treating TB patients are linked to the NTP; and/or
- The country has an excellent surveillance system, but it is not considered necessary to demonstrate that no cases or only a negligible number of cases are going unreported through a scientific study.

Other benefits that may arise from inventory studies include enhancements to data quality by improving the completeness of registration and using the results of cross-validation across multiple data sources, and sustained improvements to the timeliness and quality of reporting following study interventions such as training of health-care providers included in the study [18]. Compared with other study designs, inventory studies may also have the advantage that they can make use of data already being collected, and are cost-effective compared with other population-based sampling methods.

In countries with high-performance surveillance systems, inventory studies can provide clear evidence that cases are not being missed or that only a small fraction of cases are being missed, such that reported cases can be used a good proxy for TB incidence (see Box 1.2). In other countries, inventory studies will provide a much improved estimate of TB disease burden (see Box 1.3), and provide evidence of where and how surveillance needs to be strengthened and interventions such as PPM implemented. **Inventory studies are especially useful in countries with a high burden of TB where robust TB surveillance systems are not yet in place and/or there is high utilization of private providers [19].**

Box 1.2: Inventory and capture-recapture studies in the UK

Surveillance of TB in the UK relies on inventory methods based on matching with other data sources, such as the laboratory database, national HIV surveillance and bespoke surveys, to assess whether all cases of TB are reported to the national programme. These comparisons generally indicate between 5% to 17% under-notification. A detailed audit revealed the proportion is closer to 5%, as some cases reported as "failing to match" were either present in the Enhanced Tuberculosis Surveillance system with few identifiers or related to organisms that should not have been reported. The matching systems have been continually improved and now rely on a probabilistic programme (probabilistic matching is discussed in detail in Chapter 4).

Assessments of the completeness of reporting using capture-recapture methods have used hospitalization records (health-care administration data), laboratory records and death registrations to estimate the records that are not notified to any surveillance system. Previous studies using this approach suggested under-notification to the surveillance system of about 15.9% during the period 1999–2002. However, the methods used and thus the results produced are limited by the inherent dependencies between the data sources and the quality of the matching. Nonetheless, lessons learnt from record-linkage and capture-recapture studies are applied to the national web-based case reporting system to improve data quality.

For further details: Van Hest NA et al. Record-linkage and capture-recapture analysis to estimate the trend of incidence and completeness of reporting of tuberculosis in England 1999–2002. *Epidemiology and Infection*, 2008, 136:1606–1616.

Box 1.3: An inventory and capture–recapture study in Iraq

An inventory study was implemented in Iraq in 2011 with the objectives of estimating the level of TB under-reporting and estimating TB incidence using capture–recapture methods. Prospective longitudinal surveillance was established among all eligible public and private non-NTP providers in a random sample of 8/18 Iraqi governorates for 3 months (May–July). Record-linkage and three-source capture-recapture analysis of data were then conducted.

A total of 1985 TB cases were identified. The NTP registered 1677 patients (observed completeness 84%, i.e. 16% under-reporting). The number of incident cases was estimated at 14 500. Cases diagnosed by all providers represented 81% (95% CI, 69–89%) of estimated incident cases. The estimated ratio of notified to incident cases was 69% (95% CI, 58–76%).

The findings show that TB surveillance needs to be strengthened to reduce levels of under-reporting.

For further details: Huseynova S et al. Estimating tuberculosis burden and reporting in resource-limited countries: a capture–recapture study in Iraq. *International Journal of TB and Lung Disease*, 2013 (in press).

1.3 Goal and objectives

The main goal of an inventory study is to measure the extent to which diagnosed TB cases are reported, as a basis for certification or strengthening of TB surveillance, improved estimates of TB incidence and better diagnosis and treatment of TB patients.

The objectives of an inventory study can be defined as one or more of the following:

1. To quantify the level of under-reporting of diagnosed cases of TB to national surveillance systems.
2. To estimate TB incidence using capture–recapture methods.
3. To demonstrate that under-reporting is minimal.

The study design required to achieve these objectives will vary according to the data and databases already available in a country (details are provided in Chapter 2). In general, Objective 3 is potentially the least demanding in terms of sample size, logistics and cost, and applies to countries with high-performance surveillance systems. Objectives 1 and 2 are more demanding, and Objective 2 is especially demanding if national databases of case-based records are not already available.

Once at least one inventory study has been implemented, a fourth objective could be to assess whether levels of under-reporting have fallen (improved) since the last study.

1.4 Essential ingredients required for an inventory study to be feasible and successful

Chapters 2–6 provide guidance on how to design and implement an inventory study, and how to analyse and report the results. Before getting started, however, certain essential ingredients must be in place for a study to be feasible and successful (enabling factors that make it much easier to implement an inventory study are discussed in Section 1.5).

1.4.1 Case-based data with reliable personal identifiers

Case-based records (i.e. one record per case as opposed to aggregated data for groups or cohorts of patients) are essential because inventory studies require analysis of whether individual patients exist in different lists of cases (e.g. the list of cases reported to the NTP, and a list of all cases diagnosed in the private sector). To calculate the total number of cases diagnosed, cases that appear in more than one list (duplicates) need to be identified. Besides having one record in a database per case, personal identifiers are required to allow reliable identification of cases that appear in more than one list. Examples include unique personal identifiers, such as a TB registration number or a national ID number. Combinations of non-unique identifiers (e.g., names, dates of birth) may also be used, if they are deemed accurate and reliable and either allow individuals to be uniquely identified or enable probabilistic matching.

When reporting results, it is important to describe the overall accuracy of record-linkage and to discuss and acknowledge the implications for the estimate of under-reporting [19].

1.4.2 Standard case definitions across all care providers

A clear and consistent case definition must be used across all data sources to standardize the process of data collection, to accurately match cases across different data sources and to evaluate the proportion of cases reported from different sources [19,20]. Unambiguous and uniform criteria are needed to minimize misclassification of records. Standard case definitions for TB increase specificity and improve the comparability of TB cases reported from different data sources.

If there is variation in case definitions, results will be unreliable. For example, in the private sector where case definitions may be less clear and specific, false-positive cases may be more common (i.e. over-diagnosis).

C ase definitions must be agreed upon when planning an inventory study, before data collection starts, and should not be changed during prospective data collection, analysis and reporting. In studies using retrospective data (see Chapter 2 for the circumstances in which this study design can be used), case definitions must already be consistent across the databases to be used.

Table 1.1: TB case definitions and patient categories

Category		Definition
New TB episodes (first and recurrent* episodes)	Definite case	Culture positive for Mycobacterium tuberculosis, OR WHO-approved rapid diagnostic test (e.g. GeneXpert®) positive, OR at least one sputum specimen positive for acid fast bacilli (from a quality assured microscopy centre)
	Case	Positive histological examination (extrapulmonary TB) OR TB case without bacteriological or WHO-approved rapid diagnostic test results, put on a full course of TB treatment by a qualified physician or health worker

* Recording may occur in the same year as the year of registration. ***Recurrent cases*** have been treated for tuberculosis in the past and been declared successfully treated (cured/treatment completed) at the end of their treatment regimen. Recurrent cases include relapses due to the same *M. tuberculosis* strain as for the previous episode as well as new episodes of TB due to reinfection.

The TB case definitions recommended by WHO are available in recent treatment guidelines [20]. Case definitions are based on the level of certainty in the diagnosis: for example, whether laboratory confirmation (e.g. smear, culture, rapid molecular test endorsed by WHO) is available or not. Cases of TB can be classified according to anatomical site of disease (e.g. pulmonary or extrapulmonary), bacteriological results (including drug resistance), history of previous treatment (e.g. new or retreatment) and HIV status. The case definition can also include clinical manifestations (i.e. symptoms) and epidemiological information (e.g. person, place, and time). Case definitions for TB may vary across countries depending upon guidelines [20] and specific study definitions may be devised as well; what is critical is that the same definitions are used among all care providers.

The TB case definitions that are recommended by WHO are summarized in Table 1.1. They include adaptations to the latest WHO guidelines in the context of the December 2010 endorsement of a new rapid molecular test for TB (Xpert MTB/RIF).

1.4.3 Adequate staffing and funding

Although inventory and capture-recapture studies are usually less costly than other population-based sampling methods relevant to TB [19, 21] including TB prevalence surveys that cost between US$ 1 million and US$ 4 million each [22], adequate financial and human resources are needed for optimum results. The staffing and funding required will depend upon the scope of the study. Usually, more data sources and bigger sample sizes will increase accuracy, but will also increase the cost of data collection. The budgets of recent studies in Iraq and Yemen that covered approximately half of these countries and lasted 6 months were in the range US$ 120 000–300 000 (see Appendix 3.3 for a detailed budget).

When budgeting for an inventory study, it is important to take into account the costs of a pilot study, which can be used to explore migration in and out of the study area, to optimize data collection procedures, determine feasibility including the willingness of care providers outside the NTP to contribute, costs, and constraints during the initial preparatory stage of the study [19, 21]. Donors may be more likely to commit the necessary funding to a study when budgets are clearly defined and feasibility is demonstrated.

1.4.4 Care providers outside the existing NTP network can be mapped and convinced to participate

Non-NTP health-care providers need to be mapped and then convinced to participate in the study. In Egypt [10], Yemen and Iraq [11], it was possible to map public and private care providers in large geographical areas, and to ensure their participation in a nationally representative inventory study through repeat visits from study investigators. The short duration of an inventory study and the very limited workload imposed on health-care providers are factors that should help to encourage their participation (see also Chapter 3).

1.4.5 Expertise in sampling design, data management and data analysis

Sampling design, data management and data analysis are key components of inventory studies, with a lot of information collected from multiple data sources. A plan that clearly documents data management procedures and analytical methods is needed to ensure uniformity and continuity of data collection, smooth data entry, and data validation. A high-quality relational database (i.e. not Excel) is needed to store information. In addition to an experienced data manager, an epidemiologist and a statistician should be available for study design and data analysis. Guidance on data management is presented in Chapter 3 and guidance on design and analysis is included in Chapter 2, Chapter 4, Chapter 5 and Chapter 6.

1.4.6 At least 3 fairly independent data sources and sampling of 50% of country areas if capture-recapture analysis is planned

For inventory studies that will include use of capture–recapture methods, three fairly independent data sources should be used [8]. These data sources need to have a certain level of overlap, of around 15–30% [19, 23]. In addition, the study will need to cover 50% of a country's geographic area, for reasons explained in Chapter 2.

Three data sources are needed to enable validation of data among sources [8]. While analytical strategies to assess and take into account dependence between sources and heterogeneity are available (e.g. log linear modelling, stratification) [4, 6], they can be analytically complex, requiring the involvement of an epidemiologist and/or a statistician.

1.5 Enabling factors that greatly facilitate implementation of an inventory study

There are two major factors that will substantially facilitate the implementation of an inventory study:

- The prior availability of a national case-based electronic database that includes records for all cases reported to the NTP;
- The inclusion of personal identifiers in each record of the national case-based database that can be used for record-linkage. Unique personal identifiers, such as a national health insurance number or an identity card number, are highly preferable.

If these conditions are not met, considerable efforts will be needed to create a case-based electronic database of cases notified to the NTP. If unique personal identifiers are not in use, record-linkage that includes use of probabilistic matching software will be required (as described in detail in Chapter 4).

As more countries adopt electronic recording and reporting systems, the number of countries where these conditions are met should increase. In 2012, WHO published a comprehensive guide on the design and implementation of electronic recording and reporting for TB care and control [24].

1.6 Limitations of capture-recapture methods

Capture-recapture methods can be used to estimate TB incidence. However, four conditions must be met for the methods to be applicable [2, 7, 23, 25, 26, 27]. These are:

- There should be no change to the study population during the study. In other words, the study population is closed and there is no immigration or emigration from the geographical area studied during the time period considered. This is challenging in settings where levels of migration are high.
- Cases can be matched across data sources (i.e. there is no misclassification of records). Potential biases can result if there are differences in case definitions across sources, diagnostic errors or imperfect record-linkages (e.g. spelling or typing errors, lack of birth date or address, changes in or lack of identifiers). It is also important that all cases identified in each source are true cases (i.e. no false-positive cases).
- The probability of being included in a data source (i.e. one of the lists of diagnosed cases) should be the same for all individuals in the population. In other words, there are no subgroups with very different probabilities of being observed in one data source and re-observed in another data source.
- Data sources are independent (i.e. being in the second data source is not affected by being in the first source).

These assumptions are discussed in more detail in Chapter 6.

It is also important to realize that capture–recapture studies will not be able to identify the number of cases in the population in settings where all data sources have a zero chance of capturing some cases. That is, there is an underlying assumption that no TB cases are being missed due to failure to access health-care or failure to be recognized as a TB case when seeking care at a health-care facility. If the underlying assumption is not met, capture-recapture methods may lead to an under-estimate of TB incidence.

References

1. *TB impact measurement policy and recommendations for how to assess the epidemiological burden of TB and the impact of TB control.* Geneva, World Health Organization, 2009 (Stop TB policy paper no. 2; WHO/HTM/TB/2009.416. www.who.int/tb/advisory_bodies/impact_measurement_taskforce/en).

2. International Working Group for Disease Monitoring and Forecasting. Capture–recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology,* 1995, 142(10):1047–1058.

3. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology,* 1995, 6(1):42–48.

4. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: Applications in human diseases. *American Journal of Epidemiology,* 1995, 142(10):1059–1068.

5. Brenner H. Application of capture-recapture methods for disease monitoring: potential effects of imperfect record-linkage. *Methods of Information in Medicine,* 1994, 33(5):502–506.

6. Knowles RL et al; British Paediatric Surveillance Unit. Using multiple sources to improve and measure case ascertainment in surveillance studies: 20 years of the British Paediatric Surveillance Unit. *Oxford Journal of Public Health,* 2006, 28(2):157–165.

7. van Hest NA et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? *Epidemiology and Infection,* 2007, 135(6):1021–1029.

8. van Hest R et al. Record-linkage and capture–recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England in 1999–2002. *Epidemiology and Infection,* 2008, 136(12):1606–1616.

9. Guernier V, Guégan JF, Deparis X. An evaluation of the actual incidence of tuberculosis in French Guiana using a capture-recapture model. *Microbes and Infection,* 2006, 8(3):721–727.

10. Bassili A et al. Estimating tuberculosis case detection rate in resource-limited countries: a capture–recapture study in Egypt. *Int J Tuberc Lung Dis*, 2010, 14(6):727–732.

11. Huseynova S et al. Estimating tuberculosis burden and reporting in resource-limited countries: a capture-recapture study in Iraq. *Int J Tuberc Lung Dis,* 2013 (*in press*).

12. Dunbar R, Lawrence K, Verver S, Enarson DA, Lombard C, Hargrove J, Caldwell J, Beyers N, Barnes JM. Accuracy and completeness of recording of confirmed tuberculosis in two South African communities. *Int J Tuberc Lung Dis*, 2011, 15(3):337-343.

13. Dunbar R, van Hest R, Lawrence K, Verver S, Enarson DA, Lombard C, Beyers N, Barnes JM. Capture-recapture to estimate completeness of tuberculosis surveillance in two communities in South Africa. *Int J*

Tuberc Lung Dis, 2011, 15(8):1038-1043.

14. Bassili A, Al-Hammadi A, Al-Absi A, Glaziou P, Seita A, Abubakar I, Bierrenbach AL, van Hest NA. Estimating tuberculosis burden in resource-limited countries: a capture-recapture study in Yemen. *Int J Tuberc Lung Dis,* 2013 (*in press*).

15. TB e-recording and reporting portal [web site]. Geneva, World Health Organization, 2009 (http://www.who.int/tb/err/en/index.html; accessed 1 October 2009).

16. Nanan DJ, White F. Capture-recapture: reconnaissance of a demographic technique in epidemiology. *Chronic Dis Can,* 1997, 18(4):144–148.

17. *Global tuberculosis control 2011*. Geneva, World Health Organization 2011 (WHO/HTM/TB/2011.16).

18. Silk BJ, Berkelman RL. A review of strategies for enhancing the completeness of notifiable disease reporting. *Journal of Public Health Management and Practice,* 2005, 11(3):191–200.

19. van Hest R, Grant A, Abubakar I. Quality assessment of capture-recapture studies in resource-limited countries. *Tropical Medicine and International Health*, 2011, May 23 (doi: 10.1111/j.1365-3156.2011.02790.x. [Epub ahead of print]).

20. *Treatment of tuberculosis guidelines*, 4th ed. Geneva, World Health Organization, 2009 (WHO/HTM/TB/2009.420)21. Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group. *Morbidity and Mortality Weekly Report,* 2001, 50(No. RR-13):1–29.

21. Chang YF et al. The importance of source selection and pilot study in the capture-recapture application. *Journal of Clinical Epidemiology,* 1999, 52: 927–928.

22. *Tuberculosis prevalence surveys: a handbook*. Geneva, World Health Organization, 2010 (WHO/HTM/TB/2010.17).

23. Kruse N et al. Participatory mapping of sex trade and enumeration of sex workers using capture– recapture methodology in Diego-Suarez, Madagascar. *Sexually Transmitted Diseases,* 2003, 30:664–670.

24. *Electronic recording and reporting for tuberculosis control*. Geneva, World Health Organization, 2012 (http://www.who.int/tb/publications/electronic_recording_reporting/en/index.html; accessed April 2012).

25. Borgdorff MW, Glynn JR, Vynnycky E. Using capture-recapture methods to study recent transmission of tuberculosis. *International Journal of Epidemiology,* 2004, 33(4):905–906 [author reply 907].

26. Brenner H. Effects of misdiagnoses on disease monitoring with capture–recapture methods. *Journal of Clinical Epidemiology,* 1996, 49(11):1303–1307.

27. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. *American Journal of Epidemiology*, 1992, 135:1060–1067.

Chapter 2
Study design

Authors: Charalambos Sismanidis, Philippe Glaziou, Fulvia Mecatti, Alex Pavli, Ross Harris, Amal Bassili, Hazim Timimi, Katherine Floyd

There are four main study designs for an inventory study that are presented in this chapter. Selection of study design depends on which of the main objectives an inventory study is attempting to address, as well as the available data.

As explained in Chapter 1, the objectives of an inventory study are:

1. To quantify the level of under-reporting of diagnosed TB cases to national surveillance systems;
2. To estimate TB incidence using capture–recapture methods; or
3. To demonstrate that under-reporting of diagnosed cases of TB is minimal.

All three objectives may apply at the national or subnational level.

Depending on the defined objectives and the available data, there are three main study designs:

- **Study design 1: Simple random sampling of well-defined geographical areas followed by prospective collection of data for cases diagnosed by all health-care providers within these areas for a specified time period, followed by record-linkage with an electronic, case-based NTP database.** This study design applies to Objective 1 and could also be used for Objective 3. Appropriate documentation and standard TB case definitions must be used by all of the sampled health-care providers during the study period.
- **Study design 2: Simple random sampling of large self-contained geographical areas followed by prospective collection of data for cases diagnosed by all health-care providers within these areas for a specified time period and ensuring that two databases are produced from two separate sources (e.g. health-care providers and laboratories), followed by record-linkage across these two databases and an electronic, case-based NTP database.** This study design is suitable for Objective 1 and Objective 2. *It is much more demanding in terms of study size, logistics and cost compared with Study design 1 since the areas must be large to fulfil the condition of being "self-contained".* Self-contained geographical areas must be sampled to fulfil the conditions required for the application of capture-recapture methods, notably the underlying assumption that the study population is closed and there is no immigration or emigration from the geographical area studied during the time period considered. The reason for needing three separate databases is also that this is a requirement for capture-recapture analysis.
- **Study design 3: Retrospective analysis of all records of diagnosed TB cases held in electronic, case-based databases for a specified time period.** This study design is possible when there is an electronic, case-based database of TB cases that have been reported to the national TB surveillance system and where other electronic and case-based databases containing records for TB cases also exist (for example, a national health insurance database and comprehensive hospital databases). Standard TB case definitions must also be applied across databases, and individual records must include either a) unique identifiers that can be used for record-linkage (for example, a national identity number) or b) combinations of identifiers e.g. name, age, sex, residence. *Unlike the other three study*

designs, no sampling of health-care providers is required. If there is at least one database besides the database containing records of cases reported to the national TB surveillance system, then this study design is suitable for Objective 1 and could also be used for Objective 3. If there are at least two databases in addition to the database containing records of cases reported to the national TB surveillance system, then the study can include Objective 2 as well.

A fourth possible study design that could be used for Objective 3 is described in Appendix 2.2. It involves prospective collection of data for cases diagnosed by a simple random sample of health-care providers selected using *lot-quality assurance sampling* for a specified time period, followed by record-linkage with an electronic, case-based NTP database. Such a study design is expected to be less demanding in terms of study size, logistics and cost compared with Study design 1 because it should be the least demanding in terms of the number of health-care providers that need to be studied. However, it will provide the least precise estimate (among the three prospective designs) of the absolute level of under-reporting. Appropriate documentation and standard TB case definitions must be used by all of the sampled health-care providers during the study period. This study design is included in an appendix because it requires more theoretical development and is expected to apply in a relatively small number of countries compared with the other three study designs.

Definitions of terms used in the chapter are provided in Box 2.1.

Box 2.1: Definition of terms used in the chapter

Under-reporting: A proportion (that can also be expressed as a percentage) calculated as the number of cases diagnosed and not reported to the national TB surveillance system (often managed by an NTP) divided by the total number of diagnosed cases (the sum of reported and unreported cases).

Health-care providers: Any health-care facility, public or private, including dispensaries, private practitioners, paediatricians, small private clinics, hospitals and laboratories, where TB patients may be diagnosed.

NTP and non-NTP providers: A distinction is made between (i) providers operating directly under the NTP or similar authority responsible for national TB surveillance and (ii) other providers that may be in the private or public sector, e.g. public hospitals that are not linked formally to the NTP (see Chapter 3). These two groups are referred to as NTP and non-NTP providers respectively throughout the chapter.

All study designs require an electronic, case-based database of records for TB cases reported to the national TB surveillance system. If such a database is not available, it must be created (see also Chapter 1, Section 1.5). The prior availability of an electronic, case-based database of records for TB cases reported to the national TB surveillance system greatly facilitates the implementation of an inventory study.

Experience with the design of national inventory studies is limited and largely restricted to a small number of high-income countries where study design 3 could be used and a few countries in the Eastern Mediterranean Region where study design 2 was applied. There is no experience of using study design 1. As more inventory studies are implemented, the study design and sampling methods described in this chapter will evolve and improve accordingly. A recent analogy is the evolution of recommendations for the design, implementation, analysis and reporting of TB prevalence surveys, which were updated in 2010 from the first WHO guidelines on the topic (published in 2007) following lessons learnt from surveys completed in Asia and during preparations for surveys in Africa [1].

This chapter explains each of the study designs based on current thinking about the best available methods and drawing on experiences in the UK, the Netherlands and countries in the Eastern Mediterranean Region [2–5]. Biases resulting from unmet assumptions required for capture–recapture modelling are discussed specifically in Chapter 6.

We hope that most if not all of the chapter can be clearly understood by general readers without a background in epidemiology or mathematical statistics. Where more difficult mathematical concepts are included to inform statisticians and quantitative epidemiologists, this is clearly indicated in the text and the material is included in boxes or appendices as much as possible.

2.1 Study design 1: measuring the level of TB under-reporting by simple random sampling of well-defined geographical areas and prospective data collection among all health-care providers in those areas

Study design 1 allows quantification of the level of under-reporting with greater precision than study design 4. It falls short of study design 2 because it does not allow estimation of TB incidence using capture–recapture methods.

The basic approach in study design 1 is as follows:

- Estimation of the sample size required, i.e. the total number of all cases of TB diagnosed by all providers that will need to be included in the study to ensure a pre-defined level of precision in the measurement of under-reporting;

- Definition of well-defined geographical areas followed by simple random sampling of these geographical areas to achieve the required sample size;
- Mapping of all health-care providers in the selected geographical areas followed by prospective collection of data about TB cases diagnosed by all health-care providers in these areas during a specified period of time (the recommended period is 3 months);
- Record-linkage of cases, using the case-based databases created during the study for all providers and a case-based, electronic database of records of TB cases reported to the NTP or national TB surveillance system (see Box 2.3 and Chapter 4).

The following two subsections (Section 2.1.1 and Section 2.1.2) explain how to calculate sample sizes and select geographical areas. Study implementation including mapping of providers is covered in Chapter 3, and record-linkage in Chapter 4.

The definition of the geographical area to be used depends on the country context but in particular its administrative structure and its size in terms of the number of TB cases diagnosed. The geographical area could be a prefecture in China, a sub-district in India or a district in Ghana. Another good candidate for a “geographical area” could be the TB Basic Management Unit (BMU)[2] since its catchment area is well-defined in terms of responsibility for management, supervision and monitoring.

2.1.1 Estimating the sample size required

To estimate the sample size required, the following three components are needed:

- A choice about the **relative precision** (denoted as t) required for the estimate of the true level of under-reporting (denoted as π) that will come from the survey estimate of under-reporting (denoted as p). Relative precision refers to the width of the confidence interval for the true level of under-reporting that is approximately centered on the survey estimate p. It is expressed as a proportion (or percentage) of the true level of under-reporting π. For example, if the relative precision is 0.2 (i.e. 20%), then the 95% confidence interval for π is between p-0.2π and p+0.2π.

[2] Basic Management Unit: A BMU is defined in terms of management, supervision, and monitoring responsibility. A BMU for TB control may have several treatment facilities, one or more laboratories, and one or more hospitals. The defining aspect is the presence of a manager or coordinator who oversees TB control activities for the unit and who maintains a master register of all TB patients being treated, which is used to monitor the programme and report on indicators to higher levels. Typically, the units correspond to the government’s second subnational administrative division, which might be called, for example, a “district” or “county”. It is internationally recommended that a BMU cover a population of between 50,000 and 150,000 or up to 300,000 for large cities. (*Source Compendium of indicators for monitoring and evaluating national tuberculosis programmes* (WHO/HTM/TB/2004.344), page 10. http://whqlibdoc.who.int/hq/2004/WHO_HTM_TB_2004.344_chap1-2.pdf)

- A "prior guess" π_g of **the true level of under-reporting** π. This is based on previous such studies conducted in the country (or other countries that could be considered similar), and/or on data collected during the pilot phase of the current study, and/or on expert opinion;
- **An estimate of the design effect** – the factor by which the sample size required for a simple random sample survey must be multiplied because of the random sampling of geographical areas including groups of, rather than individual, TB cases. The selection of geographical areas means that there will be a clustering effect, i.e. the level of under-reporting will tend to be more similar within geographical areas than between geographical areas. As a result, observations within geographical areas are not independent and each individual in the sample provides less information than an individual TB patient selected at random from the whole population.

There are then two steps in the sample size calculation:

1. Calculation of the sample size that would be needed if the inventory study was a simple random sample (SRS) survey of individuals, i.e. a survey in which each individual TB patient could be sampled at random from a list of all individuals that are the subject of the survey (in an inventory study, the list of all individuals required would be a list of all diagnosed cases of TB);
2. Upward adjustment of the sample size according to the estimated design effect, to account for the fact that in an inventory study geographical areas (i.e. groups rather than individual TB cases) are sampled at random.

Step 1: Let us denote by N_{SRS} the number of individual TB cases to be included in the study and π_g the prior guess of the true level of under-reporting (expressed as a proportion). From statistical theory the standard error of an estimated proportion is given by $se(p) = \sqrt{\frac{p(1-p)}{N}}$. The 95% confidence interval for the level of under-reporting is then calculated as: $p \pm 1.96se(p)$ where $1.96se(p)$ expresses the precision. It can be deduced that the higher the required precision (that is the narrower the confidence interval), the greater the sample size will be.

We then need to express the precision as a proportion of the unknown level of under-reporting (as opposed to an absolute precision). We can denote this relative precision as $t\pi$, where t is a proportion and π is the true level of under-reporting. For example, if the required relative precision is 0.2 (or 20%) of the level of under-reporting then precision $=0.2\pi$.

Therefore, replacing π with the prior guess π_g, we require that:

$$1.96\sqrt{\frac{\pi_g(1-\pi_g)}{N_{SRS}}} = t\pi_g$$

By re-arranging the above equation, it can be deduced that the sample size N_{SRS} required for a simple random sample study can be calculated as:

$$N_{SRS} = 1.96^2 \frac{(1-\pi_g)}{t^2 \pi_g} \qquad (1)$$

Example 2.1.1: A hypothetical country example

Country X is planning an inventory study to measure the level of under-reporting according to study design 1. The prior guess about the true value of under-reporting is 25% and the relative precision required is 20%. The sample size calculation for a simple random sample study of individuals can then be calculated as:

$$N_{SRS} = 1.96^2 \frac{(1-0.25)}{0.2^2 0.25} \approx 289$$

Using the information presented in Example 2.1.1, if alternatively the prior guess of the level of under-reporting was 10% then the sample size required would be greater, at 3 458 with a relative precision of 10% and 865 with a relative precision of 20%. Table 2.1 presents more examples of sample size calculations under different assumptions to illustrate the association between sample size and a) relative precision and b) the level of under-reporting. The sample size increases: a) the greater the required precision of the estimate to be produced by the study and b) as the level of under-reporting decreases.

Table 2.1: Examples of the required sample size for different assumptions about the level of under-reporting π_g and relative precision t (both expressed as percentages)

π_g	$t = 5$	$t = 10$	$t = 20$
5	29 196	7 299	1,825
10	13 830	3 458	865
15	8 708	2 177	545
20	6 147	1 537	385
25	4 610	1 153	289
30	3 586	897	225
35	2 854	714	179
40	2 305	577	145

Step 2: The sample size calculated using formula (1) must then be inflated by a multiplying factor called the design effect (DEFF) to account for the clustering of TB cases in the selected c geographical areas. The DEFF can be expressed in one of two ways. Here, the one that is more intuitive as well as easier to use and interpret is presented [6].

DEFF can be expressed using the true level of under-reporting π, the coefficient of between-cluster variance k and cluster size m:

$$\mathrm{DEFF}=\left[1+(m-1)\frac{k^2\pi}{(1-\pi)}\right] \qquad (2)$$

The coefficient of between-cluster variance k is a measure of the spread of the cluster-specific estimates of under-reporting, π_j, $j=1, \dots , c$ which are sampled from a distribution with mean $E(\pi_j)=\pi$ and variance $Var(\pi_j)=\sigma_B^2$ and is defined as: $k=\sigma_B/\pi$ [7]. The final required sample size, accounting for clustering, is calculated as follows:

$$N=\mathrm{DEFF} * N_{SRS}$$

Example 2.1.2: Estimating the coefficient of between-cluster variance when no prior data are available

We need to account for the clustering effect of the study first discussed in Example 2.1.1, and then inflate the sample size calculated in step 1 accordingly. In the absence of data from previous relevant studies the best that can be done is to discuss with local experts who could provide information about what the plausible range of values of under-reporting would be in the country. Suppose these experts think that the level of under-reporting would be about 25% on average and would vary (we assume this to translate into a 95% confidence interval) between 5% and 45% across different geographical areas in the country where the study would be implemented. Under the Normal distributional assumption of the cluster-specific estimates of under-reporting these assumptions can be translated as $25\% \pm 2 * \sigma_B = (5\%\text{-}45\%) \rightarrow \sigma_B = 10\%$ or $k = {}^{\sigma_B}/_{\pi_g} = {}^{10}/_{25} = 0.4$.

In the context of inventory studies, the cluster size m (that is the number of all cases diagnosed by all providers in each sampled geographical area) will not be equal across the sampled geographical areas. In this context, a modification to typical sample size calculations that can be made is to use in equation (2) the **harmonic mean** of cluster sizes of all c sampled geographical areas as opposed to a fixed cluster size [8].

The harmonic mean can be calculated as:

$$m = \overline{m}_H = \frac{c}{\Sigma_{j=1}^{c}({}^{1}/m_j)}$$

where m_j and $j=1, \dots , c$ denotes the specific number of cases diagnosed by all providers in each sampled geographical area.

From equation (2) it can be deduced that the design effect depends on:

- The number of geographical areas that are selected. Reducing the cluster size (this means the total number of all TB cases diagnosed among all providers in each selected geographical area) and hence increasing the number of geographical areas selected will lower the design effect. For the same sample size in terms of statistical efficiency it is, therefore, better to have more smaller areas than less larger ones;

- The number of all cases diagnosed by all providers in each of the geographical areas that are selected. This number is not known in advance, but the number of *reported* cases from the previous year is known. Since the study design requires including all diagnosed cases in each selected area during a specified time period, and since the number of cases diagnosed will vary among areas, the cluster size cannot be controlled as part of the study design. The larger the harmonic mean of cluster sizes is, the larger the DEFF;
- The extent to which the number of cases diagnosed in each of the selected geographical areas varies. The larger the variability, the larger the DEFF;
- The extent to which the level of under-reporting of diagnosed cases of TB varies among the selected geographical areas. The larger the variability of the cluster-level estimates, the larger the DEFF.

Given the limited number of inventory studies that have been implemented to date, especially in TB-endemic countries, the size of the design effect is difficult to predict. The more such studies are implemented in the future the more relevant and reliable data will become available to inform such sample size calculations. **In the meantime it is strongly recommended to include a pilot phase and an associated interim analysis in the study design.** This will allow the quantification of the outcome of interest and of the clustering (both the coefficient of between-cluster variation *k* to be put in equation (2) and the DEFF itself) based on observed data from the country where the study is implemented. **Most importantly, a recalculation of the initial sample size can be undertaken based on observed values for *k*, DEFF and *p*** and a decision made to continue the study as is, add more sampled geographical areas (if sample size was under-estimated) or stop (if sample size was over-estimated).

Example 2.1.3: Estimating the coefficient of between-cluster variance from empirical data

Suppose we have designed the study described in Example 2.1.1 in such a way that an interim analysis is undertaken after the first 5 clusters have been recruited and followed up for 3 months. These observed cluster-level data allowed us to estimate the overall proportion of under-reporting π_g=0.30, the harmonic mean of cluster size $\overline{m}_H$=43, as well as the empirical variance of the cluster-specific percentages of under-reporting s^2=0.0274. This random error comprises two components, the within- and between-cluster variability.

$$s^2 = \frac{\pi_g(1\text{-}\pi_g)}{\overline{m}_H} + \hat{\sigma}^2_{\mathrm{B}} \Rightarrow \hat{\sigma}^2_{\mathrm{B}} = 0.0274 - \frac{0.30*(1\text{-}0.30)}{43}$$

Hence, we estimate $\hat{\sigma}_B = 0.15$ and $k = \hat{\sigma}_B / p = 0.5$. Then:

$$\text{DEFF} = \left[1+(\bar{m}_H - 1)\frac{k^2 * \pi_g}{1 - \pi_g} \right] = \left[1+(43 - 1)\frac{0.5^2 * 0.30}{1 - 0.30} \right] = 5.5$$

and the final sample size, accounting for clustering, is:

$$N = \text{DEFF} * N_{SRS} = 5.5 * 289 \approx 1590.$$

Stratification may help to decrease the design effect, but the benefit of stratification, if any, will only be known at the end of the study.

2.1.2 Stratification

Levels of under-reporting are likely to vary from area to area, depending on the number of non-NTP providers. For example, some areas within a country may have a small private sector and are therefore unlikely to have a problem with under-reporting. It is possible to stratify geographical areas in an inventory study so that these differences do not unduly skew the national results. The precision of the national estimate of under-reporting will be negatively affected by the sampling design effect due to within-area correlation [9, 10]. Stratification may be used to improve the precision of estimates for a given sample size [11]. This is known as sampling efficiency. To improve sampling efficiency, geographical areas must be stratified according to a factor that is believed to be a good predictor of under-reporting. The case notification rate of a geographical area is potentially a good predictor of under-reporting and may be used to define strata. For example, stratification may be done in such a way that all areas in the country are grouped into areas with high case notification rates, intermediate rates and low rates. Areas are then sampled separately from each of the three strata. At least two areas should be selected from each strata for stratification to have any benefit on sampling efficiency.

2.1.3 Number of geographical areas to be sampled

The decision on the number of geographical areas (or clusters) to sample to find (at least) the planned sample size *N* of TB cases is a difficult one. In the context of these studies, with unknown cluster sizes m_j, $j=1, \dots, c$ of the sampled *c* geographical areas, there are no statistical grounds on which a single required minimum number *c* can be recommended that applies in all situations and countries. However, several clusters are required for certain aspects of the analysis. The number of geographical areas *c* needs to be chosen individually for each country. It is important to remember that the larger the size of the area, the more cases would be found in them and the smaller *c* would need to be. On the other hand the larger the cluster sizes m_j are, the larger DEFF would become, increasing the total sample size of TB cases needed overall to guarantee the intended precision of the estimated under-reporting.

The total number of all TB cases (reported and not reported to the NTP) in each area is not known in advance (otherwise there would be no need for the inventory study), but the number of reported cases from the previous year is available from the routine reporting system. This reported number of cases could be used to inform the initial decision on the number of geographical areas to be sampled (which can later be refined once data collection in a few areas allows better estimation of the size of DEFF, as described above). Prior information about the number of reported cases should be available for each geographical area in the sampling frame (e.g. the number of reported cases in the previous year, in each district). Thus, one approach to selecting the total number of geographical areas to be sampled is as follows:

i. After the sample size calculation, treat the first few clusters as the pilot phase of the study, e.g. the number of geographical areas $c_1 = N_{SRS}/n$, where N_{SRS} is the sample size calculation under SRS as in equation (1) and n is a prior guess of the median value of number of cases expected to be reported over the study period across all areas in the sampling frame. In situations where the cluster sizes, and hence their median value n, are large with respect to N_{SRS}, c_1 will be small (e.g. less than 3). In this context this step is not recommended as the use of the pilot phase data for the calculation of DEFF would only be worthwhile if data from at least 4 to 6 clusters are accumulated. An estimate of the DEFF as explained in Section 2.1.1, Step 2, should then be used to estimate the number of areas to be studied.
ii. From the interim analysis of the pilot phase data recalculate DEFF and revise the sample size calculation.
iii. Calculate the total required number of geographical areas $c = N/n$, where N is the total recalculated sample size and n is the median value of number of cases expected to be reported over the study period across all areas in the sampling frame.

Geographical areas should be selected or defined in such a way that the required sample size will not be reached by the inclusion of a very limited number of areas, say one or two. For instance, if only one type of geographical area is sufficient to meet the sample size, then the studied sample will not be necessarily representative of the country. There is no single recommended minimum number of areas to be sampled that fits all scenarios. We have described in this section one way this issue could be addressed. It is advisable to also include in the judgement knowledge of factors relevant to the definition and selection of geographical areas. Such factors may include heterogeneity in the country's health and infrastructure situation, particularly if the country is very large, decentralization of laws and regulations related to TB case reporting, and logistic constraints.

Example 2.1.4: Country example: India

Let us assume that the state of Gujarat in India would like to perform a sample size calculation for an inventory study. A prior guess about the true value of under-reporting is 25% and the desired relative precision 20% (as in Example 2.1.1), which translates into N_{SRS}=289 by applying equation (1). There were a total of 18,382 TB patients registered for treatment during the first quarter of 2011. Gujarat is split into 30 districts and there are 138 BMUs. The latter constitutes the sampling unit ("geographical area" or "cluster") for the purposes of the study since districts have a large number of patients reported and it is possible that only one or two of them would need to be sampled to reach the desired sample size.

For the sake of simplicity and the purposes of this example: (i) we only present a sub-set of the BMUs, shown in Table 2.2, and (ii) it is assumed that the number of patients expected to be diagnosed in each BMU during the first quarter of 2011 is the number of patients registered for treatment inflated by the expected 25% of under-reporting (assuming the same across BMUs).

The harmonic mean of all 138 BMUs is calculated to be

$$\overline{m}_H = \frac{r}{\Sigma_{j=1}^{r}({}^{1}/{m_j})} = \frac{138}{(\ \ldots+\frac{1}{160}+\frac{1}{187}+\ldots+\frac{1}{289}+\ldots\ )} \approx 158$$

The BMU-level estimates of under-reporting are assumed to vary roughly between 5% and 45%. This translates into $25\% \pm 2 * \sigma_B = (5\%\text{-}45\%) \rightarrow \sigma_B = 10\%$ or $k = {}^{\sigma_B}/{\pi_g} = {}^{10}/{25} = 0.4$. The design effect is then calculated:

$$\text{DEFF} = \left[\ 1+(\overline{m}_H\ \text{-}1)\ \frac{k^2 * \pi_g}{1 - \pi_g}\right] = \left[\ 1+(158 - 1)\ \frac{0.4^2 * 0.25}{1 - 0.25}\right] \approx 9.4$$

This means that the final sample size, accounting for clustering, should be:

$$N = \text{DEFF} * N_{SRS} = 9.4 * 289 \quad 2717.$$

The median value of the number of patients expected to be diagnosed across all 138 BMUs is calculated as 177.5. A pilot phase of 289/177.5≈2 BMUs would not be very informative for recalculating the DEFF, so in this case an alternative is to consider the first five sampled BMUs as the pilot phase. BMUs are assigned a unique identifier as in the leftmost column of abridged Table 2.2 and randomly sampled, one by one, without replacement, until the required sample size is reached. In this example, we use the R language (an open-source freely-available computing environment well suited for data manipulations, graphing and statistical analysis – see http://www.r-project.org) to draw a random permutation of BMU identification numbers (1,…, 138) from abridged Table 2.2:

```
> sample(1:138)
 [1] 4 3 9 8 12 10 7 6 5 2 1 11...
```

The first sampled BMU is 4, the second 3, etc. At the end of the pilot phase, after selecting BMUs 4, 3, 9, 8 and 12, the number of diagnosed cases is 767. Suppose that the design effect is recalculated and DEFF=6. This means that the final sample size is N=6*289=1 734. The study would, therefore, continue after the pilot phase (and the inclusion of districts 4, 3, 9, 8 and 12) to include BMUs 10, 7, 6, 5 and 2 to reach the final required sample size (adjusted for clustering).

Table 2.2: Cases diagnosed by NTP providers during the 1st quarter of 2011[1] in the state of Gujarat in India. (Abridged table only showing 12 out of 138 BMUs in the state).

ID	BMUs	No of patients registered for treatment (N_1)	No of patients expected to be diagnosed $N_2=N_1/0.75$
…	…	…	...
1	Bardoli	120	160
2	Kamrej	140	187
3	Mandvee	103	138
4	Mangrol	71	95
5	Surat	125	167
6	Dhrangadhra	126	168
7	Limbdi	201	268
8	Muli	162	216
9	Ahwa	59	79
10	Baroda	167	223
11	Padra	154	205
12	Savli	179	239
…	…	…	...

[1] RNTCP performance report for 1st quarter 2011, accessed 27/09/2012, http://tbcindia.nic.in/perfor.html

2.2 Study design 2: measuring the level of TB under-reporting and estimating TB incidence using capture–recapture methods with prospective data collection among all health-care providers in selected areas

Study design 2 is the most challenging of the study designs discussed in this chapter because the objective is to estimate TB-underreporting and TB incidence. Capture-recapture methods and prospective data collection are required for this study design.

Application of capture–recapture methods requires that the following elements are part of the study design:

- The geographical areas that are studied must be **self-contained**. In other words, in any of the selected geographical areas, diagnosed cases of TB do not move in and out of that area. To fulfil such a requirement (when the conditions required for study design 3 are not in place) it is likely that sampling of large geographical areas is required. For example, in a recent inventory and capture–recapture study in Iraq [5], half of the country (8 of 18 Iraqi governorates) was included (see Box 2.2).
- At least three types of health-care providers must be distinguished, one of which must be the NTP (this is in contrast to the two categories of provider – NTP and non-NTP – that are distinguished in study designs 1 and 4). Examples of types of health-care providers include the NTP, hospitals, private clinics, laboratories and health insurance companies.
- A non-biased estimate of the number of cases common to the different categories of health-care providers must be obtained. For example, the number of cases common to the NTP and hospitals, or common to hospitals and health insurance or laboratories, relative to the total number of diagnosed cases. This latter requirement imposes more constraints on the sampling approach, compared with study designs 1 and 4.

To fulfil these requirements, the basic approach is as follows:

- Definition of self-contained (and likely large) geographical areas;
- Simple random sampling of the self-contained geographical areas. The default recommendation is to sample half of the identified geographical areas [12-14];
- Definition of the three (or more) categories of health-care provider to be used in the study;
- Mapping of all health-care providers in the selected geographical areas followed by prospective collection of data about TB cases diagnosed by all health-care providers in these areas during a specified period of time (the recommended period is 3 months);
- Record-linkage of cases, using the case-based databases created during the study for non-NTP providers and a case-based, electronic database of records of TB cases reported to the NTP or national TB surveillance system.

The selected geographical areas may be stratified prior to sampling based on the TB case notification rate in the area, but stratification is optional. The sample size, based on estimation of the likely level of under-reporting with a given choice of precision, can be estimated using the methods described in Section 2.1. In several high TB-burden countries (e.g. India, China, Indonesia), the sample size will be very large if 50% of the country is covered by the study.

2.2.1 Self-contained and large geographical areas

The sampled geographical areas must be self-contained when capture-recapture methods are used. This is to ensure that almost all patients diagnosed by both NTP and non-NTP providers have a chance of being identified within the sampled areas. In practice, in the context of a TB inventory study, this means that the sampled areas need to be large. If small geographical areas are sampled, it is much more likely that patients have sought medical advice from providers outside the sampled area and their records may fail to be linked. This idea is illustrated in Figure 2.1.

In Figure 2.1, the boxes on the left and right both contain the same number and arrangement of dots, each of which represents a TB record and its geographical location. The coloured dots represent NTP records and the white dots represent non-NTP records, while the blue lines link up records belonging to the same individual. In the box on the left, two large geographical areas, represented by black ovals, have been selected whereas in the box on the right, five smaller geographical areas have been selected. In the box on the left, only one linkage is missed between a non-NTP record in a sampled area and its corresponding NTP record located outside the sampled area. This represents a patient who has been to a private practice in the sampled area, and then visited a public practice where their TB diagnosis has been notified to authorities. In contrast, in the box on the left, seven linkages are missed. Two of the linkages are cases identified by the NTP outside the sampled areas but by non-NTP providers within the sampled areas, and five are cases that were identified by non-NTP providers outside the sampled areas. As can be seen, a sampling design using small geographical areas will tend to miss more linkages than a sampling design based on larger areas.

The sampling of geographical areas will result in smaller biases if relatively few records in sampled areas have linkages with records outside the sampled areas. It is thus important to consider the physical distance between providers who attend to the same individuals in relation to the size of sampled geographical areas. To minimize bias both in the estimation of under-reporting and in the estimation of incidence using capture–recapture modelling, it is important to minimize the total number of sampled geographical areas. Table 2.3 below illustrates the relationship between the size of the bias in the estimate of under-reporting and the number of sampled geographical areas, for a given sampling fraction of 50%. The bias in the estimate of incidence will be affected similarly (not illustrated).

Figure 2.1: The probability of a case seeking diagnosis and care from a provider within and outside the sampled geographical areas increases as the size of the sampled geographical areas falls

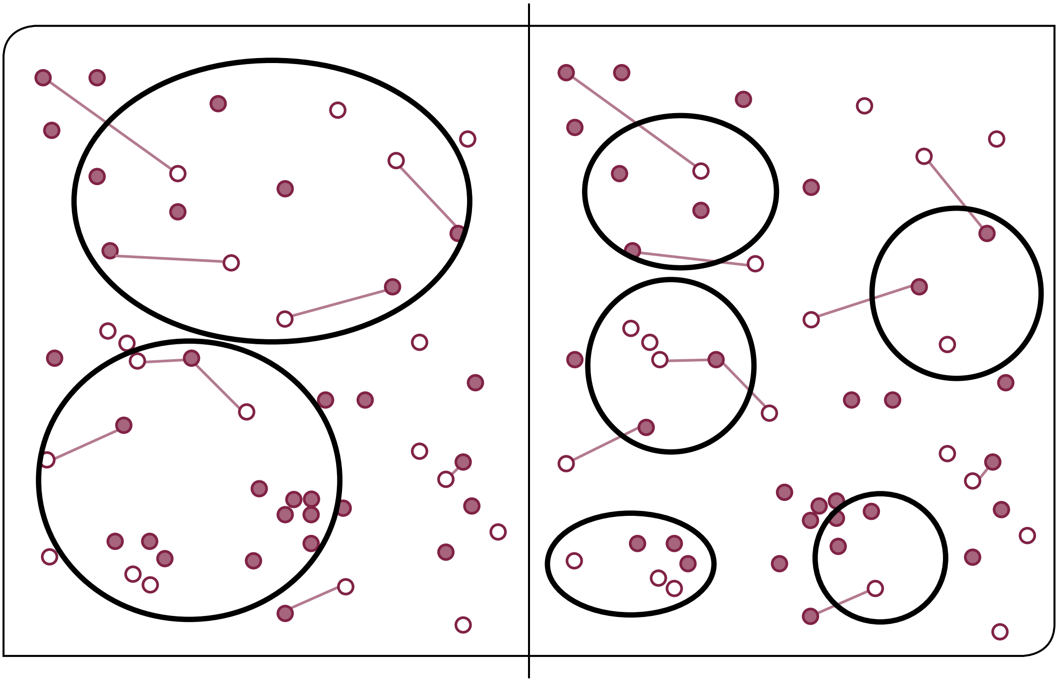


Table 2.3: Bias in the estimation of under-reporting obtained from random sampling of 50% of geographical areas and all providers and their records within areas

Results from sampling a dataset with 528 records with a level of under-reporting of 11.2% indicate a positive bias that increases as the size of geographical areas decreases, resulting in an over-estimate of the level of under-reporting. Centiles of the distribution of biases were obtained from sampling 1000 times. When linked, records tend to be within close proximity of each other, and the sampling of geographical areas is considerably more effective than simple random sampling of providers and their records. The relative bias (rightmost column) will almost always be large compared with the relative precision of the estimate.

Number of geographical areas	Number of sampled areas	Sampling fraction (%)	Estimated under-reporting (%) (true value: 11.2%)	Sampling bias (2.5th-97.5th centiles)	Bias relative to the true value (%)
64	32	50	13.8	+2.5 (2.2 – 2.8)	23%
8	4	50	12.9	+1.6 (1.2 – 1.9)	15%

2.2.2 Definition of geographical areas

Since the selected geographical areas must be 'self-contained' in terms of the provision of TB diagnosis and care, areas need to be defined so that patients have little or no propensity to seek TB diagnosis and care outside of the area. An example of a self-contained geographical area is an entire province including a capital city. Geographical areas should be clearly defined prior to sampling. As geographical areas must be large, appropriate areas include rural areas covering 1–10% of the national population. Contiguous urban areas and their surrounding districts or geographically-connected cities should not be spread over different areas.

Box 2.2: The example of Iraq

An inventory study was undertaken in Iraq in 2011, to measure the level of TB under-reporting in the country, and perform a capture-recapture analysis (see Chapter 6) to estimate TB incidence. The country is split into 18 governorates which are further divided into 124 districts. Each district comprises one BMU. A stratified sampling design was used, based on the smear-positive TB case notification rate at the level of the governorate. Four strata were defined (based on the 25th, 50th (that is, median) and 75th centile values of the smear-positive notification rates) and two governorates selected with simple random sampling from each stratum. Sampled governorates covered 53% of the country's total population.

All facilities were enrolled in the study during the period May–July 2011 in each of the sampled governorates and all TB cases diagnosed during this period in these facilities were included in the study. The TB case definition in facilities was the same as the one used by the NTP.

Laboratory and TB register forms identical to those used by the NTP were introduced to all participating facilities for the collection of data. All facilities were visited weekly to ensure data collection, completeness and accuracy.

Record-linkage of TB cases from facilities diagnosed during the period May–July 2011 and NTP records from two quarters before and one quarter after the study period was done.

The results were used to estimate that TB incidence in Iraq was approximately 14,500 new cases in 2011 and that the level of under-reporting was 16%.

2.3 Study design 3: retrospective analysis of existing databases

A retrospective study design is feasible in countries where all of the following conditions are met:

- A case-based national TB database with electronic records for all cases reported to the national TB surveillance system is already in place;
- Other national, electronic case-based databases with records for TB cases (e.g. a national health insurance or a comprehensive hospital database) exist;
- Both sets of databases have unique identifiers, such as a national identity number, that can be used for record-linkage;
- Standard TB case definitions are used in all databases.

Records of TB cases across databases are linked for a specified period of time in the past. Since databases are national, no sampling of health-care providers is required. Recent examples of countries where inventory studies using this design have been implemented and used to estimate TB under-reporting and TB incidence include the Netherlands (in 1997) [2] and England and Wales (during 1999–2002) [3].

Common characteristics of the three prospective study designs (1, 2 and 4) and an important note about record linkage are shown in Box 2.3.

Box 2.3: Common characteristics of the three prospective study designs in which sampling is used

Common characteristics of study designs 1, 2 and 4

1. Direct random selection of diagnosed cases cannot be performed because there is no source of information about all diagnosed cases (otherwise the study to measure underreporting would not be needed). Selection of geographical clusters of health-care providers is suggested instead.
2. The recommended time period for data collection is around 3 months (for theoretical details see Appendix 2.2).
3. Diagnosed cases may have records maintained by different providers. During the study period, all records from all of the providers selected for inclusion in the study are compiled.
4. Standard case definitions must be used by all health-care providers in the selected geographical areas.
5. An electronic NTP case-based database is required for record-linkage, to identify which of the diagnosed TB cases identified in the inventory study were reported.
6. Record-linkage needs to include records of cases reported before and after the study period, to ensure that cases are correctly classified as reported or unreported; the default recommendation is 3 months before and 3 months after the study period (for theoretical details see Appendix 2.2). The reason for this is to allow for: (i) the time taken to report a TB case to the national TB surveillance system and (ii) the pattern of patient health-seeking behaviour. The NTP records from the 3 months on either side of the study period are used to classify cases that were diagnosed within the sample as either reported or not reported.

Common characteristic of study designs 1 and 2

1. Geographical areas are sampled using simple random sampling (SRS) and then **all** health-care providers within the selected areas are eligible for inclusion in the study.

   The geographical areas that are sampled at random should be areas that can be readily defined, for example because they are based on existing administrative boundaries. Examples include districts and counties.

Important note for record linkage

Record-linkage needs to include investigation of whether cases diagnosed within the selected geographical areas were subsequently reported *outside* those areas. The reason for needing to look for matches with NTP records outside the geographical areas selected for the study is that some of the TB cases diagnosed by non-NTP providers in the selected geographical areas may seek medical care at an NTP provider located outside the area after their diagnosis. This is in contrast to Study design 2, in which the selected areas are "self-contained". In Study designs 1 and 4, it is possible that cases diagnosed by a non-NTP provider are subsequently reported in another geographic area – the one where they sought care with the NTP.[1] Unless the NTP records for the relevant location outside the study area are part of the record-linkage process, a case could be wrongly classified as unreported.

To minimize the chance of this happening, NTP records from outside the selected geographical areas need to be used when record-linkage is done. For example, if a selected geographical area is a city centre, it is important to include NTP records for the outlying suburbs in the record-linkage process as well. Similarly, records from large cities surrounding selected geographical areas that are mostly rural areas will need to be included if it is judged or known that the population living in the selected area is likely to seek specialized medical care in such cities.

Patterns of health-care seeking behaviour and the extent to which health-care providers outside the study area are used can be explored as part of a pilot study, to better understand the extent to which reporting outside the study areas is likely to occur (and where it is most likely to occur). Collection of data on where the cases diagnosed in the selected geographical area subsequently sought care and their normal place of residence can also help to reduce the amount of effort required to link study records with NTP records outside the study areas.

It should be stressed that records of cases reported outside the selected geographical areas are only used to classify cases that were diagnosed within the selected geographical area as reported or not reported.

[1] A good example is a patient with TB who attends a non-NTP private practice in a city centre where their TB diagnosis is made but not reported to the NTP authorities. After a few weeks, the patient goes for a check-up with a public provider near their residence in an outlying suburb and here their case is reported to the NTP. If the geographical area sampled in the inventory study was limited to the city, then the case would be counted as 'not reported' in the inventory study.

Appendix 2.1 Time to notification

It is important to note that bias with regards to time to notification is dependent on both the time to notification and study duration. Both a shorter time to notification and longer study duration result in less bias in the estimation of under-reporting, as can be seen in Table A.2.1.1. In the table, the time to notification is modelled as an exponential distribution and it is assumed that non-NTP case detection occurs at a constant rate. From the table it can be seen that for any given study duration, a shorter time to notification will lead to fewer cases being notified after the completion of the study. Similarly, for a given time to notification, longer study durations will result in fewer cases reported after the study completion date. For example, if the average time to notification is four weeks and the study duration is also four weeks, then 63% of cases will be reported after the study period. However, if the average time to notification remains at four weeks, and the study duration is increased to twelve weeks, then only 32% of notifications will occur after the study period.

Table A.2.1.1: Percentage of cases diagnosed by non-NTP providers in whom reporting to NTP occurs after the study period, out of all reported non-NTP cases

	Study duration*	4 weeks	8 weeks	12 weeks	24 weeks
Average time to report					
1 week		25	13	8.3	8.3
2 weeks		43	25	17	17
4 weeks		63	43	32	32
8 weeks		79	63	52	52

* The study duration in the table does not include the recommended additional 3 months of inclusion of NTP records before and after the study period in order to reduce the proportion of cases wrongly classified as not reported.

Cases are assumed to be diagnosed at a constant rate over time and uniformly distributed over the study period. Time to notification is then modelled as exponentially distributed. The expectation of the proportion Q of non-NTP cases with notification occurring after the study period is obtained as follows:

$$E(Q(t;\lambda)) = \int_{0}^{s} te^{-\lambda t}\, dt$$

where s is the duration of study, t is the time from non-NTP case record occurrence to the end of study and λ is the reciprocal of the average time to report to NTP, or the rate of reporting.

Appendix 2.2 Study design 4: prospective data collection and lot quality assurance sampling to assess if tb under-reporting is minimal

Study design 4, based on the methods of lot quality assurance sampling (LQAS), is appropriate when the objective of the inventory study is to demonstrate that under-reporting is minimal. This objective is most relevant in countries in which the national surveillance system functions well and the aim is to demonstrate that this is the case.

The basic approach in study design 4 involves four steps:

- Sampling of health-care providers from a comprehensive list of all health-care providers in the country (for a national assessment) or within a particular part of the country (for a subnational assessment). Since a list of all health-care providers is required, this must be already available (or easy to compile) if a national assessment is to be done. For a localized assessment, mapping might be feasible prior to sampling from the list of all providers;
- Prospective collection of data about TB cases diagnosed by the selected health-care providers for a specified period of time (e.g. 3-month period);
- Classification of each of the sampled health-care providers as either "acceptable" or "unacceptable" according to pre-specified thresholds of *levels of under-reporting* that are "acceptable" and "unacceptable";
- Classification of the level of under-reporting in a country or part of a country as either "acceptable" or "unacceptable". The classification of "acceptable" applies if the *number of unacceptable providers* is below an allowable and pre-specified threshold. The classification of "unacceptable" applies if the number of unacceptable providers is above the allowable and pre-specified threshold.

By design, such studies do not provide a precise measure of the absolute level of under-reporting. However, the advantage is that smaller numbers of health-care providers need to be studied compared with study designs 1 and 2. Furthermore, even if the study finds a country "unacceptable", that is, under-reporting is found to be higher than anticipated, then data from this design could inform sample size calculations for further investigations into the level of under-reporting in the form of study design 1 and 2.

This is the first time that the LQAS approach is considered for use in the context of inventory studies, so to date no practical experience exists. The clustered sampling of cases through health-care providers is a complication that leads to increased misclassification errors with the LQAS approach. The amount of this increase will depend on how correlated the patients

within a health facility are with respect to their under-reporting status. As field experience and empirical data accumulate the quantification of these unknown factors, as well as the testing of the different suggested solutions to design issues presented in this section, will be refined accordingly.

It is intended that this trial edition of the inventory guide will be used to test study design 4 in one or two countries with high-performance surveillance systems, where study design 3 cannot be done, with experience and lessons learned then used to refine the methods set out here.

A2.2.1 Sampling and required number of health-care providers

The LQAS approach was first introduced in the manufacturing industry in the 1920s as a way of checking the quality of the output from the mass production lines of factories [15]. It was subsequently adopted in other fields and began to be used in the health sciences in the late 1980s [16], with a wide variety of applications to date including assessment of immunization coverage (the most common topic), cancer screening, antimalarial treatment and oral rehydration therapy [17–20]. In the context of a TB inventory study, LQAS could be used to test whether the true level of under-reporting in a country (or subnational area) is less than a *pre-specified* acceptable level.

In LQAS theory, the pre-specified level of under-reporting that is considered "acceptable" is expressed as a proportion. For example, if the level of under-reporting that is considered acceptable is 5%, then the acceptable level of under-reporting is expressed as 0.05 (this is π_g in Box A2.2.1).

Box A2.2.1: LQAS in mathematical notation and theory

The true level of under-reporting in a country (or subnational area) can be denoted as π and the acceptable, pre-specified level of under-reporting as π_g.

Although it has been stressed in the published literature that statistical hypothesis testing is not part of LQAS, there are clear analogies between LQAS and hypothesis testing that can help to explain LQAS methods [21, 22], and for this reason the analogy is used here.

In the context of a TB inventory study, the desirable outcome in LQAS is to establish that π is less than π_g. In hypothesis testing terms, this can be expressed as rejecting the null hypothesis (H_0) in favour of the alternative (H_α):

$$H_0 : \pi \geq \pi_g \text{ Vs. } H_\alpha : \pi < \pi_g$$

The “acceptable” threshold (expressed as the proportion π_g) is then translated into a threshold (denoted *d*) of allowable numbers of sampled health-care providers (which are the “lots”) that are unacceptable in terms of their level of under-reporting. This value *d* is referred to as the “decision rule” in the LQAS literature.

The selection of the value of *d* is as much an informed choice as it is a calculation. The value of *d* is based partly on statistical considerations, through the selected values for the threshold π_g and the statistical error that is allowed; partly on logistical and feasibility considerations; and partly on what study investigators consider acceptable or unacceptable in relation to the total sample size of required health-care providers. The statistical error (using the hypothesis-testing analogy, this is the probability α of the type I error) is the probability of incorrectly rejecting the null hypothesis, that is concluding that the number of unacceptable health-care providers is less than or equal to *d* and classifying a country as “acceptable” when the level of under-reporting is actually above π_g. A typical value to choose for α is 5%.

From statistical theory, say that the number of health-care providers is a random variable X that follows a hyper-geometric distribution (*d* “successes” of unacceptable health-care providers out of *n* draws without replacement from a finite population of size *N*, controlled for at π_g). Then, using its cumulative distribution function, the type I error can be expressed as the maximum allowable probability of observing up to *d* unacceptable providers (and concluding under-reporting is less than the acceptable level *p*):

$$Prob(X \leq d) \leq a \Rightarrow \Sigma_{i=0}^{d} \frac{\binom{N\pi_g}{i}\binom{N(1-\pi_g)}{n-i}}{\binom{N}{i}} \leq a \qquad (1)$$

The selection of sample size *n* (the number of health-care providers that need to be sampled) and the decision threshold *d* are guided by calculations performed according to inequality (1) that meet the chosen restrictions for π_g and α, but also feasibility considerations as well as what seems acceptable to study investigators. The stricter the study team decides to be in terms of selecting small values for *d*, the smaller the required sample size *n* will be. The larger *d* becomes, the larger the total sample size *n* also becomes, hence weakening the advantage of small sample sizes that LQAS compared with other study designs in which samples sizes are estimated based on the precision with which the outcome of interest is measured (in this case, the level of TB under-reporting).

For further details see Rhoda DA et al. [21] and Pagano M et al. [22]

The "acceptable" threshold (expressed as a proportion) is then translated into a threshold (denoted d in Box A2.2.1) of the allowable *number* of sampled health-care providers (which are the "lots") that are unacceptable in terms of their level of under-reporting. This value *d* is referred to as the "decision rule" in the LQAS literature. If the number of sampled health-care providers found to be unacceptable is the same or below the *pre-defined* critical value *d* then the country (or subnational area) is considered to be acceptable in terms of under-reporting. **Agreement on the values for π_g and d is the key issue in studies of this design.**

The cluster sampling of TB cases from health-care providers means there is correlation between cases from the same providers in terms of their under-reporting status (the outcome of interest), which increases the misclassification type I error α associated with the LQAS analysis. The amount of this increase is unknown, but can be informed, ideally, from empirical data. In the absence of empirical data, computer simulations need to be employed to study it [23].

The selection of the value of *d* is *as much an informed choice as it is a calculation*. The value of *d* is based partly on statistical considerations, through the selected values for the threshold π_g and the statistical error that is allowed; partly on logistic and feasibility considerations; and partly on what study investigators consider acceptable or unacceptable in relation to the total sample size of required health-care providers (Box A2.2.1). The stricter the study team decides to be in terms of selecting small values for *d*, the smaller the required sample size *n* will be. The larger *d* becomes the larger the total sample size *n* also becomes, hence weakening the advantage of small sample sizes that LQAS has over other study designs. Examples of sample size calculations for different values of *n* and *d* are illustrated in Table A2.2.1.

Box A2.2.2: R code that can be used to explore different sample size scenarios

```
nsize <- function (p = 0.05, N = 1000, d = 1, alpha = 0.05){
            s <– N
            for (n in N:1){
                  m <– N – n
                  k <– trunc (p * N)
                  if (dhyper (d, n, m, k) > alpha) break
                  s <– n
            }
            return (s)
}
```

Table A2.2.1: Examples of sample size calculations of *n* health-care providers to be included in the study, for varying total numbers *N* of health-care providers in the country (or subnational area), varying values of decision interval *d*, a pre-specified level of "acceptable" under-reporting of 5% and different assumptions about the statistical error (5% or 10%)

Total number of health-care providers in the country N		*d = 0*	*d = 1*	*d = 2*	*d = 3*	*d = 4*	*d = 5*
100	α = 0.05	45	65	80	92	99	100
	α = 0.1	37	57	74	88	98	100
150	α = 0.05	52	76	96	114	129	142
	α = 0.1	42	65	86	105	122	137
200	α = 0.05	51	76	97	117	135	152
	α = 0.1	41	64	84	104	123	141
300	α = 0.05	54	80	103	125	145	165
	α = 0.1	42	66	87	108	128	147
500	α = 0.05	56	83	108	131	153	175
	α = 0.1	43	68	90	111	131	151
1000	α = 0.05	57	86	112	136	159	181
	α = 0.1	44	69	92	113	134	153
10 000	α = 0.05	59	89	115	140	164	187
	α = 0.1	45	71	93	115	135	155

Table A2.2.1 shows how the number of health-care providers *n* required for the study increases as (i) the total number of providers in the country *N* (or subnational level) increases (although *n* increases much more slowly compared to *N*) and (ii) the decision interval *d* increases; conversely, the larger the allowable statistical error α, the lower the required sample size.

The sample sizes in Table A2.2.1 were calculated using the R computer code shown below (Box A2.2.2). Different scenarios can be explored according to different country contexts using this code. R is freely available for download at: http//www.r-project.org

To use the code shown in Box A2.2.2, it should be directly copied and pasted into an R console. The code generates a function named **nsize**. The function code is minimalist and does not include checks for improper parameter values. The function can then be called upon for different values of the acceptable threshold level of under-reporting π_g, the decision threshold *d*, the total number of health-care providers in the country *N*, and the probability α of the type I error. In the first example in Box A2.2.3 below, **nsize** returns a sample size of health-care providers n=76 under the default assumptions π_g=0.05, N=200, d=1 and α=0.05. In the second example in Box

A2.2.3, the returned sample size n=57 corresponds to assumptions π_g=0.05, N=1,000, d=0 and α=0.05. From these examples it is clear that the parameter with the most influence on the sample size is d.

Box A2.2.3: Examples of sample size calculations using the code provided in Box A.2.2.2

```
nsize(p=0.05, N=200, d=1, alpha=0.05)
[1] 76
nsize(p=0.05, N=1,000, d=0, alpha=0.05)
[1] 57
```

Health-care providers should be sampled in a representative manner from an exhaustive list of all health-care providers in the country under a simple random sampling design, with an equal probability of selection for each provider and inclusion of all diagnosed TB cases during the study period.

Box A2.2.4: Two scenarios that illustrate the role of using confidence intervals and not just point estimates to assess whether a provider is "acceptable" or "unacceptable"

Scenario 1 – use of the point estimate only
Threshold: under-reporting is less than or equal to 5%.
A health-care provider diagnoses 10 patients during the study period and has a single patient not reported to the NTP. This gives a point estimate of under-reporting of 10% (>5%). This provider is classified as unacceptable.

Scenario 2 – use of the confidence interval
Threshold: the 95% confidence interval for the level of under-reporting includes the value 5%.
A health-care provider diagnoses 10 patients during the study period and has a single patient not reported to the NTP, giving a point estimate that the level of under-reporting is 10% and an accompanying 95% confidence interval of 0.2–45%. This provider is classified as acceptable.

A2.2.2 Classification of health-care providers

As explained in Section A2.2.1, the LQAS approach informs the decision of the required number of health-care providers n that need to be sampled, as well as the decision interval d of the number of allowable unacceptable providers, to determine if the country (or part of the country) has a level of under-reporting that is less than a pre-specified acceptable level (e.g. 5%). Following sampling, health-care providers are recruited into the study and data collection is established for a certain period of time (e.g. 3 months as indicated in Box 2.3). The level of under-reporting during that time is measured after record-linkage (described in detail in Chapter 4) as the percentage of diagnosed TB cases by the provider that also appear in the NTP database over the total number of diagnosed TB cases by the provider.

It is not possible to know in advance how many TB cases each provider will diagnose during the study period. Furthermore, the number of cases diagnosed across providers will be quite variable, especially if follow-up time is restricted to be the same for all providers.[3] A strategy that characterizes health-care providers as acceptable or not, taking into consideration this variation in the number of diagnosed patients, is required - in particular to ensure that study results and conclusions are not unduly influenced by providers in which only small numbers of TB patients are diagnosed (see Box A2.2.4). These are some possible approaches:

1. Classify a health-care provider as acceptable or unacceptable based on the 95% confidence interval around the best estimate of under-reporting, rather than the point estimate itself.
2. Restrict the sampling frame of health-care providers to include only providers that diagnose a minimum number of TB cases. Defining what this minimum number should be needs to be informed by experience in the field and empirical data.
3. Calculate the overall national (or sub-national) level of under-reporting as a weighted average of the level of under-reporting of the sampled health-care providers and compare with π_g. The total number of diagnosed cases could be used as the weight for each health-care provider. This is a hybrid LQAS approach that requires more field experience to be further studied and validated.

The rest of this section elaborates further on approach 1. We classify a health-care provider based on the 95% confidence interval around the best estimate of under-reporting, rather than the point estimate itself. Once monitoring of all providers is completed for the pre-specified study period, they can all be classified as acceptable, if the 95% confidence interval around the estimate of under-reporting includes the chosen threshold (e.g. 5%), or unacceptable, if the 95% confidence interval around the estimate of under-reporting does not include the chosen threshold. If the sum of the total number of unacceptable providers is less than or equal to the decision interval d then the country overall (or the part of the country evaluated with the study) is considered to have a minimal level of under-reporting. If the total number of unacceptable providers is more than d then the level of under-reporting is above the minimal threshold π_g.

[3] We recommend fixing the same study period across all providers to avoid temporal differences that could bias results, but also as a practical solution to the logistical complications with providers that are very slow at recruiting cases.

Table A2.2.2 provides some examples of confidence interval calculations around the percentage of under-reporting of a health-care provider, based on the binomial exact statistical approach. The number of unacceptable health-care providers from the total sampled will decide if national (or sub-national) level of under-reporting overall is acceptable or not.

Table A2.2.2: Example calculations for the percentage of patients under-reported in a health-care provider, and their 95% confidence intervals (based on the binomial exact approach)

Total number of patients diagnosed	*Total number of patients not reported*	*Percentage of health-care provider under-reporting (95% confidence interval)*	*Classification of health-care provider (5% value included in the confidence interval)*
5	2	40 (6–85)	**Unacceptable**
5	1	20 (1–72)	**Acceptable**
10	3	30 (7–65)	**Unacceptable**
10	1	10 (0.2–45)	**Acceptable**
20	5	25 (9–49)	**Unacceptable**
20	3	15 (3–38)	**Acceptable**
30	5	17 (6–35)	**Unacceptable**
35	6	17 (7–34)	**Unacceptable**
40	6	15 (6–30)	**Unacceptable**
45	6	13 (5–27)	**Acceptable**
50	7	14 (6–27)	**Unacceptable**
70	8	11 (5–21)	**Acceptable**
80	10	13 (6–22)	**Unacceptable**
90	10	11 (5–19)	**Acceptable**
100	11	11 (6–19)	**Unacceptable**

A2.2.3 A country example

A country example that illustrates the methods described in Section A2.2.1 and Section A2.2.2 is provided in Box A2.2.5.

Box A2.2.5: A country example

Country X has a low burden of TB disease and is widely thought to have one of the best performing TB surveillance systems in the world, with a strong NTP network spanning all (*N*=1240) health-care providers that diagnose TB in the country. The NTP believed that the level of under-reporting in the country was about π_g=5%, and decided to embark on a nationwide inventory study to produce the scientific evidence needed to demonstrate that this was the case. Different scenarios for the required sample size were investigated, and the country finally chose a sample size of 87 health-care providers, based on a decision interval *d*=1 of the number of allowable health-care providers that would be deemed unacceptable (defined as a confidence interval around the level of under-reporting that did not include the value 5%) and a margin of statistical error (of misclassifying the country as having a level of under-reporting of up to 5% when in reality it was higher) of α=0.05.

Provider	n_1/n_2[1]	Percentage of under-reporting (95% confidence interval)	Classification
1	0/21	0 (0–16)	*Acceptable*
2	1/5	20 (1–72)	*Acceptable*
3	0/0	Not evaluable	*Not evaluable*
…[2]	…	…	…
86	3/10	30 (7–65)	*Unacceptable*
87	0/12	0 (0–26)	*Acceptable*
88	0/4	0 (0–60)	*Acceptable*

[1] n_1=number of TB cases diagnosed by the provider not found in the NTP lists, after record-linkage, n_2=number of total diagnosed TB cases;

[2] All providers not shown in the table have diagnosed at least one TB case during the study period and have all been classified as acceptable

From the total list of all 1240 health-care providers in the country, 87 were randomly selected to take part in the study for a period of 3 months each. The study team visited each of the selected providers and enrolled them in the study, asking them to produce and submit to the study team a list of all diagnosed TB cases during the study period. The abridged table above shows the main results from enrolled health-care providers.

Out of all sampled health-care providers, 1 was deemed unacceptable. Another did not diagnose any TB cases during the study period, hence it was not possible to assess the level of under-reporting, and was replaced by an additional health-care provider (which is why the total number of providers is 87+1). The evidence produced from this study suggests that the level of under-reporting in the country is less than 5%.

References

1. *Tuberculosis prevalence surveys: a handbook*. Geneva, World Health Organization, 2010 (WHO/HTM/TB/2010.17).

2. van Hest NA et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture–recapture analysis? *Epidemiology and Infection*, 2007, 135(6):1021–1029.

3. van Hest R et al. Record-linkage and capture recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England in 1999-2002. *Epidemiology and Infection*, 2008, 136(12):1606–1616.

4. Bassili A et al. Estimating tuberculosis case detection rate in resource-limited countries: a capture-recapture study in Egypt. *Int J Tuberc Lung Dis*, 2010, 14(6):727–732.

5. Huseynova S et al. Estimating tuberculosis burden and reporting in resource-limited countries: a capture–recapture study in Iraq. *Int J Tuberc Lung Dis,* 2013 (*in press*).

6. Thomson A. et al. Measure of between-cluster variability in cluster randomized trials with binary outcome. *Statistics in Medicine,* 2009, 28:1739-1751.

7. Collett D. *Modelling binary data*. Chapman & Hall/CRC. 2003.

8. Hayes RJ, Moulton LH. *Cluster randomized trials*. London, Chapman and Hall, 2009.

9. Tillé Y. *Sampling algorithms* [Springer Series in Statistics]. New York, Springer, 2006.

10. Thomson A, Hayes R, Cousens S. Measures of between-cluster variability in cluster randomized trials with binary outcomes. *Statistics in Medicine*, 2009, 28:1739–1751.

11. Handbook of capture-recapture analysis. Edited by Amstrup CS et al. Oxfordshire, Princeton University Press, 2005.

12. Schmidtmann I. Estimating Completeness in Cancer Registries – Comparing Capture-Recapture Methods in a Simulation Study. *Biometrical Journal*, 2008, 50(6):1077-1092.

13. Kose T et al. Extending the Lincoln-Petersen Estimator for Multiple Identifications in One Source. *Biometrical Journal*, 2012 (preprint).

14. Levy PS, Lemeshow S. *Sampling of populations:methods and applications* [Wiley Series in Probability and Statistics]. London, Wiley, 1999.

15. Dodge H, Romig H. A method of sampling inspection. *Bell System Technical Journal*, 1927.

16. Lanata CF et al. Lot quality assurance sampling in health monitoring [Letter to the editor]. *Lancet*, 1988,

331:122–123.

17. Robertson SE, Valadez JJ. Global review of health care surveys using lot quality assurance sampling (LQAS), 1984–2004. *Social Science and Medicine*, 2006, 63:1648–1660.

18. Pezzoli L et al. Cluster-sample surveys and lot quality assurance sampling to evaluate yellow fever immunisation coverage following a national campaign, Bolivia, 2007. *Tropical Medicine and International Health*, 2009, 14:355–361.

19. Pezzoli L et al. Clustered lot quality assurance sampling to assess immunisation coverage: increasing rapidity and maintaining precision. *Tropical Medicine and International Health*, 2010, 15:540–546.

20. Hedt BL et al. Multidrug resistance among new tuberculosis cases: detecting local variation through lot quality-assurance sampling. *Epidemiology*, 2012, 23(2):293-300.

21. Rhoda DA et al. LQAS: user beware. *International Journal of Epidemiology*, 2010, 39:60–68.

22. Pagano M et al. Commentary: understanding practical lot quality assurance sampling. *International Journal of Epidemiology*, 2010, 39:69–71.

23. Olives C et al. Cluster designs to assess the prevalence of acute malnutrition by lot quality assurance sampling: a validation study by computer simulation. *Journal of the Royal Statistical Society A*, 2009, 172:495-510.

Chapter 3
Preparing and implementing an inventory study

Authors: Amal Bassili, Philippe Glaziou, Alexandra Pavli, Emily Bloss, Hazim Timimi, Mukund Uplekar, Knut Lonnroth, Ibrahim Abubakar, Katherine Floyd

Now that we have presented an overview of inventory studies and the principles of study design, it is time to turn to the practical steps that are needed to prepare and implement an inventory study. The first step is to develop a study protocol. Once the protocol has been finalized and approved (including by an ethics committee), the study team needs to be established, including identification and/or recruitment of staff with the appropriate qualifications and experience. In most cases, technical assistance will be required from a local research institution or external partner as it is not likely that the NTP will have enough experienced staff to implement an inventory study, especially if they have other full time responsibilities. As soon as the study team is in place, preparations for data collection can begin. In prospective study designs (three of the four options discussed in Chapter 2), a pilot study will be needed to test the proposed methods for mapping health-care providers and collecting data. Following the pilot and adjustment of study methods based on lessons learnt, full implementation (mapping providers and collection of data) can begin. During prospective data collection, it is essential that data are properly collected and managed, and that the study is appropriately monitored. The final stage is data analysis and the preparation of a final study report.

This chapter presents each of the major components of an inventory study. Further details related to the analysis of data are covered in Chapter 4 and Chapter 5.

3.1 Study protocol

The study protocol is an essential document that is fundamental to a high-quality inventory study; it should guide the work from start to finish. The protocol should outline the rationale for the study, its goal and objectives, methods for sampling, data collection and analysis, the collaborating partners involved, the study timeframe, the budget required (including for staff), ethical considerations and how findings will be disseminated and used. A good protocol will help to secure funding, and clear instructions and objectives in the protocol will assist team members to work together to achieve common goals. It also ensures consistency in the implementation of the study and helps to uphold the rights of study participants. The study protocol must be reviewed by an ethics committee and approved by public health authorities before the study is implemented [1].

3.2 Inventory study team

An inventory study is a large undertaking. It must be performed accurately and on budget, within a limited timeframe. This section outlines the key staff members of the inventory study team, including descriptions of their roles and responsibilities and the qualifications needed.

3.2.1 Steering committee

A steering committee is an advisory committee typically composed of high-level executives and stakeholders. It is recommended that a steering committee be established, particularly as the results of the inventory study may be used to strengthen TB surveillance, implement PPM approaches and re-estimate the burden of the disease. In an inventory study, the steering committee should include stakeholders such as the NTP, the public health service, local research institutions and possibly the sponsor agency providing funding. The steering committee has the final responsibility for the study design, the study protocol, the quality of the study and the final study report. It is also responsible for preparing the study protocol.

3.2.2 Key staff members

Principal investigator. The principal investigator (PI) is a member of the steering committee and is the person responsible for all study activities.

- Job description: the PI assembles an inventory study team that has all the expertise needed to design, implement and analyse the study.
- Qualifications: at least 5 years of managerial experience in the field of public health; strong managerial skills; an extensive knowledge of population-based studies and of TB control principles.

Data manager. The data manager leads the data management unit and reports to the PI.

- Job description: the data manager coordinates data management and is responsible for data validation, storage and backing up of data.
- Qualifications: proven extensive experience with data management in the context of research studies; appropriate skills for building, managing and maintaining databases; analytical skills; and skills in the maintenance of adequate documentation.

Epidemiologist. The epidemiologist reports to the PI and is involved from the planning stages to the preparation of the final report.

- Job description: the epidemiologist uses their knowledge of health systems and TB to help design the study, establish case definitions and ensure the study design is consistent with the study goal and objectives. They provide guidance during data collection, help to analyse and interpret results and contribute to writing the final report.
- Qualifications: at least an MSc degree in epidemiology; proven experience with population-based epidemiological studies, preferably including studies about TB; extensive knowledge of the health care system in the country.

Statistician. The statistician reports to the PI and is involved from the planning stages to the preparation of the final report.

- Job description: the statistician advises on study design, notably sampling design, works with the data manager to design and implement data collection forms and record-linkage procedures, and plans and implements data analysis. They may also help to conduct data quality checks during prospective data collection and will contribute to preparing the final report.
- Qualifications: at least an MSc in statistics; proven experience with the design and analysis of studies with complex sampling designs.

Field team leaders. These are required for inventory studies in which prospective data collection is required (likely to be most studies in TB endemic countries). Field team leaders supervise the work performed by the field team members within a specific geographical area (or areas). An example is provided in Box 3.1.

- Job description: the job entails visiting the selected geographical area, organizing the mapping of providers, contacting and enrolling providers in the survey, and liaising with local and provincial authorities.
- Qualifications: at least 2 years of experience in field work for survey or research projects; managerial skills; and knowledge of TB control principles.

Box 3.1: The composition of field teams in the 2010 Inventory study in Yemen and methods used to encourage participation

In the 2010 inventory study in Yemen, there was one study coordinator for each district. Additional field officers were recruited in some districts with large number of non-NTP facilities. The role of these staff was to orient the non-NTP providers about the study, enroll them in the study after obtaining informed consent, visit each provider twice per month to ensure proper completion of the forms and to collect completed forms. They also regularly checked the status of notification of diagnosed cases and checked the diagnosis made by non-NTP care providers if the detected cases were not notified to the NTP. The field teams were supervised by the governorate coordinators (one per study governorate). The governorate coordinators followed the implementation of the survey in their governorate and made supervisory visits to all or some selected districts.

In the preparatory phase of the inventory study, letters were sent from the Ministry of Public Health (MOPH) to private and public health-care providers in the study governorates to inform them about the study objectives and to ask them to collaborate with the field teams. The letters requested them to provide the field teams with the requested information to complete their research study, but did not request them to refer or notify their cases to the NTP. Field workers wore badges labeled with the study title, their names and role in the study. They also carried copies of the MOPH letters to show them to the health-care providers during the first visit. This approach helped to ensure a very high participation rate in the study among non-NTP providers. The short duration of the field work (3 months) also helped to achieve a high participation rate.

Incentives were paid to field officers (see Appendix 3.3) per visit, with the amount paid according to the distance to the non-NTP facility from the district centre where the TB Basic Management Unit (BMU) was located. For example, 400 YR (just under US$2) was the incentive per visit to a facility within the centre of the district and YR 2000 (US$ 9) was the incentive for facilities located in the outskirts of the district. The per diem of the governorate coordinator during field visits was YR 3200 (US$ 15); the visit would last 2 days since travel to the study district was required. The monthly cost of field and supervisory visits in Hodeida governorate was YR 488,000 (US$ 2270); further details are provided in Appendix 3.3. Incentives were also given to NTP staff, who were made aware that study activities should be part of their routine daily work in TB control. As a result, the NTP has been able to sustain the engagement of non-NTP providers after the completion of the study with limited resources.

Field team members. The field team members work under the direct supervision of the field team leaders.

- Job description: mapping providers within selected geographical areas; visiting providers regularly during data collection; and reporting to their assigned field team leader.
- Qualifications: experience in field survey or research work.

Study monitor. The study monitor is usually from an external organization and monitors the study impartially to ensure compliance with the study protocol, to ensure that quality is assured and that ethical standards are met. The study monitor works in close collaboration with field team members, the data manager and the PI. The study monitor reports to the Steering Committee and to the sponsor.

- Job description: works in close collaboration with the PI; ensures compliance of operations with the study protocol and completeness and accuracy of the data.
- Qualifications: at least one year of experience in field research work or clinical trials; extensive diplomatic skills to ensure effective interactions with the PI and field team members.

3.2.3 Technical assistance

An assessment of the skills and expertise of study staff should be conducted before initiating the study to ensure capacity for its design, implementation, analysis, and dissemination. If additional assistance is needed, a technical agency or research institute with experience in conducting inventory studies should be approached.

3.3 Field activities: mapping of providers and data collection

This section applies to inventory studies with a prospective design only (see Chapter 2). In countries where a national TB surveillance database as well as other national case-based databases (e.g. a national health insurance database and a comprehensive hospital database) with unique identifiers (such as a national identity number) are available, and standard TB case definitions are used in these databases, field-work to collect data prospectively is unnecessary.

Once the study protocol has been written and the team established, it is possible for field team members to begin mapping and enrolling providers, providing them with instructions and collecting information about the TB cases that they diagnose and treat.

Study duration (typically 3 months) should be calculated in accordance with principles in Chapter 2.

3.3.1 Mapping and enrolment of providers

All care providers within the geographical areas selected for the study who may prescribe TB diagnostic tests are potentially eligible to participate in the study (the selection of geographical areas is explained in Chapter 2). Eligible providers are described in Table 3.1. It is important to note that because TB can affect any organ, private clinicians working in a variety of specialties should be consulted and enrolled in the study. Eligible providers within selected areas are then invited to participate in the study and to document details about new TB cases they detect over the study period. In large hospitals, it may be necessary to ask the heads of medical wards to co-ordinate their staff so that all practitioners who might prescribe TB diagnostic tests participate in the study. The study participants are then defined as the health-care providers and laboratory staff that deliver care to TB suspects and patients in the selected areas and who consented to participate in the study. An example template of a form to help with the mapping of health-care providers is illustrated in Table 3.2.

Table 3.1: Types of TB service providers

Public	Private
• Public hospitals • University hospitals (governmental) • Public health insurance institutions • Ministry of Interior or Justice (prisons) • Ministry of Defence (military) • Laboratories • Health clinics • Others (depending on the country setting)	• Private hospitals • Private teaching hospitals (private universities) • Private clinics • Individual private practitioners • NGOs • Laboratories (including microbiology laboratories performing microbiological TB diagnostic tests, pathology laboratories performing histopathological diagnosis of extrapulmonary TB and radiology units) • Others such as private pharmacists if they prescribe medicines and indigenous practitioners if they diagnose TB and prescribe TB medicines

Table 3.2: Example template for mapping of providers

Geographical area: ____

Name of Field team leader: ____

Facility name	Facility type	Facility specialization	Address	Name of the contact person	Email	Telephone	Workload of TB suspects per week	Preferred time for visit	Consent given (Y/N)	Remarks

3.3.2 Instructions to providers

Providers enrolled in the study should be given instructions stating that the purpose of the study is to document the number of TB cases that they detect in order to improve the assessment of the burden of disease caused by TB (see Appendix 3.4). If reporting is mandatory by law, then the instructions should not make any reference to mandatory reporting of TB; forms and documents should not have visible links to NTP policies and documents. The reason is to limit the tendency of study participants to change their behaviour simply because they know they are being observed as part of a study (known as the “Hawthorne effect”). The Hawthorne effect would result in more TB cases being reported to the NTP by providers who would otherwise have failed to report cases.

Providers should also be informed that the investigations they order to diagnose TB will be reviewed by the study investigators to ensure the use of consistent case definitions (see Chapter 1 for WHO case definitions). Non-NTP and laboratory providers will be instructed to enter information into the non-NTP and laboratory registers respectively, which are further described in Section 3.4 on data collection. The forms collected by the field team members will be submitted to the field team leader of each geographical area for review before submission to the data manager.

3.3.3 Non-notified cases

When cases are recorded by providers in the data collection forms used in the inventory study, the subsequent steps taken will depend on the laboratory where the diagnosis was performed. If investigations were done in a quality-assured laboratory then they can be accepted as the final, definitive results. If the investigations were done in a laboratory with undocumented quality, the provider and patient will be contacted to ask the patient to confirm the diagnosis, for example by sending specimens to another quality-assured laboratory, using an **identity card** specific to the study patients to avoid registration of that case in an NTP register. The field team members are responsible for informing non-NTP providers about the correct regimen and guidelines if they are not in place, and as much as possible, should ensure that cases diagnosed at non-NTP facilities are treated at these facilities according to NTP guidelines.

Incentives may be needed to ensure the compliance of non-NTP providers with study requirements. At the end of the study, the identified non-NTP cases should be officially registered, if they have not already been done so, and their treatment outcome should be reported.

3.4 Data collection and management

Data collection and data management are crucial aspects of implementing an inventory study: the accuracy and reliability of the final results are dependent on these processes. A data management plan documenting all data management procedures should be developed before the study to ensure all data management activities are correctly and uniformly followed. A data management unit headed by an experienced data manager should also be established to take overall responsibility for data management activities and to ensure their quality.

It is necessary to create at least two electronic registers such as an NTP register; and a non-NTP register and/or a laboratory register. Each of these registers is explained in more detail below and example templates are illustrated in Figure 3.1 and Figure 3.2. The data collection includes a patient ID used with a bar code, national ID number, or specific patient ID code that will be used to compile complete records (Figure 3.3), for example to combine laboratory results with register information. Ideally, a personal identification number, such as a national social security number unique to each resident, will be assigned to each TB case. However, in many countries there is no unique identifier that can be used, and proxy identifiers will have to be used, such as name, date of birth and sex (see Chapter 4 and [1]).

3.4.1 NTP register

If it is not already available at the NTP, an electronic case-based database needs to be developed for the TB patients notified to the NTP during the study period (e.g. 3 months) and the 3 months before and after the study period (see Chapter 2). This is necessary to allow for cross-checking and confirming the notification status of non-NTP cases (see also Chapter 1 and Chapter 4).

Figure 3.1: An example of a register for patients with presumptive and confirmed TB to be used by non-NTP providers

Geographical area:________________ Type of non-NTP provider:__________ Officer name:________________

District:________________ Specialty:__________ Signature:________________

Patient ID (bar code)	**Computer ID**	**Date**	**Name of TB Suspect (pulmonary or extrapulmonary)**				**Age (years)**	**Sex (M/F)**	**Complete Address and telephone**	**Requested investiga-tions (from the same facility or through referral)** 1=sputum smear examination 2=X-ray 3=both (1+2) 4=culture 5=other	**Action taken** 1=referred 2= diagnosed and referred for treatment 3=treated	**Place of referral (1)**	**Reason of referral (2)**	**Provisional diagnosis** 1=ss+ PTB 2=ss- PTB 3= Pulmonary, bacterio-logically positive 4= Pulmonary, not specified 5=EP 6=other (non-TB)	**To be completed by investigator**	
			First	Father's	Grandfather's	Fourth name									Final Diagnosis 1=ss+ PTB 2=ss- PTB 3=Pulmonary, bacteriologi-cally positive 4=Pulmonary, not specified 5=EP 6=other (non-TB))	Date of initiation of anti-TB treatment at the NTP (record date)

Definitions and codes

Type of non-NTP provider:
Public:

- Public hospitals
- University hospitals (governmental)
- Ministry of Interior (prisons)
- Ministry of Defence

Private:

- Private hospitals
- Private teaching hospitals (private universities)
- Private clinics
- NGOs

Specialty:

- General Practitioners (GP)
- Chest physicians
- Internal Medicine specialists
- Informal provider
- Other

Sex: M=male; F=female

Place of referral (1):

1. District TB centre (NTP)
2. public lab
3. private lab
4. other provider
5. other, specify

Reason of referral (2):

1. confirming diagnosis
2. treatment
3. NA (in case of treatment)

Figure 3.2: Example of a laboratory register

Geographical area:________________ Type of non-NTP laboratory:________ Officer name: ________________

District:________________ Signature:________________

Patient ID (bar code)	ID number	Lab. serial No.	Date specimen received	Name				Age (years)	Sex (M/F)	Complete address (all patients) and phone, mobile)	Referred by (1)	Results of sputum smear micros-copy (2)			Other laboratory investiga-tions	To be completed by investigator	
				First	Father's	Grandfather's	Fourth name					1	2	3		Final Diagnosis 1=ss+ PTB 2=ss- PTB 3=Pulmonary, bacteriologi-cally positive 4=Pulmonary, not specified 5=EP 6=other (non-TB)	Date of initiation of TB treatment

Definitions and codes

Type of non-NTP provider:
Laboratory:

- private
- public

Referred by (1):
[1] self
[2] community: treatment supporter, etc.
[3] public provider
[4] private provider
[5] traditional healer, Informal practitioner
[6} drug store/pharmacies
[7] other

Sex: M=male; F=female

Results of sputum smear microscopy (2):
(NEG): 0 AFB/100 fields;
(1-9 bac) exact number if 1 to 9 AFB/100 fields;
(pos1): 10-99 AFB/100 fields;
(pos2): 1-10 AFB/ field;
(pos3): > 10 AFB/ field

Figure 3.3: Examples of patient identification cards

Using National ID No.

National ID No. ###################	Patient full name:
Referring Facility:	Date:

Using Bar code

Bar code ###################	Patient full name:
Referring Facility:	Date:

Using a Study Identifier

IDNO □ □ □ □	Patient full name:
Referring Facility:	Date:

In terms of the variables for which data are recorded, this electronic TB register will be identical to the TB registers routinely used by the NTP, which are still paper-based in many countries. Key variables included in the register are patient identifiers, geographical area, type of TB case (for case definitions see Chapter 1) and category of provider (for example, NTP, private or public non-NTP).

3.4.2 Non-NTP register

An electronic case-based register of all TB cases diagnosed by non-NTP providers must also be created. Records for TB cases from non-NTP providers should be classified by source: for example, private providers, hospitals, laboratories or health insurance, and then entered into the database system. Key variables included in the register are patient identifiers, geographical area and type of TB case. An example is shown in Figure 3.1.

3.4.3 Laboratory register

A paper-based laboratory register must be established in each of the non-NTP laboratories and should be completed by a laboratory technician designated by the head of the laboratory. An example template is illustrated in Figure 3.2. Key variables included in the register are patient name, contact address with telephone number, date of birth, sex, source of referral, date of receipt of specimen, number and results of specimens examined, and other investigations, for example histological examinations. The national ID number can also be added to the form if there is universal coverage in the country.

3.4.4 Database software

The choice of software (not Excel) to build the necessary relational database should be guided by the experience of the data manager. Preferably, the software should include robust security. It is essential that the software programme offers mechanisms to ensure that invalid values cannot be entered. Extensive guidance on data management principles and good practices is available in *Tuberculosis prevalence surveys: a handbook.* [1] and in *Electronic recording and reporting for tuberculosis control* [2].

3.5 Monitoring

Monitoring and data quality assurance are carried out through routine data verification and validation, supervisory visits and adequate training. They are essential for the successful implementation of the record-linkage phase of an inventory study described in Chapter 4.

3.5.1 Routine data verification and validation

Registers should be checked for their completeness and the quality of collected data, such as correct filling of information according to the right codes and internal consistency. Facilities with a low workload can be contacted by telephone until they have a sizable number of forms to be collected. Having the cell phone numbers of the providers is crucial for monitoring the registration of cases and to agree on the timing of each visit. Electronic records should be quality assured to ensure their completeness and accuracy [1].

3.5.2 Supervisory visits

Supervisory visits are conducted by the PI and the study monitor together with the data management unit. A supervision plan is developed based on the number of geographical areas to be sampled (discussed in Chapter 2). The frequency of visits could be weekly, fortnightly or once per month according to the number of geographical areas, the number of non-NTP providers, and logistic considerations such as transportation facilities, distances to be covered and cost. A standardised process for supervisory visits is strongly recommended, facilitated by a standard checklist that is used at each visit. A template for the checklist is available in the Annex. The template includes information about the extent of participation of the mapped providers, data quality, verification of the status of diagnosis and registration made at the non-NTP facilities, and ensuring adequate diagnosis and treatment of all cases in both NTP and non-NTP facilities. A feedback report is then prepared and discussed with field team leaders. The supervisory team should then follow-up on the implementation of the corrective measures. An example of a supervision checklist is illustrated in Appentix 3.1.

3.5.3 Adequate training with emphasis on the potential sources of errors

All field team members should adequately understand the details of the study, how to brief the non-NTP providers on filling in the forms, how to review the collected forms, and sources of errors and bias. The pilot stage of the study may be used to identify issues and to strengthen the training phase.

3.6 Ethical considerations

Several ethical issues must be considered when implementing an inventory study. Inventory studies involve the collection of confidential data from patients in order to generate aggregated indicators. Surveillance activities are typically not considered research and do not require ethical approval. In some countries, inventory studies will not be considered research while in others, they will be considered research.

The purpose of ethics approval is to secure the safety of study participants, to ensure confidentiality and privacy and to protect the rights and dignity of people involved in the study. Research with human beings is guided by bioethical principles outlined in internationally recognised documents including the Nuremberg Code [3] and the World Medical Association's Declaration of Helsinki [4]. The Nuremberg Code upholds the principle of voluntary participation and informed consent. The Council for International Organizations of Medical Sciences (CIOMS) has published a document called I*nternational Ethical Guidelines for Epidemiological Studies* [5]. This document is relevant to epidemiological studies and asserts the centrality of beneficence, respect for persons, and highlights the need to treat populations and individuals fairly.

The inventory study team should consult with the appropriate ethical review committee and ensure the study design complies with ethical requirements and gain ethical approval if necessary. WHO has issued general guidance on ethics in the context of TB control and epidemiological research that should be reviewed [6, 7].

In an inventory study, the team performing the study is different from the group of data custodians responsible for the secure collection, use and disclosure of the original data. In most countries, nominal information collected for administrative and regulatory purposes may be used later for further analysis and research so long as: the project is ethically acceptable and scientifically valid; data are only released under formal arrangements; appropriate confidentiality and security safeguards are guaranteed; and the results of the linkage procedures should only to be released to other researchers without identifiers. However, in some countries, data or information acquired by an agency for "purely statistical purposes" can be used only for statistical purposes and cannot be shared with identifiers for any other purpose without the informed consent of the involved persons [8, 9]. In this regard, it is important to note that NTP and other TB registries serve other purposes than the pure estimation of TB burden. Case-management, for example, is one of the key objectives in recording such data.

Inventory studies aim to identify TB patients diagnosed in public or private services that have not been notified to the NTP registry. Apart from estimating TB burden, the result of these studies can immediately contribute to improving national TB surveillance systems by clarifying the particular characteristics of records of TB cases not notified to the NTP and where and why the compulsory notification system failed. Non-notified patients found as a result of these studies may also directly benefit from the study. For instance, confirmatory diagnostic procedures may be performed, inappropriate treatment schemes may be changed in favour of the nationally recommended ones and access to free treatment may be granted. The public good which will result from inventory studies and linkage procedures largely outweighs the minimal risk of potential loss of privacy.

Appendix 3.1 Supervision checklist

Geographical Area	**Date**	**Review period**
1. Coverage		
Public Non-NTP providers in the district	*Number of existing private non-NTP providers in the district*	*Number of private non-NTP providers engaged in the study*
Prisons		
Military services		
University hospitals		
Public hospitals		
Others (add rows and specify)		
Private Non-NTP providers in the district	*Number of existing private non-NTP providers in the district*	*Number of private non-NTP providers engaged in the study*
Private clinics		
Private hospitals		
NGOs		
Private university hospitals		
Faith-based services		
Private polyclinics		
Others (add rows and specify)		
2. Data quality	**Public non-NTP providers**	**Private non-NTP providers**
Hard copies of forms		
Total number of collected cards in the district		
Number with missing or incomplete information		
Type of missing/incomplete variables (list)		
Number with errors		
Type of errors (list)		

Timeliness of collecting the cards (Number of field workers who submit the filled cards before the deadline out of all)		
Cross-checking the forms with database (random sample using ID code)		
Number with errors		
Type of errors (list)		
3. Verification of diagnosis made by non-NTP providers	**Public non-NTP providers**	**Private non-NTP providers**
Total number of suspects identified by non-NTP providers		
Number of suspects for whom diagnosis was verified by the NTP		
Number of cases diagnosed by non-NTP providers for whom diagnosis was verified by the NTP		
4. Verification of the status of registration at NTP: quality, misclassification	**Public non-NTP providers**	**Private non-NTP providers**
Is the status of registration of cases diagnosed by non-NTP providers and with a diagnosis confirmed by the NTP regularly checked in the TB register by the district TB coordinator?		
Are cases with a confirmed diagnosis that were not originally registered in the district TB register being registered in a separate register at the district level (and not in the main district TB register) until the end of the study? What is their number? (*see below note)		
5. Ethical issues	**Public non-NTP providers**	**Private non-NTP providers**
Informed consent was taken from all participants		
NTP informs non-NTP provider about the correct regimen/and guidelines, when needed		
Cases diagnosed by non-NTP providers and with diagnosis confirmed by the NTP treated by non-NTP providers according to NTP guidelines?		

Appendix 3.2 Budget template

Activities	Input description/ TOR	Measurement unit	Quantity	Frequency	Unit cost (US$)	Total cost (US$)
Human resources and field work						
Survey coordinator	Coordination and monitoring the project	Person months	1	6		
1 epidemiologist	Data management	Person months	1	6		
1 statistician	Data management	Person months	1	6		
Field supervisors	Supervisory visits at district level	Person days	12	6		
Field subsidies for district supervisors and non-NTP coordinators	Data collection from non-NTP (every 2 weeks)	Person days	195	7		
Subtotal						
Training and workshops						
Participants	Field and district supervisors, 5 non-NTP coordinators	Per diem	53	4		
Facilitators	Project coordinator, project manager	Per diem	2	4		
Refreshment and logistics		Per workshop	1	4		
Subtotal						
Local Travel						
Air travel	From central unit to far districts	Per travel	3	3		

Local transportation	Cost of fuel	Per visit	207	7		
Subtotal						
International technical assistance						
M&E missions of international experts	International technical assistance	Per person per mission	3	2		
Subtotal						
Office equipment and stationery						
Printing of forms	Data collection forms	Lump sum	1	1		
Stationery		Lump sum				
Subtotal						
Data entry, management, analysis and reporting						
Data entry	Data entry for NTP and non-NTP	Person month	2	7		
Final Data cleaning and management	International technical assistance	Per person	1	1		
Data analysis and reporting	International technical assistance	Person month	2	2		
Subtotal						
Communication facilities						
Telephone cards for cell phones, mailing etc.	Communication facilities	Every 2 weeks	207	7		
Subtotal						

Dissemination of results						
Meeting	One meeting for dissemination of results	Per event	1	1		
Subtotal						
Contingency cost	5% of the total					
Total budget						

Appendix 3.3 Budget for monthly visits and incentives in the 2010 inventory study in Yemen

Total incentives for 2 monthly visits	Incentives for the field visits by the survey teams (Yemeni Rials (YR)) US$1=215 YR					Number of targeted facilities		Hodeida governorate
	Total incentives for one visit for all facilities in the districts	Total incentives for the visits in the district outskirts	Total incentives for the visits in the district centre	Incentives for the visit per facility in the district centre	Incentives for the visit per facility in the district centre	Facilities in district outskirts	Facilities in district centre	Districts
58400	29200	0	29200		400	0	73	Elmina
60000	30000	8000	22000	2000	400	4	55	Elhok
56400	28200	12000	16400	2000	400	6	41	Elhaly
18400	9200	0	9200		400	0	23	Bagel
4800	2400	0	2400		400	0	6	Elgerahi
6400	3200	0	3200		400	0	8	Elkanawos
19600	9800	1600	8400	800	400	2	21	BetElfakih
3600	1800	800	1200	800	400	1	3	Elzeydeya
8000	4000	0	4000		400	0	10	Elmansouria
4800	2400	0	2400		400	0	6	Hess
4000	2000	0	2000		400	0	5	Elhegela
12800	6400	2400	4000	800	400	3	10	Elmerawgha
44000	22000	13600	8400	800	400	17	21	Zobeid
9600	4800	800	4000	800	400	1	10	ElTahteya
4800	2400	800	1600	800	400	1	4	Elderihmi
20000	10000	4800	5200	800	400	6	13	Eldoha
5600	2800	0	2800		400	0	7	Ellehya
3200	1600	0	1600		400	0	4	Elmoneira
12800	6400	0	6400		400	0	16	ElZahra
8800	4400	1600	2800	800	400	2	7	Elmokhlaf
366000	**183000**	**46400**	**123600**			**43**	**343**	**TOTAL**

Appendix 3.4 Instructions to health-care providers (Yemen 2010)

Healthcare providers' name:

Subject: Registration of tuberculosis cases detected in the public and private health care facilities.

The Ministry of Public Health, with technical support from the World Health Organization, is conducting a project aiming at registering all tuberculosis cases detected in public and private healthcare facilities in your governorate.

In this connection, we urge you to:

- Register all tuberculosis suspects or cases, pulmonary and extra-pulmonary, identified and diagnosed in your facility using the standard forms enclosed herewith and as explained by the field researchers;
- Collaborate with the field researchers in conducting this study and facilitate their visits (every 2 weeks) aiming at monitoring progress, evaluating the quality of field work implementation and collecting the completed forms;
- Monitor and supervise the completion of the forms by your staff and solve problems faced by the field workers, if any.

We look forward for your active participation for controlling this public health problem.

Signed by: NTP manager, directors of surveillance and primary health care.

Copies to: Minister, director generals of the targeted governorates and all field workers during their initial visit to map the facility and enrol the healthcare provider in the study.

References

1. *Tuberculosis prevalence surveys: a handbook*. Geneva, World Health Organization, 2007 (also available at: http://www.who.int/tb/advisory_bodies/impact_measurement_taskforce/resources_documents/thelimebook/en/index.html; accessed July 2012).

2. *Electronic recording and reporting for tuberculosis control*. Geneva, World Health Organization, 2012 (http://www.who.int/tb/publications/electronic_recording_reporting/en/index.html; accessed April 2012).

3. U.S. Government Printing Office, 1949. Permissible medical experiments on human subjects. [Nuremberg Code]. In: Trials of War Criminals before the Nuremberg Military Tribunals under Control council Law No. 10, Vol. 2. p.181-182. Washington, D.C.

4. *Declaration of Helsinki*. Ferney-Voltaire, France, World Medical Organisation, 1964.

5. *International ethical guidelines for epidemiological studies*. Geneva, Council for International Organizations of Medical Sciences, 2009.

6. *Guidance on ethics of tuberculosis prevention, care and control*. Geneva, World Health Organization, 2010 (also available at: http://www.who.int/tb/features_archive/ethics/en/index.html; accessed November 2011).

7. *Guidelines for surveillance of drug resistance in tuberculosis*. Geneva, World Health Organization 2009. http://www.who.int/tb/publications/2009/surveillance_guidelines/en/index.html

8. Baker R et al. What proportion of patients refuse consent to data collection from their records for research purposes? *British Journal of General Practice*, 2000, 50(457):655–656.

9. Huang N et al. Record-linkage research and informed consent: who consents? *BMC Health Services Research*, 2007, 7:18.

Chapter 4
Record-linkage

Author: Ana Bierrenbach

This chapter is intended primarily for epidemiologists and statisticians who are involved in or have an interest in inventory studies. However, the key principles and methods are expected to be understood by a more general audience.

Record-linkage is a process that aims to accurately identify whether two or more records relate to the same individual. When consolidating several different databases into one central database, record-linkage is used to find duplicates. Record-linkage can be performed manually by visually comparing records. However, this becomes labour intensive, tedious and inefficient as the number of records increases. The process of record-linkage is usually performed on a computer; technological advances in software have made it possible to perform efficient and accurate record-linkage within or between large datasets.

Record-linkage is an essential part of inventory studies used to estimate TB under-reporting. Inventory studies require different sources of patient information to be collated in order to obtain an accurate estimate of diagnosed TB cases, including data from public case notifications, private clinicians and laboratory registries. In the course of the disease, a patient may attend several health-care facilities and be counted more than once when the data from public and private practices is compiled in an inventory study. Record-linkage enables the identification of all records belonging to the same individual, so that each case is only counted once. It is important to remove duplicates because if this is not done, the number of diagnosed TB cases counted in the study may be artificially increased and the results may be skewed.

This chapter outlines how to perform record-linkage as part of a TB inventory study of under-reporting. It presents the two major types of record-linkage (deterministic linkage and probabilistic linkage) and describes the main steps in the record-linkage process, including the tools and techniques that can be used to implement it. An example of record-linkage based on the kind of data that would be generated from an inventory study in Egypt is used to illustrate the methods that are described. Finally, the different record-linkage software programs are summarized.

4.1 Deterministic and probabilistic record-linkage

In deterministic record-linkage, two records are said to match if one or more individual identifiers, such as name or age, are identical. In contrast, if identical individual identifiers are not available, probabilistic record-linkage matches two records based on a score that reflects the probability that the records relate to the same person. The choice of deterministic or probabilistic record-linkage depends on the purpose of the linkage. Some researchers believe that the hierarchical rules in deterministic record-linkage give better control over the specificity of the matches [1]. If this is so, where the vital status of individual patients is being studied, it is be best to use deterministic methods. However, when the linkage is done for the purpose of studying population-based characteristics, probabilistic methods could be used given that the observed false-positives and false-negatives would tend to cancel out. However, not many comparative studies have been performed to accept or reject this idea.

4.1.1 Deterministic record-linkage

In deterministic record-linkage, the detection of duplicates involves finding all records that contain exactly the same data in one or more fields. The process is far easier if there is one personal identifier that is common to all databases used. A personal identifier should have the characteristics outlined below. It must be:

1) Unique – no two people should share the same identifier;
2) Universal – available to every member of the population under study;
3) Permanent – should remain unchanged throughout the study period;
4) Accurate – allow no, or very little, mistakes;
5) Reasonable – bring forth no objection to its disclosure to intended users;
6) Simple;
7) Known.

The United States Social Security number and the United Kingdom National Health Service number are examples of personal identifiers that could be used in inventory studies performed in these countries. If a single, common personal identifier is not available, a combination of variables can be put together to allow the unique discrimination of records, for example, the combination of "name of patient", "name of patient's mother" and "date of birth". It is advisable not to base deterministic record-linkage on a single identifier such as the "name of patient" variable, as homonyms exist in most countries.

4.1.2 Probabilistic record-linkage

Probabilistic linkage is defined as record-linkage that is based on a score that reflects the probability that the records relate to the same person or entity. Probabilistic methods can be divided into classical [2] and newer approaches [3]. The classical approach follows the theory that was developed by Fellegi and Sunter in 1969 [2]. The probabilities are calculated taking into consideration how similar each of several matching fields are and also how frequent the values observed on the matching fields are in relation to those of the other records in the files being compared, or even to the values observed in the population as a whole. To compare the similarity of matching fields, probabilistic linkage software programs make use of the edit distance and similarity measures discussed later in this chapter.

Probabilistic record-linkage is uncertain in nature and should only be used in the absence of unique, exact and reliable individual identifiers.

Computer-based calculations are made based on the **discriminating power** of each field. Agreement or disagreement weights are assigned for each field and these weights are usually added together to get a combined score that represents the probability that the records refer to the same person. This is how the classical Fellegi and Sunter model works [2]: it simply sums up all the weights to produce a total weight for each pair, and then uses two thresholds to classify the pair into one of the three classes: matches, non-matches or possible matches. The results of a classification are stored in an ordered way in the dataset to be presented to the users. A detailed discussion of match weights and probability matching can be found in many publications [2-5]. The sensitivity of the linkage process can usually be improved by performing several cycles (or passes) over the data using different matching and blocking fields, and submitting the matching variables to different algorithms for approximate string comparisons.

4.2 The main steps in the record-linkage process

Record-linkage consists of several steps: pre-processing; the selection of matching variables; blocking or indexing; searching and scoring, and a manual review. These steps are outlined in detail below. Apart from the scoring step which applies to probabilistic matching only, all other steps apply to both deterministic and probabilistic linkage processes. The differences in their use and relevance are highlighted in each subsection.

4.2.1 Pre-processing

Pre-processing of data is an essential first step in the record-linkage process. The data to be used may be recorded in different formats and may contain errors, inconsistencies and missing items. The aim of the pre-processing phase is to clean and standardize the data used in the record-linkage process. Possible data transformations include:

- Removal of commas and other punctuation marks. This is particularly important if the

file has to be transformed into a delimited format, such as "Comma separated variable format" (CSV), in order to be used by the linkage software program;
- Removal of leading, trailing or internal unnecessary blanks;
- Correction of upper/lower case variations;
- Removal of numbers from variables that should be purely composed of letters and vice-versa;
- Removal of accent marks;
- Removal of terms indicating lack of information (e.g. "don't know", "unknown", "unidentified", "n/a", ..);
- Standardization of date formats;
- Standardization of terms used in addresses (e.g. "St." could be replaced by "Street", "Sq." by "Square",..);
- Standardization of the order of the address elements;
- Replacement of obvious spelling variations with standard spelling for common words and names.

Some record-linkage software programs have basic pre-processing functions. However, to perform thorough pre-processing of the data, a more robust alternative is to make use of general purpose **programming languages** like Python, Perl and Ruby, or more specialised languages such as R. Moreover, knowledge of **regular expressions** is essential to allow for matching complex patterns of text with minimal effort [6].

Several methods outlined below may be used in the pre-processing stage to increase the ability of record-linkage to find related records.

4.2.1.1 Parsing

Parsing involves partitioning a string variable into its component parts, for example into the first, middle and last names of a person. Parsing can be used to increase the sensitivity of the deterministic linkage in finding related records: individual fields make better duplicate detection candidates if the field is split into two or more fields.

4.2.1.2 Substringing

Substringing is another method used to increase sensitivity. It involves separating a determined number of characters from a string. Obviously, other pieces of data would still need to be used in order to determine whether a link should actually be classified as a match. A trade-off exists between an increase in sensitivity and the undesired consequence of loss of specificity.

4.2.1.3 Phonetic coding systems

Phonetic coding involves coding a string based on how it is pronounced, so that it can be linked despite minor spelling or typographical differences. The **Soundex system** is used for phonetic coding. It was developed in 1918 and is based on how names are pronounced in English. Its code consists of the first letter of the name followed by three numerical digits representing the remaining consonants. Similar sounding consonants share the same digit. The following steps are used to transform a name into a Soundex code:

1) Keep the first letter of the name and drop all other occurrences of a, e, h, i, o, u, w, y
2) After the first letter, replace consonants with digits as follows:
   - b, f, p, v => 1
   - c, g, j, k, q, s, x, z => 2
   - d, t => 3
   - l => 4
   - m, n => 5
   - r => 6
3) Replace adjacent letters coded with the same number as a single number

Continue until you have one letter and three numbers. If the name is short with less than three consonants to be replaced, fill in 0s until there are three numbers.

Some examples of Soundex codes are outlined Table 4.1 below.

Table 4.1: Examples of Soundex codes

gandhi	G530	jeberson	J162	john	J500	mohamed	M530
gandii	G530	jeferson	J162	johny	J500	monat	M530
gante	G530	jeferzon	J162	jonas	J520	muhamed	M530
ghandi	G530	jefferson	J162	jonathan	J535	muhammed	M530
ghandy	G530	joversen	J162	jonnotahn	J535	nuhamed	N530

A far more discriminative code may be created by adding together the codes for the first, middle and last names of a person, e.g. *Mohammed Abdullah*=>M530A134.

The Soundex system works well with Anglo-Saxon and many European names, but not for names which are short, as is the case for many very common names, names with a high percentage of vowels, or some names that are not Anglo-Saxon in origin. Variants of the coding system have been developed to address these limitations, such as Phonex, NYSIIS and Double-Metaphone.

Phonetic coding is often used to create blocking variables in the probabilistic linkage process. In the deterministic linkage process, it creates modifications of the original matching variables to be compared in between records.

4.2.2 Selecting matching variables

Suitable variables must be selected to match records. The variable should be chosen based on its ability to discriminate between different records. If a variable has many different values then it has a higher discriminative power. For example, a comparison between two different records containing the same last name has greater discriminating power if the name is rare rather than common. Variables with high proportions of missing values are not very useful for matching variables. As methods for probability matching depend on making comparisons between each of several variables with identifying information, variables like "name", "date of birth", "name of mother" and "address" are commonly used jointly. If it was not possible to satisfactorily arrange the order of the elements in the "address" variable in the pre-processing phase, it may be best not to choose it as a matching variable. It may be used as a matching variable in later linkage cycles or it may only be used in the manual verification of uncertain linked pairs.

4.2.3 Blocking

Potentially, each record in one file has to be compared with every record in a second file. For files with large numbers of records, the total number of possible pairs is too large for practical computation. While the number of records to be compared increases linearly, the computational task, and therefore the time spent in performing searches, increases quadratically. To reduce the number of comparisons, blocking(indexing) techniques are typically used. The data sets are split into smaller blocks using one or more blocking variables, and only within these blocks are records compared between the files.

"Sex" may be a good blocking variable in the sense that not many records are likely to be wrongly classified or to have missing variables in this field. However, blocking by sex only splits the file into two parts. Choosing "district of residence" as a blocking variable will certainly have a higher impact in increasing the efficiency of searching. However, the records of patients that have moved from one district to another during the study period will not be paired. Phonetic codes of names or last names, date of birth and postcode are commonly used as blocking variables.

4.2.4 Searching and Scoring

This is the core of the linkage process and the phase in which the computer will do the work for you. In the probabilistic linkage process, the computer program searches probable pairs of records, estimating the probability that the pairs relate to the same person, calculating linkage scores and displaying the results to the user in an ordered way.

In the deterministic linkage process, several consecutive matching strategies are combined in a hierarchical way, beginning with the use of the more specific / less sensitive ones [8]. The computer program does not actually calculate scores, but it does display to the user in which strategies the pairs of records were created.

Both processes make use of so-called "edit distance measures". These measures are essential in probabilistic linkage as the numerical values they return are used to compute matching weights for string fields like names and addresses. In the deterministic linkage process, they may be used as part of one or more matching strategies, as shown below.

Edit distance measures, also known as string comparator metrics, are used to compare strings at the character or the term level. The **Levenshtein distance** [3] works on the character level and counts the minimum number of deletions, substitutions and insertions required to transform one string into the other. The greater the distance, the more different the strings are.

As longer strings are more likely to have greater Levenshtein distances, some corrections for the length of string have been proposed. For example, a **similarity measure** between 0 and 1 may be obtained by subtracting the distance from the length of the longer string and then dividing the result by the length of the longer string.

An example of Levenshtein distance and similarity measure is illustrated in Box 4.1.

Box 4.1: An example of Levenshtein distance and similarity measures

Frankenstein X Fronckensteen

Distance = 3 (substitution of "a" with "o", insertion of "c" and substitution of "i" with "e")

Similarity = 0.77 = (13-3)/13

The Levenshtein distance and some of its simple variants, like the **Damerau-Levenshtein distance** and the **Hamming distance** are mostly suited for handling spelling or typographical errors [4, 5]. The Damerau-Levenshtein distance allows not only deletions, substitutions and insertions but also the transposition of two adjacent characters, like *John* and *Jonh* while the **Hamming distance** allows only substitutions.

Other string comparator measures have been developed to deal with more complex variations in name representations [7]. They take into account not only the length of the strings being compared, but also the kind of errors that are more frequently made and where in the string (beginning/middle/end) these kinds of errors are more likely to occur. Different kinds of errors in different places of the string are given different weights. There are some string comparator measures, like the **Smith-Waterman distance**, that give a smaller "penalty" for abbreviations in the middle names (*John Kevin Smith* X *John K Smith*) and for simple transpositions of the terms in a string (*John Smith* X *Smith John*). Others, like the **Jaro-Winkler distance**, are fit for names that tend to have errors towards the end of the string (*Abdullah* X *Abdullag*), among other attributes.

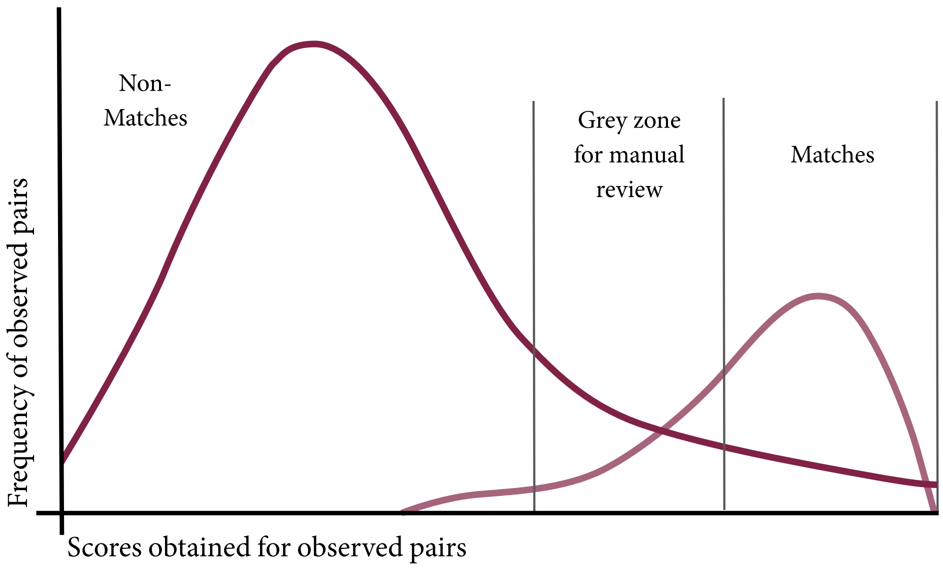
An example of a hierarchical algorithm used in deterministic linkage is provided in Table 4.2.

Table 4.2: An example of a hierarchical algorithm used in deterministic record-linkage

Rules	Name of patient	Date of birth	Name of patient's mother
1	Exact	Exact	Exact
2	Exact	Exact	Same Soundex code
3	Exact	Levenshtein distance of 1	Exact
4	Exact	Levenshtein distance of 1	Same Soundex code
5	Same 4-character substring	Exact	Exact

In a typical probabilistic record-linkage, the histogram of the score frequency of observed pairs will show a bimodal distribution (Figure 4.1). There is a bigger mode of pairs with low scores and a smaller mode with high scores. Pairs around and below the bigger mode can usually be classified as non-matches without a need for further review. The same is true for pairs around and above the smaller mode, which can be directly classified as matches. The problem lies in the so called grey zone in between the two modes, where pairs demand further review so as to be satisfactorily classified.

Figure 4.1: Distribution of scores obtained for observed pairs



It takes some experience with probabilistic linkage to be able to select score cut-off points that will separate the matches from non-matches. The process tends to be a trade-off between being confident that all matches are correctly identified and not leaving too much work for the manual review process. It is preferable to start with the default cut-off values recommended by the selected software program.

4.2.5 Manual review

The final step in record-linkage is the manual review. The manual review is the process of manually looking over uncertain linkages in the grey zone and then classifying them as matches or non-matches. When manually reviewing data, intuition is used in combination with an intrinsic knowledge of the frequency of names and addresses in the population to help decide whether paired records relate to the same person, even if they contain slight variations or missing information.

In theory, the person undertaking the review has access to additional data from variables not used in the searching and scoring phase, or even data external from the files being compared, which enables them to resolve the linkage status. Importantly, after all pairs have been classified, it is best to **not** delete the unwanted duplicated records. Instead, duplicated records should be **marked** as duplicates and kept on file, to allow a subsequent reassessment. Any record-linkage exercise should be accompanied by full documentation of the methods used. The documentation is necessary for two main purposes: to allow peer review and to provide a record of what has been done for possible replication in the future.

4.3 An example of record-linkage

We will now use a practical example to illustrate the process of record-linkage in a TB inventory study. Let us consider a hypothetical inventory study in Egypt, resulting in two large files that need to be cross-checked for the presence of patients common to both files. The first file is from the NTP (the National TB Control Programme);the second is from National Mycobacteria Laboratory registries. As the files are large, the identification of possible matches across the two files would not be a task to be done by visually comparing records, no matter how trained and committed the local study team members may be. Record-linkage should be done electronically using appropriate software programs. The following are fictitious examples of four records from each of the files:

NTP registry

Name of patient	Date of Birth	Address
1. Mohammed Abdullah	24/05/1972	10, El Tahrir Square
2. Adib Omar Karim	17/07/1945	Tora Prison
3. Piotr Sviatopolk-Mirskii	10/03/1925	150, Hagar Nawatia
4. John K. Smith	13/09/1989	1121(A), 26th July Street

Laboratory registry

Name of patient	Date of Birth	Address
1. Mohamed Abdalah	May, 1972	P.O. Box 134
2. Adib Amr Korayem	July, 1945	Tora Prison Complex
3. Piotr Sviatopolk-Mirsky	-	-
4. Jonathan K. Smith	September, 1989	26th July St, 1121

A **deterministic record-linkage** approach using unmodified fields would probably find some matches across the two files if the name of the patient or the address were used as matching variables. However, the records shown above do not have exact **matching fields** in any of the variables and therefore would not be paired. Fortunately, at least part of this problem could be solved if we were to **pre-process** the files before the linkage, putting the data into standardized coding formats, eliminating inconsistencies such as spelling errors and punctuation and inconsistent use of upper and lower cases. Pre-processing increases the likelihood of matching records and is essential prior to both deterministic and probabilistic methods.

If we first look at the "name of patient" variable, we must consider that the differences in the spelling of the first three names may be exclusively due to problems in the transcription from the Arabic to Latin alphabet for the first two records, and from the Cyrillic to Latin alphabet for the third record. If we take the first record and consider the four most common versions of the name Mohammed and the six most common versions of the name Abdullah we would get twenty-four legitimate versions of this patient's name, without even adding other possibilities due to misspellings and abbreviations. Such spelling problems would be less likely to happen if the original alphabet or language of the country were used in the files.

As the misspelling errors for the first and second records are very common and easy to predict, one possible approach could be to generate a variable in which the various versions of the name would be modified into only one, and use this new variable as our matching field. For example, all twenty-four translation versions for the first pair could be uniformly presented as *Mohammed Abdullah*. Similar modifications could be made for the Arabic names most commonly used in Egypt, and this would greatly impact in the usefulness of the "name of patient" variable as our matching field.

If we next look at the "date of birth" variable, it is noted that the laboratory file does not have information on the days on which patients were born, just which months and years. The only way to standardize this variable on both files with the intention that it can be used as an exact matching field would be to modify the one from the NTP registry, so that it would also have just months and years. We are losing some information by doing this, but we are also gaining the possibility of pairing these records so that the decision about whether or not they relate to the same person can be taken on a later phase of the linkage process.

The "address" variable can also be standardized on the use of abbreviations for street and square and the order that the elements of the address appear. Address is obviously not an identifier *per se*, but it may help decide whether or not pairs of records identified using other variables relate to the same person.

Once the files are pre-processed, they would typically be imported into a specialised software program if **probabilistic linkage** is to be used. The software would use the information from the selected matching fields to calculate a **linkage score** that indicates, for any pair of records, how likely it is that they both refer to the same person. Potential pairs would then be automatically accepted or rejected based on a defined threshold, a **cut-off value**. Since there would probably still be a grey area of uncertain matched pairs above the threshold, they would need to undergo **manual review** in order to determine whether they actually relate to the same person.

Do our four pairs of records relate to the same four patients? It is not so easy to classify pairs of records based on only these three variables:

1. *Mohammed Abdullah* is a very common name in Egypt, the date of birth stored in the laboratory registry is incomplete and the two addresses are not comparable. However, if TB is not a very prevalent disease in the studied area, it would be improbable to find two TB patients named Mohammed Abdullah born on the same month and notified in the same time period.
2. *Adib Omar Karim* is not a common name and it is unlikely that two different TB patients with this name would be born in the same month of the same year, and have Tora Prison as their common address. Therefore, the pair of records probably do relate to the same patient. In this situation, it may also be helpful to know how many prisoners are detained in Tora Prison.
3. The third pair of records is also likely to belong to the same person, since *Piotr Sviatopolk-Mirskii* is a very uncommon name in Egypt, and also in Russia. It would be virtually impossible to have two TB patients in Egypt with such a name. The only possibility would be for them to belong to the same family, for example a father and son. This possibility has to be considered, particularly as we are dealing with an infectious disease that spreads within a household.
4. *John* and *Jonathan K. Smith* could be twin brothers living at the same address, flat mates with no family relation or even the same person if, for example, the name Jonathan was misunderstood or just misspelled while the NTP record was being generated.

To know for sure if these pairs relate to the same patients, if there are only a few pairs to check, we could go back to the health care facility or laboratory where the patient was notified to obtain more identification details, or we could contact the patients themselves. However, if there are many doubtful pairs to check, it would be more practical to start by checking the differences and similarities across other variables that might be available in both files, for example, telephone number, name of the patient's mother or the dates of notification and sputum submission. The manual or semi-automated verification of linked records (or **links**) found by the linkage software is often called **post-processing**, and the verified pairs are often called matched records (or

matches, i.e. relating to the same person or entity).

In summary, record-linkage is used to detect pairs of matched records. Errors in the linkage process will impact the results of inventory studies by producing an overestimation of TB burden when records that correspond to the same patients do not link due to missing or inaccurate data (false negatives), or an underestimation of TB burden when unrelated records are mistakenly matched (false positives).

4.4 Record-linkage software programs

Several software packages that implement computational models for record-linkage have been developed over the last two decades. Time and effort is required to decide which one better meets the specific needs of an inventory study. Some operational aspects and methodological issues need to be considered including the data management skills of staff responsible for performing the linkage procedures, the availability of hardware and the operating system used.In relation to the software program the following issues need to be considered:

- Record-linkage methodology;
- Features included, for example pre- and post-processing functions - this may determine whether the software can stand alone or needs to be complemented by other programs or languages;
- Data and system requirements;
- Performance in terms of memory use and anticipated time necessary to perform tasks;
- Flexibility to deal with different file formats, different languages and different scripts (e.g. Roman, non-Roman) used in the matching fields;
- Availability of documentation with comprehensive description on the functions, mathematical models, system options and procedures available;
- User-friendliness and language of the interface;
- Cost of licenses, personnel experience and preferences, training costs for personnel and for technical assistance.

Some software comparisons and ratings have been developed that take into account different operational aspects and methodological issues [3, 10-12]. The main features of a few of the open-source probabilistic linkage software programs are presented below (based on versions available in October 2012).

4.4.1 FEBRL

Freely Extensible Biomedical Record-linkage (FEBRL) is an open-source software program developed by the Australian National University. It does data standardization (parsing and cleaning) and implements probabilistic record-linkage to perform de-duplication (same database) and record-linkage between two files. It comes with very good documentation and can be used to import files in different formats, such as TXT and CSV. A large number of functions for blocking and comparing are available. Other applications are needed to further processing the output files. It runs on Mac, Linux and Windows operating systems. The different installation requirements are well explained in the installation notes.

Available at : http://sourceforge.net/projects/febrl/

4.4.2 RELAIS

RELAIS is a freely available software program developed by the Italian National Statistics Institute. RELAIS is based on MySql, and allows the use of different formats of the files being compared. RELAIS uses Java and R languages, which makes it quite flexible and adaptable. It comes with excellent documentation, high performance and its interface is quite easy to use. As the focus of the software is the matching step, no pre and post-processing functions are available, which may be a problem if there is a lack of skilled data management personnel in the study team (although as noted in Chapter 3, a data manager should be a core member of the team in an inventory study). It runs on Linux and Windows operating systems.

Available at: http://www.osor.eu/projects/relais

4.4.3 LinkPlus

LinkPlus is a freely available software program developed by the United States Center for Disease Control and Prevention. As with the other software on this list, it performs de-duplication and record-linkage. The accompanying documentation is not very extensive. The interface is very user friendly and intuitive, and the second version of the software includes some post-processing features. The software works mostly with text delimited files, such as CSV. Pre-processing has to be done mostly in other applications, which may limit its use for users with restricted data managing skills. Linkplus runs on Windows.

Available at: http://www.cdc.gov/cancer/npcr/tools/registryplus/lp.htm

4.4.4 OPENRECLINK

OPENRECLINK is an open-source software programme developed by researchers at the Federal University of Rio de Janeiro in Brazil. The interface is available in English and Portuguese.

The current manual is in Portuguese, but an English version is being developed. In the newer version of the software, the interface has been greatly improved. Features are available for performing many pre- and post-processing functions. RECLINK allows the easy implementation of sequential linkage cycles in which different inputs may be chosen, including different matching and blocking variables and cut-off values. OPENRECLINK requires Mac, Linux or Windows. I

Available at: http://reclink/sourceforge.net/

4.4.5 R

The RecordLinkage R package implements record-linkage in the R environment. Flexibility and high performance are its main advantages, but the interface is not user friendly and requires the use of the command line. Pre-processing can be done using core R functions and knowledge of the R language is a requirement. R runs on a variety of Unix platforms (including Linux), Windows and Mac.

Available at http://www.r-project.org and the RecordLinkage package at
http://cran.r-project.org/web/packages/RecordLinkage/index.html.

References

1. Clark DE. Practical introduction to record-linkage for injury research. *Injury Prevention*, 2004, 10(3):186–191.

2. Fellegi IP, Sunter AB. A theory for record-linkage. *Journal of the American Statistical Association*, 1969, 40:1183–1220.

3. Herzog TN, Scheuren F, Winkler WE. *Data quality and record-linkage techniques*. New York and London, Springer, 2007.

4. Newcombe HB. *Handbook of record-linkage: methods for health and statistical studies, administration, and business*. Oxford and New York, Oxford University Press, 1988.

5. Jaro MA. Probabilistic linkage of large public health data files. *Statistics in Medicine*, 1995, 14(5–7):491–498.

6. Friedl JEF. *Mastering regular expressions*, 3rd ed. Sebastapol, CA, O'Reilly, 2006.

7. Bilenko M, Mooney RJ. *Learning to combine trained distance metrics for duplicate detection in databases* (available at http://www.cs.utexas.edu/~ml/papers/marlin-tr-02.pdf ; 2002.

8. Pacheco AG et al. Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in Brazil. *American Journal of Epidemiology*, 2008, 168(11):1326–1332.

9. Bierrenbach AL et al. Duplicates and misclassification of tuberculosis notification records in Brazil, 2001-2007. *Int J Tuberc Lung Dis*, 2010, 14(5):593–599.

10. Campbell KM, Deck D, Krupski A. Record-linkage software in the public domain: a comparison of Link Plus, The Link King, and a 'basic' deterministic algorithm. *Health Informatics Journal*, 2008, 14(1):5–15.

11. Silva AD et al. *Study of record-linkage software for the 2010 Brazilian Census post enumeration survey* (available at: http://isi2011.congressplanner.eu/pdfs/450055.pdf; 2010.

12. Ferrante A, Boyd J. A transparent and transportable methodology for evaluating Data Linkage software. *Journal of Biomedical Informatics*, 2012, 45(1):165-172.

Chapter 5
Data analysis and reporting

Authors: Philippe Glaziou, Ross Harris, Charalambos Sismanidis, Amal Bassili, Fulvia Mecatti and Rob van Hest

Data analysis and reporting is the final stage in the implementation of an inventory study. The data from the study should first be described and analysed and then the final inventory study report must be prepared. The final report should include the results of the study and the level of under-reporting determined, as well as the study methods, limitations and recommendations for how to improve TB surveillance and the extent to which PPM efforts need to be strengthened. The following subsections provide an overview of how to describe and analyse the data. Full details of how to perform record-linkage and how to undertake a capture-recapture analysis (if appropriate based on the study design) are provided in Chapter 4 and Chapter 6, respectively. Chapter 1 describes the conditions that must be met for capture-recapture analysis to be applied and Chapter 2 describes the associated study design.

5.1 Data description

The completeness of the core data should be documented. For each selected geographical area, a table summarizing the number of cases recorded by each type of health-care provider should be provided in the study report. Among eligible providers, the number who refused to participate should also be documented. Record matching, which is described in detail in Chapter 4, will generate counts of cases by source. Counts of cases should be produced for each of the geographic areas included in the study, following the example in Table 5.1.

Table 5.1: Example of output from the record-matching analysis, Egypt 2007 [1]

Geographical area-level data obtained from the authors. Column *freq* indicates the number of TB cases for each combination of sources and for each geographical area. For example, in area 1, there were 11 cases seen only by private providers, 61 seen by the NTP only, 16 seen by the NTP and private providers, and 12 seen by the NTP and public providers.

ntp	public	private	area	freq
0	0	1	1	11
1	0	0	1	61
1	0	1	1	16
1	1	0	1	12
0	0	1	2	7
1	0	0	2	46
1	0	1	2	5
1	1	0	2	6
1	1	1	2	1
0	0	1	3	21
0	1	0	3	5
1	0	0	3	131
1	0	1	3	18
1	1	0	3	58
0	0	1	4	2
1	0	0	4	9
1	0	1	4	1

The ratio of the total number of NTP cases to the total number of diagnosed cases provides a crude estimate of the level of under-reporting. The data in Table 5.1 can easily be summarised as in Table 5.2 below. The crude estimate of the overall level of under-reporting in this example is 11.1%.

Table 5.2: Aggregated data, Egypt 2007

Column *under-reporting* is calculated as the ratio of non-NTP cases over the *total* number of diagnosed cases. Four geographic areas (clusters) were selected by sampling. To do this, first four strata were defined based on quartiles of case notification rates. Each strata contained 6 or 7 geographical areas (column nAreas), and one area was selected in each strata.

nAreas	NTP	**non-NTP**	**total**	**under-reporting**
1	89	11	100	0.11
2	58	7	65	0.108
3	207	26	233	0.112
4	10	2	12	0.167

5.2 Adjustments for sampling design

If providers were sampled (through the sampling of geographical areas), as will often be the case in an inventory study (Chapter 2), then the sampling approach needs to be taken into account. Failure to do so will almost certainly understate the uncertainty surrounding the point estimate of under-reporting (that is, its standard error) and may also affect the point estimate itself. Variance estimation for estimators depends upon the sampling plan specifics and requires approximate methods, generally Taylor series linearization or replication techniques. Variance estimation for ratios of population totals obtained from cluster sampling with or without stratification may be done using the method of Taylor series linearization (also known as the Delta method) [5] as shown in examples below.

Table 5.3: Aggregated data, Iraq 2011, preliminary results

Column *ntp* shows the number of NTP-reported cases during the period of the survey, *total* is the total number of individual TB cases diagnosed during the study period, including those reported to NTP, *stratum* is an identifier for strata and column *fpc* (finite population correction) shows the total number of areas within each stratum (additional details about these data are provided in Table 6.5, Chapter 6).

area	**ntp**	**total**	**stratum**	**fpc**
Basrah	207	225	1	4
Duhok	90	121	1	4
Misan	56	77	2	4
Najaf	72	82	2	4
Baghdad	693	866	3	4
Sulaymania	177	209	3	4
Diwanyia	170	183	4	5
Wasit	208	217	4	5

To illustrate sampling design effects and implement design-adjusted analyses, we will use unpublished preliminary data from a recently completed inventory study conducted in Iraq (data obtained and used with permission from the principal investigator). We will use the free statistical software ‘R’ which is explained more fully at the end of Chapter 4. A total of 18 geographical areas constituted the sampling frame, covering the whole country except for areas unreachable due to security concerns. The geographical areas were grouped into 4 strata defined by quartiles of the case notification rates and 2 areas from each stratum were selected. Table 5.3 shows the study data by sampled area.

The data are stored into a comma-delimited text file and are first loaded into R.

```
irq <- read.csv(‘iraq.csv’)
```

Let us first assume that areas were sampled through simple random sampling, ignoring the stratification. In the following, we will use the R package “survey” to compute totals and ratios adjusted for sampling design effects. The following lines show R’s output in a console, commands are echoed following > at the beginning of a line (in bold face). Comments are added in the code below (following the # symbol) to provide brief explanations. More details on a specific command can be obtained by typing help(command) in an R console.

```
> irq$fpc2 <- sum(irq$fpc) / 2 # total number of areas in the sampling frame
> library(survey) # loads the survey package
> dclus <- svydesign(id=~1, fpc=~fpc2, data=irq2) # specifies sampling design
> summary(dclus) # summary statistics from object dclus
Independent Sampling design
svydesign(id = ~1, fpc = ~fpc2, data = irq2)
Probabilities:
       Min.   1st Qu.  Median  Mean   3rd Qu.   Max.
       0.471   0.471   0.471   0.471   0.471   0.471
Population size (PSUs): 17
Data variables:
[1] “gov”  “ntp”  “total”  “stratum”  “fpc”  “nonntp”  “pw”  “fpc2”  “pw2”
[10] “under”
```

A finite “population” size is introduced, defined by the total number of areas in the sampling frame [2, 3]. Totals are computed as follows [4]:

```
> svytotal(~total, dclus)
      total   SE
total  4208 1124
```

The total number of diagnosed cases estimated in the country over the study period was 4208 (Standard Error (SE) 1124). Since a large proportion of the areas were sampled (8 out of a total of 17), standard errors are smaller than would otherwise be observed without a finite population correction. Under-reporting is obtained through a similar approach:

```
> irq$nonntp <- irq$total - irq$ntp # assign a new variable to data frame irq
> u.dclus <- svyratio(~nonntp, ~total, dclus)
> u.dclus$ratio # prints the proportion nonntp
        total
nonntp 0.1551
> confint(u.dclus) # returns a 95% confidence interval for the proportion nonntp
              2.5 % 97.5 %
nonntp/total 0.1138 0.1963
```

The estimated level of under-reporting in Iraq is 15.5% (95% CI 11.3 – 19.7), assuming simple random sampling of areas. Stratification before sampling, if based on a factor (or combination of factors) associated with the outcome, results in reduced standard errors.

```
> irq$pw <- irq$fpc / 2 # population weights, set equal within strata
> dstrat <- svydesign(id=~1, strata=~stratum, weights=~pw, fpc=~fpc, data=irq)
> svytotal(~total, dstrat, deff=TRUE)
      total   SE DEff
total  4160  943 0.75
> u.dstrat <- svyratio(~nonntp, ~total, dstrat, deff=TRUE)
> u.dstrat$ratio
        total
nonntp 0.1502
> confint(u.dstrat)
              2.5 % 97.5 %
nonntp/total 0.1143 0.1861
> deff(u.dstrat)
[1] 0.6823
```

The design-adjusted level of under-reporting is 15% (11.4 – 18.7%). Stratification improved the sampling efficiency by reducing within-stratum variability. The design effect is 0.75 for the estimated total number of diagnosed cases, and 0.68 for the level of under-reporting, compared with a design effect of 1 with the use of simple random sampling. A smaller than 1 design effect is not expected in the case of study design 1 (see Chapter 2) where clustering effects will negatively affect sampling efficiency.

5.3 Missing data

Two types of missing data may occur: failure to recruit an identified provider (e.g. refusal to participate) and failure to obtain all requested information from an enrolled provider (incomplete records). These can be referred to as unit non-response and item non-response, respectively.

The bias from unit non-response can be mitigated by adjusting sampling weights, thus modelling the non-response as part of the sampling mechanism.

For item non-response (e.g. incomplete information on TB diagnostic test results), two types of approach can be considered. One is to model the non-response as part of the sampling mechanism. This type of approach is suitable when data are missing completely at random (MCAR), i.e., the missing data are not related to the value of the outcome and to an individual characteristic that is a risk factor for the outcome (e.g. type of provider, geographical area). If data are MCAR, the analysis can be restricted to the complete records and an unbiased estimate of under-reporting will be obtained. However, it is very unlikely that MCAR will be true overall.

The second approach is to impute the missing data using the observed information on each record as a guide to plausible values for the missing information. The necessary assumption underlying multiple imputation of missing data states is that all outcome differences between complete and incomplete records are explained by the variables for which data are complete. When the probability that a record has missing data for a variable that is related to other recorded individual variables such as "geographical area" or "type of provider", the missing data are said to be missing at random (MAR). Within groups of records with the same auxiliary variables, the probability of data being missing on the outcome variable is not associated with its value. While the MAR assumption is not a priori plausible, multiple imputation to address item non-response is expected to reduce bias when compared with an analysis that simply drops the missing records. Multiple imputation of missing data can be implemented in R [5] (e.g. using package mice [6] or package Amelia [7]) and is described in [8] and accompanying web materials. Analysis with sampling design adjustments can be made on multiple imputed datasets using the survey package [4].

References

1. Bassili A, Grant AD, El-Mohgazy E, Galal A, Glaziou P, Seita A, Abubakar I, Bierrenbach AL, Crofts JP, van Hest NA. Estimating tuberculosis case detection rate in resource-limited countries: a capture-recapture study in Egypt. *Int J Tuberc Lung Dis*, 2010, 14(6):727-32.

2. D. G. Horvitz DG, Thompson DJ. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 1952, 47(260):663- 685.

3. Tillé Y. *Sampling Algorithms*. Springer Series in Statistics. Springer, New York, 2006.

4. T. Lumley. *Survey: analysis of complex survey samples*. 2006. R package version 3.26.

5. http://www.r-project.org

6. Stef van Buuren, Karin Groothuis-Oudshoorn. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, 2011, 45(3):1548.

7. James Honaker, Gary King and Matthew Blackwell. Amelia: Amelia II: A Program for Missing Data. 2011. R package version 1.5-4. http://gking.harvard.edu/amelia/

8. *Tuberculosis prevalence surveys: a handbook*. Geneva, World Health Organization, 2010 (WHO/HTM/TB/2010.17).

Chapter 6
Capture-recapture modelling

Authors: Philippe Glaziou, Ross Harris, Rob van Hest, Emily Bloss, Amal Bassili, Fulvia Mecatti, Ibrahim Abubakar

This chapter is intended primarily for epidemiologists and statisticians who are involved in or have an interest in inventory studies. It is much less accessible for general readers.

This chapter describes capture-recapture modelling in the context of TB incidence estimation. Familiarity with the R computing language will help to follow the examples.

Capture-recapture (CR) methods were originally developed in the area of wildlife management [1], but they are now used in a variety of applications [2-5] and the captured units are no longer just animals. For example, in an epidemiological application, the units are humans with a certain disease and the capture occasions are reporting lists or registers. CR is a technique that can be used to indirectly estimate TB incidence, provided that certain conditions are met.

The following ratio will be estimated: $\frac{\text{diagnosed cases}}{\text{diagnosed cases+missed cases}}$

Through record-linkage of different lists of TB cases originating from different providers, it is possible to assess the number of TB cases not recorded in any data source and get a better estimate of the true number of cases than by using one source only [6-10]. CR studies have been performed in the field of TB surveillance [11-26], using simple two-source CR models or applying three-source CR models.

CR studies in the context of tuberculosis surveillance have very important limitations that require careful consideration. CR limitations have already been highlighted in Chapter 1 and are described in fuller detail at the end of this chapter. At least three sources of information are recommended (NTP and at least two non-NTP sources), for reasons outlined in the sections below. In settings where only two sources of information are available (NTP and non-NTP sources), it is not recommended to perform capture-recapture modelling.

6.1 Models using two-data sources

A simple two-source CR model estimates the number of TB cases not captured by any one of two sources of TB patients (e.g. the NTP list and the private doctors list). Record-linkage of the two sources will provide a count of cases diagnosed in list A but not in list B (n_{10}), in list B but not in list A (n_{01}), and in both lists (n_{11}), as displayed in Table 6.1. The number of cases absent from both lists, i.e., undiagnosed cases (n_{00}) is not observed. The number N_A of cases in list A is the sum of n_{10} and n_{11}. Likewise, the number NB of cases in list B is the sum of n_{01} and n_{11}. N is the sum of n_{00}, n_{01}, n_{10}, n_{11}.

Table 6.1: Distribution of cases in a two-list CR model

Symbol n_{00} in the greyed cell refers to cases that are not directly observed. The purpose of CR modelling is to estimate n_{00}.

List A	List B		
	Not identidied	Identified	
Not identidied	n_{00}	n_{01}	
Identified	n_{10}	n_{11}	N_A
		N_B	N

If the event defined by a case appearing in list A is independent from the event defined by a case appearing in list B, n00 may be expressed as

$$n_{00} = \frac{(n_{10}\, n_{01})}{n_{11}} \qquad (1)$$

and, under the same independence assumption, the total number of cases whether identified or not identified in list A and/or B, can be expressed as

$$N = \frac{(N_A\, N_B)}{n_{11}} \qquad (2)$$

A correction for small numbers has been proposed

$$N = \frac{(N_A + 1)\,(N_B + 1)}{n_{11} + 1} - 1 \qquad (3)$$

If the individuals in the two lists N_A and N_B are positively correlated, then those individuals captured in the first list are more easily captured in the second list. That is, we would expect that $N > N_A N_B / n_{11}$. As a result, N obtained from equations (2, 3) underestimates the true size if both samples co-vary positively. Conversely, it overestimates for negatively co-varying samples. A similar argument is also valid for a general number of samples. There is a potential bias with any approach that assumes independence between the samples.

Alternatively, log-linear models can be fitted to estimate n_{00} [27]. The equations below are a reparameterization of the four expected values:

$$\begin{aligned} \log E(n_{\mathrm{ij}}) &= u + u_A I(i=1) + u_B I(j=1) + u_{AB} I(i=j=1) \\ \log E(n_{00}) &= u \end{aligned} \qquad (4)$$

where u_{AB} is an interaction parameter, I is a list/sample membership indicator function and E denotes expectation. When n_{00} is unobserved, the above parameterisation contains only 3 parameters u, u_A and u_B to describe the three observed cells in the 2X2 contingency table. From the first three equations, one can estimate u, u_A and u_B. But u_{AB} cannot be estimated and the only way to estimate n_{00} is to apply a constraint such as $u_{AB} = 0$. This is equivalent to assuming independence as in equations (2) and (3).

$$\log E(n_{\mathrm{ij}}) = u + u_A I(i=1) + u_B I(j=1)$$

In the context of applications of CR to TB surveillance, interactions between data sources are expected because TB service providers tend to collaborate. Being on one type of provider's list, an individual may learn of, or may be approached by another type of provider of TB diagnostic and care services. Also, some cases may not be "catchable" by any source. The probability that an individual is caught in any sample is (i) a property of the individual [28], which has some distribution over the population, and (ii) dependent upon their history of capture in another sample. Co-dependence between A and B will result in a non-zero value for u_{AB} in the two-source log-linear model.

Log-linear models can be fit (with two lists, models require the independence assumption and models do not offer advantages over equations (2) or (3)) in most statistical packages, such as R/S+, SAS or Stata. Estimates of the parameters u are obtained by maximizing a Poisson loglikelihood (see next section).

In the context of TB surveillance, two-source models are not recommended [29] and in order to account for pair-wise dependences at least three separate data sources should be included in CR models.

6.2 Models using three-data sources

The two-source log-linear model can easily be extended to 3 or more sources. Using three lists A, B and C, the number of cases can be expressed in a **saturated** log-linear model as

$$\log E(n_{ijk}) = u + u_A I(i=1) + u_B I(j=1) + u_{AB} I(i=j=1) + u_{AC} I(i=k=1) \\ + u_{BC} I(j=k=1) + u_{ABC} I(i=j=k=1)$$

where u_{AB}, u_{AC} and u_{BC} represent two-way interaction terms and u_{ABC} the three-way interaction. From this model, the number of TB cases not observed in any list is expressed as

$$\log E(n_{000}) = u \qquad (5)$$

Simpler models than the saturated model, including any combinations of pair-wise interaction terms can also be fit. The simplest model will include no interaction term and is only valid in the case of complete independence between lists A, B and C. The analysis of data from a CR study amounts to finding the best fitting model and estimating the number of missed cases from the chosen model. The best model will generally be selected as the one showing the lowest Akaike Information Criterion (AIC), expressed as [30]

$$AIC=G^2 - 2df \qquad (6)$$

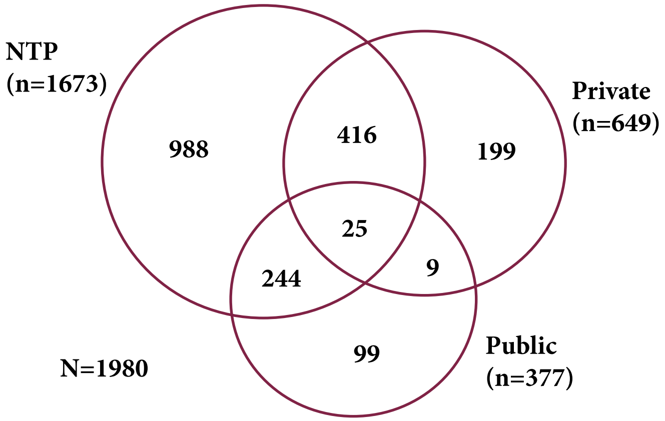
where the term G^2 is the deviance, i.e., a measure of how well the model fits the data and *2df* is a penalty for the addition of parameters and model complexity. The deviance and AIC are computed by most statistical packages.

6.3 An example of capture-recapture modelling: Iraq

6.3.1 Three source model: Iraq

An example using recently analysed data [data used with permission from the study investigators] will show how the estimation can be implemented using R, a freely available open-source computing language (http://www.r-project.org). A CR analysis was done in Iraq using three sources of information: the National TB Programme list of TB cases (NTP), a public list of TB cases obtained independently from the NTP and a private list of TB cases obtained from private physicians. The data are shown in a Venn diagram in Figure 6.1.

Figure 6.1: Distribution of observed number of tuberculosis cases (all forms, N = 1980) in sampled geographical areas in Iraq, 2011



The table below (Table 6.2) shows how the data should be tabulated and loaded into a statistics package. Indicator variables **ntp, public** and **private** are coded 1 to denote observation in the corresponding list and zero to denote absence from the corresponding list.

Table 6.2: An example of data from CR study in Iraq using 3 lists

ntp	public	private	freq
1	1	1	25
1	1	0	244
1	0	1	416
0	1	1	9
1	0	0	988
0	1	0	99
0	0	1	199

A common way to input this small dataset into R is to write the above table in a spreadsheet programme saving the file in a text format with comma delimiters (save as “csv”). The following command loads the dataset in R and stores it in a “dataframe” object, named **dta**. The <- symbol is R’s general assignment symbol.

```
dta <- read.csv(‘iraq_data.csv’)
```

We now inspect the R object named **dta** to ensure that the dataset was correctly imported, by simply typing its name in the R console. R then prints to screen the contents of **dta**, as shown below:

```
> dta
  ntp pub priv freq
1   1   1    1   25
2   1   1    0  244
3   1   0    1  416
4   0   1    1    9
5   1   0    0  988
6   0   1    0   99
7   0   0    1  199
```

We will first fit the simplest three-list log-linear model with complete independence of the three lists, that is, with no interaction term. The fitted model is saved as an object named **fit0**.

```
fit0 <- glm(freq ~ ntp + public + private, family=poisson, data=dta)
summary(fit0)
```

The summary command returns details of the fitted model in a conventional fashion, as shown below. The command summary returns model parameters and their standard errors. The intercept is the parameter *u* to be estimated in equation (5).

```
> summary(fit0)

Call:
glm(formula = freq ~ ntp + pub + priv, family = poisson, data = dta)

Deviance Residuals:
      1        2        3        4        5        6        7
-5.7306   3.9950   2.4065  -4.8688  -1.9899   0.6863   1.2188

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.25364    0.07372   84.83   <2e-16 ***
ntp          0.70469    0.06867   10.26   <2e-16 ***
public      -1.72829    0.05971  -28.95   <2e-16 ***
private     -1.04800    0.05159  -20.31   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

f(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 2189.110  on 6  degrees of freedom
Residual deviance:   84.213  on 3  degrees of freedom
AIC: 138.83

Number of Fisher Scoring iterations: 4
```

We can now compute the "independent" model predicted number of TB cases not observed in any list (n_{000}) with a 95% confidence interval. We will use **confint**, an R function that extracts confidence intervals for linear and generalized linear model parameters. The number of missing cases is given by exp(intercept), that is, by inverting equation (5). The intercept (and other model coefficients) can be extracted from the fitted object **fit0** using the R function **coef**, addressing only the first returned element: **coef(fit0)[1]**.

```
ci <- confint(fit0)
missing0 <- exp(coef(fit0)[1])
```

Confidence intervals for the log number of missing cases are now stored in **ci**. We can extract specific values, for instance, **ci[1,2]** extracts the value at the intersection of row 1 and column 2 in the object **ci**. Exponentiating these values gives confidence intervals for the total cases.

```
> coef(fit0)
(Intercept)         ntp         pub        priv
  6.2536400   0.7046878  -1.7282940  -1.0479955
> missing0
(Intercept)
   519.9018
> ci
                2.5 %     97.5 %
(Intercept)  6.1080161  6.3970745
ntp          0.5712141  0.8405037
public      -1.8466576 -1.6125524
private     -1.1497721 -0.9475034
> exp(ci[1,1])
[1] 449.4462
> exp(ci[1,2])
[1] 600.0869
```

We will now fit the **saturated** model.

```
fits <- glm(freq ~ ntp * public * private, family=poisson, data=dta)
summary(fits)
> summary(fits)

Call:
glm(formula = freq ~ ntp * public * private, family = poisson,
    data = dta)

Deviance Residuals:
[1]  0  0  0  0  0  0  0

Coefficients: (1 not defined because of singularities)
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)          6.2779     0.4168  15.061  < 2e-16 ***
ntp                  0.6178     0.4156   1.486   0.1372
public              -1.6828     0.4045  -4.160 3.19e-05 ***
private             -0.9846     0.4108  -2.397   0.0165 *
ntp:public           0.2843     0.3982   0.714   0.4753
ntp:private          0.1196     0.4066   0.294   0.7686
public:private      -1.4133     0.2180  -6.484 8.95e-11 ***
ntp:public:private       NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2.1891e+03  on 6  degrees of freedom
Residual deviance: 4.0190e-14  on 0  degrees of freedom
AIC: 60.622

Number of Fisher Scoring iterations: 3
```

The three-way interaction term has been automatically dropped. Small numbers of cases common to some of the lists can cause estimation problems, so the saturated model should always be viewed with caution, and in some cases may be extremely unreliable. In this example, cell sizes are as low as 9 and hence standard errors are quite large, although there do not appear to be problems in the model fitting as such. We should however examine models including fewer interaction terms, and the guiding rule is to select the best model based on the lowest AIC that also satisfies notions of biological plausibility. The process of model selection can be automated (for instance, via the R package MASS), but it is recommended to check models with all combinations of interaction terms, with particular attention to interaction terms for which a positive or negative dependence is expected by the study investigators.

It is important to try and understand between source interactions prior to fitting models. For instance, when the private and public systems are largely distinct, a negative co-variation may be expected between these sources. Alternatively, if doctors run public clinics in the morning and private clinics in the afternoon, there could be positive private-public dependence. A positive dependence between the public and NTP sources, with patients diagnosed through public sources being referred to NTP, may be a common occurrence.

Shown below are AIC statistics for the 8 possible models that may be fitted, depending on which of the two-way interaction terms are included:

Model	ntp:public	ntp:private	public:private	df	AIC
M1	y	y	y	7	60.6
M2	y	y		6	115.8
M3	y		y	6	58.7
M4		y	y	6	59.2
M5	y			5	135.3
M6		y		5	139.9
M7			y	5	58.2
M8				4	138.8

Models are divided quite sharply into two groups: those that do not include the public:private interaction, and have an AIC score of over 100, and those that do include the public:provate interaction, all of which have AIC scores around 60 or below. The lowest AIC score is for the model that includes public:private only (M7), but the models also including ntp:public (M3) and ntp:private (M4) have very similar scores, and would generally be considered to have as much support as the lowest scoring model. The saturated model is also quite similar, but does score a bit higher and is generally not desirable due to the greater potential for estimation problems. The choice then comes down to plausibility, and the investigators felt that an ntp:public interaction was most likely, with public patients being more likely to be referred to the NTP. Hence the model M3 was selected as the final model. Note that fortunately in this case, the predicted number of missing cases was similar between the 3 candidate models. If similar scoring models give quite different results, then extreme care should be taken in the model selection.

The following lines show R's output for the model M3. The command is echoed (in bold face) in the console, with a > sign at the beginning of the line.

```
> summary(M3)

Call:
glm(formula = freq ~ ntp + public + private + ntp:public + public:private,
    family = poisson, data = dta)

Deviance Residuals:
       1          2          3          4          5          6          7
 0.14950   -0.04733   0.00000   -0.24023   0.00000    0.07456    0.00000

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     6.15830    0.09188  67.029   <2e-16 ***
ntp             0.73738    0.08619   8.555   <2e-16 ***
public         -1.57069    0.13403 -11.719   <2e-16 ***
private        -0.86500    0.05845 -14.800   <2e-16 ***
ntp:public      0.17520    0.14285   1.226     0.22
public:private -1.44637    0.18906  -7.650    2e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2189.10991  on 6  degrees of freedom
Residual deviance:    0.08786  on 1  degrees of freedom
AIC: 58.71

Number of Fisher Scoring iterations: 3
```

This model shows a positive, but non-significant, dependence between NTP and public sources, and a negative dependence between public and private.

We then extract estimates of missing and total cases:

```
# final estimation
ci <- confint(M3)
found = sum(dta$freq)
missing = exp(coef(M3)[1])
missing.low = exp(ci[1,1])
missing.high = exp(ci[1,2])

N = found + missing
N.lo = found + missing.low
N.hi = found + missing.high
```

The number of estimated cases is 2453 (95% confidence interval: 2374 – 2545), 1980 were found in total and 473 were estimated missing (394 – 565).

```
> N
   2452.625
> N.lo
[1] 2374.091
> N.hi
[1] 2545.045
```

6.4 Alternative models: the example of Iraq

Van Hest showed [31] that for the purposes of estimation of disease incidence, parsimonious three-source log-linear models are preferable but alternative models can be used for comparison purposes, particularly when unrealistically high estimates are generated by saturated log-linear models. Three types of models for closed populations are of particular interest, described by Chao [32], Darroch [33] and Rivest [34]. The latter is based on Darroch's model using a mixture Poisson distribution. Those models (and others) are implemented in the R package **Rcapture** [35], for which well written documentation is available. The following code demonstrates how to implement with **Rcapture** the lowest AIC model shown above and then, how to fit alternative models. A model matrix needs to be specified in order to fit models with specific interaction terms.

```
library(Rcapture) # loads the Rcapture package
mat <- histpos.t(3) # generates model matrix for “independent” models
mx <- cbind(mat, mat[,1] * mat[, 2], mat[, 2] * mat[, 3]) # creates a model matrix
cc <- closedp.mX(dta, dfreq=TRUE, mX = mx) # fits models for closed population

cc # prints the returned object to screen
> cc

Number of captured units: 1980

Abundance estimation and model fit:
                  abundance  stderr  deviance  df    AIC
Customized model     2452.6    48.6     0.088   1  58.71

# alternative models
cc2 <- closedp(dta2, dfreq=TRUE)
> cc2
```

```
Number of captured units: 1980

Abundance estimations and model fits:
              abundance  stderr  deviance  df       AIC
M0               3044.5    71.1  1630.550   5  1681.172
Mt               2499.9    44.6    84.213   3   138.835
Mh Chao          3044.5    71.1  1630.550   5  1681.172
Mh Poisson2      2362.5    58.6  1559.273   4  1611.895
Mh Darroch       2157.6    45.6  1559.273   4  1611.895
Mh Gamma3.5      2061.6    30.2  1559.273   4  1611.895
Mth Chao         2499.9    44.6    84.213   3   138.835
Mth Poisson2     2209.8    37.6    38.725   2    95.347
Mth Darroch      2101.8    32.3    38.725   2    95.347
Mth Gamma3.5     2044.1    24.1    38.725   2    95.347
Mb               2002.8     5.7   490.410   4   543.032
Mbh              1544.8    45.5   123.284   3   177.907

Note: 1 eta parameter has been set to zero in the Mh Chao model
Note: 1 eta parameter has been set to zero in the Mth Chao model
```

The selected log-linear model stored in object **cc** provides satisfactory estimates and we do not retain alternatives models. The interested reader is invited to study R package's documentation while exploring the many possibilities offered by the package.

6.4.1 Problems with two source models

To demonstrate problems with two-list CR problems applied to TB surveillance, the data from Iraq will be collapsed into a two-list CR problem by combining the **private** and **public** lists into a single list **non.ntp**, as shown in Table 6.3.

Table 6.3: An example of data from CR study in Iraq using 2 lists (public non-NTP and private lists collapsed into one list)

ntp	non.ntp	freq
0	1	307
1	0	988
1	1	685

```
dta$non.ntp <- as.numeric(dta$public | dta$private) # creates new variable non.ntp
dta2 <- as.data.frame(xtabs(freq ~ ntp + non.ntp, data=dta))[-1, ] # drops the first row
(dta2) # prints the collapsed dataset to screen
> (dta2)
  ntp non.ntp Freq
2   1       0  988
3   0       1  307
4   1       1  685
n00 <- with(dta2, freq[1] * freq[2] / freq[3]) # uncorrected missing
n00c <- with(dta2, (freq[1] + 1) * (freq[2] + 1) / (freq[3] + 1) - 1) # corrected

fitm2 <- glm(freq ~ ntp + non.ntp, family=poisson, data=dta2) # two-list loglinear
ci <- confint(fitm2) # confidence intervals of model estimates
found <- sum(dta2$freq) # total observed
missing.m2 <- exp(coef(fitm2)[1])
missing.m2.low = exp(ci[1,1])
missing.m2.high = exp(ci[1,2])
```

Our main results are summarized in Table 6.4. The collapsed dataset used above actually gives a similar result to the 3-source log-linear model, as the public:private interaction has been "collapsed out" and the ntp:public interaction is not strong. However, examining other 2-source models, we found that if public and private sources only are used, the estimated number of cases is 7196- a huge over-estimate, due to the strong negative interactions between these sources not being accounted for. As it is highly likely that dependencies will exist between data sources, we do not recommend the use of two-list CR models in applications to TB surveillance.

Table 6.4: Model results

Two-list model results are likely to be unreliable due to dependence between the two lists not being accounted for, although results are similar in this example.

	Model #	**Unobserved**	**Total size** N (95% confidence interval)
Two-list models		n_{00}	
Uncorrected (equation 2)	1	443	2423
Corrected (equation 3)	2	443	2423
Log-linear	3	443	2423 (2361 – 2493)
Three-list models		n_{000}	
No interaction term	4	519	2500 (2429 – 2580)
Saturated	5	533	2512 (2223 – 3246)
ntp X public, public X private	6	473	2453 (2374 – 2545)
Model 6, cluster adjusted*	7	473	2453 (2184 – 3076)
Model 6, fully design adjusted*	8	460	2439 (2213 – 2888)

*see next section

6.4.2 Adjusting for sampling

In Iraq 18 governorates were stratified by quartiles of notified TB rate, and two governorates randomly sampled from each strata. The cluster-level data is shown in Table 6.5. It is important to consider the effect of sampling on analyses: without accounting for the clustered sampling, one essentially assumes that simple random sampling was conducted *at the individual level* i.e., all individuals had an equal probability of being sampled, and the data are representative of the entire population. This is clearly not the case: large sections of the population were not sampled from (the non-sampled governorates) and there is no information on these parts of the population in the data. Intuitively, one may see that there should therefore be more uncertainty in our estimates than those obtained above, and it can easily be shown that naïve analysis does indeed under-estimate standard errors if there are differences between clusters [36].

It is therefore recommended to adjust for the clustered sampling design. The adjustment will inflate standard errors from model coefficients, correctly accounting for the uncertainty in estimates. If there is inter-cluster variation, i.e. the populations across clusters are quite different, then this increase might be quite large. Conversely, the sampling for the Iraq data was stratified, and accounting for this may then reduce standard errors- at least in comparison with estimates obtained under basic cluster random sampling. The aim of stratification is to obtain similar clusters within each strata (low within-strata variation) and this will help to reduce uncertainty in the final estimates: the more similar clusters are within strata, the greater the improvement.

This reduction in standard errors will be correct (compared with the incorrect analysis, which just ignores the design), and the ratio of variance under stratified design vs. a simple random sample of clusters is the *design effect*.

These analyses may be carried out in R and are described in Lumley [37] and implemented in the R package **survey** [38].

Table 6.5: Cluster-level data, Iraq

Data are displayed by strata with percentage of population sampled and number of governorates in strata; then by governorate, with population size.

NTP	Public	Private	Freq
STRATA 1; pop. sampled=37.2%, #govs=4			
Basrah, population: 2609000			
1	1	1	8
1	1	0	50
1	0	1	79
0	1	1	0
1	0	0	70
0	1	0	11
0	0	1	7
0	0	0	
Duhok, population: 985000			
1	1	1	4
1	1	0	13
1	0	1	33
0	1	1	2
1	0	0	40
0	1	0	10
0	0	1	19
0	0	0	
STRATA 2; pop. sampled=47.6%, #govs=4			
Misan, population: 1030000			
1	1	1	0
1	1	0	22
1	0	1	7
0	1	1	1
1	0	0	27
0	1	0	10
0	0	1	10
0	0	0	

NTP	Public	Private	Freq
STRATA 3; pop. sampled=84.1%, #govs=4			
Baghdad, population: 7341000			
1	1	1	5
1	1	0	61
1	0	1	93
0	1	1	0
1	0	0	534
0	1	0	55
0	0	1	118
0	0	0	
Sulaymania, population: 1703000			
1	1	1	3
1	1	0	41
1	0	1	82
0	1	1	6
1	0	0	51
0	1	0	9
0	0	1	17
0	0	0	
STRATA 4; pop. sampled=31.7%, #govs=5			
Diwanyia, population: 1125000			
1	1	1	3
1	1	0	12
1	0	1	85
0	1	1	0
1	0	0	70
0	1	0	2
0	0	1	11
0	0	0	

Najaf, population: 1216000			
1	1	1	0
1	1	0	2
1	0	1	5
0	1	1	0
1	0	0	65
0	1	0	1
0	0	1	9
0	0	0	

Wasit, population: 1156000			
1	1	1	2
1	1	0	43
1	0	1	32
0	1	1	0
1	0	0	131
0	1	0	1
0	0	1	8
0	0	0	

Firstly, we load the data into R and repeat the naïve analysis, without accounting for sampling design, to check that results are the same.

```
clu <- read.csv(“H:/WHO CRC/Iraq 2011 analyses/CRC_iraq_gov_R.csv”)
unadjs<- glm(n ~ ntp + priv + pub + ntp:pub + pub:priv, offset=log(pop), fami-
ly=poisson, data=clu)
summary(unadjs)
Call:
glm(formula = n ~ ntp + priv + pub + ntp:pub + pub:priv, family = poisson,
    data = clu, offset = log(pop))
[OUTPUT OMITTED]
Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -10.50007    0.09188 -114.286   <2e-16 ***
ntp           0.73738    0.08619    8.555   <2e-16 ***
priv         -0.86500    0.05845  -14.800   <2e-16 ***
pub          -1.57069    0.13403  -11.719   <2e-16 ***
ntp:pub       0.17520    0.14285    1.226     0.22
priv:pub     -1.44637    0.18905   -7.651    2e-14 ***
---
 [OUTPUT OMITTED]
> ci <- confint(unadjs)
> missing = exp(coef(unadjs)[1])*17164779
> missing.low = exp(ci[1,1])*17164779
> missing.high = exp(ci[1,2])*17164779
> missing
(Intercept)
    472.625
> missing.low
[1] 394.0914
> missing.high
[1] 565.0447
```

The estimated number of missing cases is identical, as are the model coefficients, apart from the intercept. The intercept changes as the log of the governorate population size has been incorporated as an offset to obtain per-person incidence rates from the model, and this is later rescaled by the total population size of the sampled governorates (~17 million). The reason for this is explained later.

Next, the cluster sampling is accounted for. We use the variable gc to identify sampling units (governorates), and incorporate a finite population correction (FPC), which is the proportion of the entire population sampled.

```
> library(survey)
> dclu1 <- svydesign(id=~gc, fpc=~totfracsamp, data=clu)
> dfit1 <-svyglm(n ~ ntp + priv + pub + ntp:pub + pub:priv, offset=log(pop), fam-
ily=poisson, design=dclu1)
> summary(dfit1)
Call:
svyglm(n ~ ntp + priv + pub + ntp:pub + pub:priv, offset = log(pop),
    family = poisson, design = dclu1)

Survey design:
svydesign(id = ~gc, fpc = ~totfracsamp, data = clu)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.5001     0.4292 -24.465  0.00167 **
ntp           0.7374     0.3303   2.232  0.15526
priv         -0.8650     0.2974  -2.909  0.10066
pub          -1.5707     0.3644  -4.310  0.04985 *
ntp:pub       0.1752     0.1652   1.060  0.40012
priv:pub     -1.4464     0.1440 -10.047  0.00976 **
---
[OUTPUT OMITTED]

> ci <- confint(dfit1 )
> found = sum(dta$n)
> missing = exp(coef(dfit1)[1])*17164779
> missing.low = exp(ci[1,1])*17164779
> missing.high = exp(ci[1,2])*17164779
> missing
(Intercept)
    472.625
> missing.low
[1] 203.796
> missing.high
[1] 1096.068
```

Model estimates are identical, but standard errors have increased dramatically and there is now substantial uncertainty in the estimated number of missing cases. This indicates that there is significant between-cluster variation, and results therefore reflect this. This may be improved upon by capitalising on the similarity of clusters within strata. Accounting for the stratified design, we then have:

```
> dclu2 <- svydesign(id = ~gc, strata=~stratum, weights=~ngov, fpc=~fracsamp,
data=clu)
> dfit2 <-svyglm(n ~ ntp + priv + pub + ntp:pub + pub:priv, offset=log(pop), fam-
ily=poisson, design=dclu2)
>
> ci <- confint(dfit2 )
> missing = exp(coef(dfit2)[1])*17164779
> missing.low = exp(ci[1,1])*17164779
> missing.high = exp(ci[1,2])*17164779
> missing
(Intercept)
     459.84
> missing.low
[1] 232.8884
> missing.high
[1] 907.9579
```

The point estimate has now changed slightly as weights (the number of governorates per-strata) and stratum-specific FPCs are included. The standard error has also been reduced compared with the previous analysis. The reduction is not great, indicating that there is still a fair amount of within-strata variation, but it is a marked improvement. Results from the analyses above are displayed in Table 6.4.

Box 6.1: A note on offsets and model intercept

When analysing the governorate-level data, it is important to consider the scaling used for the log-linear model. Recall the basic (saturated) log-linear model:

$$\begin{aligned}\log E(n_{ijk}) = u + u_A I(i=1) + u_B I(j=1) + u_C I(k=1) + u_{AB} I(i=j=1)\\ + u_{AC} I(i=k=1) + u_{BC} I(j=k=1) + u_{ABC} I(i=j=k=1)\end{aligned}$$

Here the intercept, u, represents the (log) number of unobserved cases. In the governorate level case, this will vary naturally across governorates based purely on the population size: larger governorates are likely to have higher numbers of TB, irrespective of the actual TB rate. Ignoring this will result in artificially high variation between governorates, and therefore standard errors; so the log population size must be included as an offset. The intercept then represents the per-person rate of unobserved TB in the population, which should be comparable between governorates. Results are then scaled by the total population size to obtain the required estimate of unobserved cases.

6.5 Limitations

The limitations of CR studies in estimating TB incidence depend, as in any CR study, on the violation of the underlying assumptions.

1. Violation of the perfect record-linkage assumption (i.e. no misclassification of records) depends on the availability of a unique identifier in all lists, or sufficient proxy identifiers (see Chapter 4).

2. CR studies based on sampling of geographical areas will fail to link records for the same cases in a neighbouring or other non-sampled area, resulting in misclassification of records. The problem is partly alleviated by sampling large geographical areas capturing a large proportion of a country's providers. CR studies using complete lists covering the whole country are preferable but may not be practical for logistical and cost reasons.

3. Almost zero probability of being observed by any source. We can only estimate the size of a subpopulation that contains only observable individuals. For instance, if in a studied country, the vast majority of TB cases are diagnosed passively upon presentation with a certain set of symptoms (as opposed to through active screening), then we will not be able to estimate the number of asymptomatic TB cases or the number of cases with atypical

symptoms for which there is no national recommendation to systematically prescribe TB investigations. Also, cases with no geographical access to care will not be accounted for.

4. The closed population assumption (i.e. no immigration or emigration, no TB deaths and no new TB cases in the time period studied) is reasonable in countries with a well-organised TB programme, as the opportunities for notification, laboratory confirmation of the diagnosis, or hospitalisation, are determined within a relatively short period of time. In other settings, the assumptions may not be met.

5. The assumption of independence between the different TB registers (cases lists) is likely violated in most settings. The probability of being observed in one list is expected to be affected by being observed in another list, since TB services are often organised around close collaboration between clinicians, microbiologists and other public health professionals, such as TB physicians and nurses. Therefore, two-list CR models are not recommended in the context of TB surveillance and at least three lists should be used. In order to avoid model fitting problems and very large standard errors, it is recommended that a minimum of 20–30 cases be common to any combination of two lists. Some countries may not have the possibility of generating three or more lists of cases with sufficient overlap between pairwise combinations, e.g. countries with no private providers of TB diagnostic and care services and complete coverage of public services by the National TB programme. In such countries, CR studies of TB incidence may not be appropriate.

6. 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, sex, socio-economic status, location of disease and severity of the disease normally cause different probabilities of being observed in a list.

7. Cases should be uniformly defined across all lists. Laboratory-confirmed TB cases have a positive culture for *Mycobacterium tuberculosis*. For other lists there is a risk of false-positive records, for example due to infection with *Mycobacteria Other Than Tuberculosis*, or a final diagnosis other than tuberculosis, which are not necessarily removed or corrected in hospital discharge codes. It is therefore important to critically assess diagnostic confirmation in all lists.

References

1. Seber GAF. *The Estimation of Animal Abundance and Related Parameters*. Macmillan, New York, 2nd edition.

2. Abeni D, Brancato G, Perucci C. Capture-Recapture to Estimate the Size of the Population with Human Immunodeficiency Virus type 1 Infection. *Epidemiology,* 1994, 5:410–414.

3. Darroch JN, Fienberg SE, Glonek GFV, Junker BW (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. *J Am Stat Assoc,* 1995, 88:1137–1148.

4. Ebrahimi NB. On the Statistical Analysis of the Number of Errors Remaining in Software Design Document After Inspection. *IEEE Trans Softw Eng,* 1997, 23:529–532.

5. Briand LC, Emam KE, Freimut BG, Laitenberger O. A Comprehensive Evaluation of Capture-Recapture Models for Estimating Software Defect Content. *IEEE Trans Softw Eng,* 2000, 26(6):518–540.

6. Mukerjee AK: Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data source. *Commun Dis Public Health,* 1999, 2:143-4.

7. 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.

8. 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.

9. 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.

10. Van Buynder P. Enhanced surveillance of tuberculosis in England and Wales: circling the wagons? *Commun Dis Public Health,* 1998, 1:219-20.

11. 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.

12. 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.

13. 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.

14. 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.

15. 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.

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 co-infection in the province of Seville, Spain, 1998. *Eur J Epidemiol,* 2001, 17: 737-42.

17. 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.

18. 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. 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. *Epidemiol Infect*, 2008, 136(12):1606-16.

25. Cojocaru C, van Hest NA, Mihaescu T, Davies PD. Completeness of notification of adult tuberculosis in Iasi County, Romania: a capture-recapture analysis. *Int J Tuberc Lung Dis*, 2009, 13(9):1094-9.

26. Bassili A, Grant AD, El-Mohgazy E, Galal A, Glaziou P, Seita A, Abubakar I, Bierrenbach AL, Crofts JP, van Hest NA. Estimating tuberculosis case detection rate in resource-limited countries: a capture-recapture study in Egypt. *Int J Tuberc Lung Dis*, 2010, 14(6):727-32.

27. Fienberg SE. The multiple recapture census for closed populations and incomplete 2^k contingency tables.

Biometrika, 1972, 59(3):591-603.

28. Comack RM. A test for equal catchability. *Biometrics,* 1966, 22:330-342.

29. International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record estimation. *Am J Epidemiol,* 1995, 142:1047-58.

30. Sakamoto, Y., Ishiguro, M., and Kitagawa G. (1986). Akaike Information Criterion Statistics. D. Reidel Publishing Company.

31. Van Hest NA, Grant AD, Smit F, Story A, Richardus JH. Estimating infectious diseases incidence: validity of capture-recapture analysis and truncated models for incomplete count data. *Epidemiol Infect*, 2008, 136(1):14-22.

32. Chao A. Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometriks,* 1987, 43(4):783-791.

33. Darroch JN, Fienberg SE, Glonek GFV, Junker BW. A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. *J Am Stat Assoc,* 1993, 88:1137-1148.

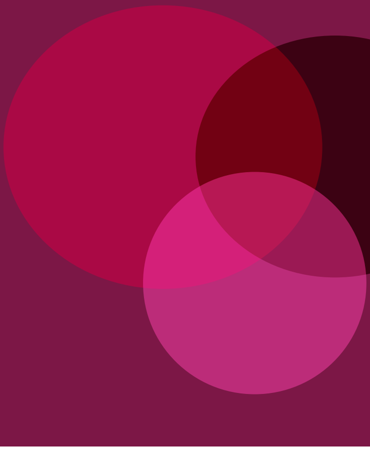
34. Rivest LP, Baillargeon S. Applications and extensions of Chao's moment estimator for the size of a closed population. *Biometrics,* 2007, 63(4):999-1006.

35. Baillargeon S, Rivest LP. Rcapture: Loglinear Models for Capture-Recapture Experiments. R package version 1.2-0. 2009. http://CRAN.R-project.org/package=Rcapture

36. Kerry SM, Bland MJ. The intracluster correlation coefficient in cluster randomisation. *BMJ,* 1998, 316:1455.1

37. Lumley T. Complex Surveys. A Guide to Analysis Using R. Wiley Series in Survey Methodology. Wiley 2010. ISBN 978-0-470-28430-8.

38. T. Lumley (2011) *Survey: analysis of complex survey samples*. R package version 3.24-1.
http://cran.r-project.org/web/packages/survey/index.html



Estimation of TB incidence is a major challenge in many countries due to under-reporting and under-diagnosis of TB cases. This guide describes and explains how to design, implement and analyse an inventory study to measure TB under-reporting. It also explains how to apply capture-recapture methods to estimate TB incidence.

The guide is intended for use by NTP managers, researchers including epidemiologists and statisticians, and agencies that provide financial and technical support for inventory studies. Its aim is to catalyse many more inventory studies worldwide, leading to better measurement of the burden of TB disease and, in turn, better TB prevention, diagnosis and treatment services.

World Health
Organization

For further information about tuberculosis contact:
Information Resource Centre HTM/STB
World Health Organization
20 Avenua Appia
1211 Geneva 27
Switzerland
Email: tbdocs@who.int
Web site: www.who.int/tb

ISBN 978 92 4 150494 2