Accuracy and completeness of notification of tuberculosis in two high incident communities in Cape Town, South Africa

by
Rory Dunbar

Thesis presented in partial fulfilment of the requirements for the degree Masters of Science in Medical Sciences (Epidemiology) at the University of Stellenbosch

Supervisor: Dr, Jo Barnes
Co-supervisor: Prof, Nulda Beyers
Faculty of Health Sciences
Division of Community Health

December 2011

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the authorship owner thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: December 2011

ABSTRACT

Introduction: Tuberculosis (TB) treatment registers and laboratory records are essential recording and reporting tools in TB control programmes. Reliable data are essential for any TB control programme but under-registration of TB cases has been well documented internationally, due to under-reporting of patients on treatment or failure to initiate treatment. The accuracy and completeness of routinely collected data are seldom monitored.

Aim: This study used record linking to assess the accuracy and completeness of TB treatment register data and the feasibility of estimating the completeness of bacteriological confirmed pulmonary TB registration in two high incident communities in South Africa with capture-recapture methods.

Methods: All cases of bacteriologically confirmed TB defined as 2 smear-positive results and/or at least one culture-positive result were included. Record linking was performed between three data sources: (1) TB treatment registers; and (2) all smear and culture results from (a) the nearest central laboratory, and (b) the referral hospital laboratory. To estimate the completeness of TB treatment recording three-source log-linear capture-recapture models were used, with internal validity analysis.

Results: The TB treatment registers had 435 TB cases recorded of which 204 (47%) were bacteriologically confirmed cases. An additional 39 cases that were recorded as non-bacteriological cases in the TB treatment register, were reclassified as bacteriologically confirmed. In addition, there were 63 bacteriologically confirmed cases identified from the laboratory databases which were not recorded in the TB treatment register. The final total number of bacteriologically confirmed TB cases across all 3 databases was 306, an increase of 50% over what had initially been recorded in the TB treatment register. The log-linear capture-recapture model estimated the number of bacteriologically confirmed TB cases not found in any of the data sources at 20, resulting in a total number of bacteriologically confirmed TB cases of 326 (95% CI 314-355). The completeness of registration of bacteriologically confirmed pulmonary TB cases was 79% after record linking and 75% after the capture-recapture estimate.

Conclusions: The results presented in this thesis highlighted the concern regarding the accuracy and completeness of routinely collected TB recording and reporting data. A high percentage of

bacteriologically confirmed cases from both laboratories were not recorded in the TB treatment registers. Capture-recapture can be useful, but not essential, for evaluation of TB control programmes, also in resource-limited settings, but methodology and results should be carefully assessed. The present study estimated the extent of the problem of underreporting of TB in South Africa and identified challenges in the process. Interventions to reduce underreporting of TB are urgently needed.

OPSOMMING

Inleiding: Registers van tuberkulose (TB) behandeling en laboratoriumrekords is noodsaaklike instrumente in die dokumentering van en verslagdoening oor TB beheerprogramme. Betroubare data is onontbeerlik vir enige TB beheerprogram maar onderregistrasie van TB gevalle is internasionaal goed gedokumenteer. Die akkuraatheid en volledigheid van roetine data word selde gemoniteer.

Doel: Hierdie studie het rekordkoppeling gebruik om die akkuraatheid en volledigheid van data in TB behandelingsregisters te ondersoek. Voorts is die uitvoerbaarheid van die vangs-hervangsmetodes vir die beoordeling van die volledigheid van bakteriologies bevestigde pulmonale TB registrasie in twee hoë-insidensie gemeenskappe ondersoek.

Metodes: Alle gevalle van bakteriologies bevestigde TB, gedefinieer as 2 smeer-positiewe resultate en/of ten minste een kultuur-positiewe resultaat, is in die studie ingesluit. Rekordkoppeling is onderneem tussen drie databronne: (1) TB behandelingsregisters; en (2) alle smeer- en kultuurpositiewe resultate van (a) die naaste sentrale laboratorium, en (b) die verwysende hospitaallaboratorium. Om die volledigheid van TB behandelingsrekords te ondersoek is drie-bron log-lineêre vangs-hervangs modelle gebruik met interne geldigheidsontleding.

Resultate: Die TB registers het 435 aangetekende TB gevalle bevat waarvan 204 (47%) bakteriologies bevestigde gevalle was. 'n Bykomende 39 gevalle wat as nie-bakteriologies bevestigde gevalle aangeteken was in die TB register is hergeklassifiseer as bakteriologies bevestig. Daar is ook 63 bakteriologies bevestigde gevalle geïdentifiseer vanuit die laboratorium databasisse wat nie in die TB register aangeteken was nie. Die finale totale aantal bakteriologies bevestigede TB gevalle oor al drie databasisse heen was 306, 'n toename van 50% in vergelyking met wat aanvanklik in die TB register aangeteken was. Die log-lineêre vangs-hervangs model het die aantal bakteriologies bevestigde gevalle wat nie in enige van die databronne gevind kon word nie as 20 gevalle geskat, wat gelei het tot 'n totaal van 326 (95% VI 314-355) bakteriologies bevestigde gevalle. Die volledigheid van registrasie van bakteriologies bevestigde TB gevalle was 79% na rekordkoppeling en 75% na die vangs-hervangs skatting.

Gevolgtrekkings: Die resultate wat in hierdie tesis voorgelê is beklemtoon die besorgdheid oor die akkuraatheid en volledigheid van die aanmelding en optekening van roetine TB data. 'n Hoë

persentasie van bakteriologies bevestigde gevalle van beide laboratoriums is nie in die TB register opgeteken nie. Vangs-hervangs kan nuttig wees, maar nie noodsaaklik nie, in die evaluasie van TB beheerprogramme, ook in hulpbron-arm omgewings, maar die metodologie moet omsigtig beoordeel word. Die huidige studie het die omvang van die probleem van onderrapportering van TB in Suid-Afrika beraam en uitdagings in die proses geïdentifiseer. Intervensies om onderrapportering te verminder word dringend benodig.

ACKNOWLEDGEMENT

A special word of thank you has to go to Dr Jo Barnes and Prof Nulda Beyers, for all their assistance and patience. Their insight has been invaluable the past years and they always kept me on the correct track. Without their guidance this thesis would never have been the final product that it is. Prof Donald Enarson also requires a special word of thank you, with his incredible insight and experience in TB and research in general. Prof Enarson has been, and still is, one of those individuals that are an endless source of knowledge.

For all those that assisted me with knowledge, analysis and data especially Rob van Hest with his insight in capture-recapture, Suzanne Verver with her insight into Epidemiology and Tuberculosis and assisting me with so many things from travel arrangements and funding. Florian Marx with his constant drive for finding answers and knowledge was a source of motivation throughout this study especially on those difficult days when it seemed that there would be no end to this thesis. I also wish to thank City of Cape Town Health Directorate, especially Judy Caldwell, and National Health Laboratory Systems (NHLS). The team at NHLS, Sue Candy and her developers, were always willing to assist with data requests.

Stellenbosch University and the Department Paediatrics and Child Health were always willing to assist financially with courses, travel costs and the rental of special equipment when travelling nationally and internationally. For this I am truly thankful. I also wish to thank the South African Centre for Epidemiological Modelling and Analysis (SACEMA) and KNCV Tuberculosis Foundation for financial support throughout different stages of my studies.

An extremely big thank you to all my colleagues at Desmond Tutu TB Centre for all their support. I especially wish to thank the research nurses, Susan van Zyl and Danite Bester, for all their hard work in the clinics. They always accepted another list of requests with a smile, even though working out in the sites is never easy. Not to forget Elizabeth Du Toit and Mareli Claassens for reading through this thesis and helping with spelling and grammar, and just for listening when I needed some advice.

Lastly thank you to my family for all their support though all these years. I wish to thank my wife, Kim, in particular for keeping me motivated through all the long nights and weekends. Especially thank you for allowing me to neglect you so much in the past couple of years. Then there is also

my sister in law, Cathy Ward, who also had the misfortune of being asked to read through this thesis and correct spelling and grammar.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AFB	Acid-Fast Bacillus
AIC	Akaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
ARTI	Annual Risk of Tuberculosis Infection
BCG	Bacillus Calmette-Guerin
BIC	Bayesian Information Criterion
BMU	Basic Management Unit
CI	Confidence Interval
df	degrees of freedom
DOTS	Directly Observed Treatment Short-course
DST	Drug Susceptibility Testing
ETR	Electronic TB Registers
GRLS	Generalised Record Linking System
HIV	Human Immunodeficiency Virus
IUATLD	International Union Against Tuberculosis and Lung Disease
LA	Local Authority
LPA	Line Probe Assay
MDGs	Millennium Development Goals
NDoH	National Department of Health
NHLS	National Health Laboratory Service

NTP	National TB Programme
NYSIIS	New York State Identification and Intelligence System
PGWC	Provincial Government of the Western Cape
PHC	Primary Health Care
PTB	Pulmonary TB
TB	Tuberculosis
TST	Tuberculin Skin Tests
WC	Western Cape
WHA	World Health Assembly
WHO	World Health Organisation

FOREWORD

National TB Programmes try to improve TB case detection rates by carefully monitoring treatment outcomes. However, little is known about the proportion of TB cases diagnosed who never start treatment or how accurate and complete TB registration in the NTP is. In order to understand and determine how accurate and complete TB case registration is, the epidemiology of TB should be understood. This thesis proposes to assess the accuracy and completeness of TB treatment register data and to estimate a more accurate TB case registration through capture-recapture.

This thesis follows the recently approved format of presentation of literature reviews, then the original research data in the form of published papers followed by an overall synthesis and recommendations. Chapters 1 and 2 contain reviews of two separate aspects of the topic, namely a review of the epidemiological and managerial aspects of TB while Chapter 2 reviews the capture-recapture approach. Since the published papers contain their own reference lists, the rest of the sections in this thesis were also provided with separate reference lists to avoid confusion over numbering.

The first chapter focuses on a review of the known epidemiological background to TB. The epidemiology of TB was described using the epidemiological triangle as a framework. The factors influencing TB exposure and TB infections and the progression to TB disease are briefly described. An introduction to the descriptive epidemiology of TB follows, focusing on the measures of TB infection, TB prevalence and TB incidence. An overview of the state of TB globally and in South Africa is given, followed by an explanation of the structure and functioning of TB services within South Africa. TB recording and reporting is discussed using the "onion model" as a framework.

Chapter 2 is a literature review of the data processing methods used in this thesis in order to address the problem of accuracy and completeness of TB recording and reporting. An overview of record linking is given followed by a description of how record linking was used in this thesis. The application and limitations of capture-recapture methods are discussed including the requirements needing to be fulfilled before using these methods. Chapter 2 concludes with a brief comparison of the current study and a similar study conducted in Egypt.

Chapter 3 provides a brief overview of the aim and objectives of this thesis. The aim of this study was to assess the accuracy and completeness of TB case recording and reporting. Two objectives

were therefore set to address the aim. The first objective was to determine the accuracy and completeness of the data recorded in the TB treatment registers and laboratory records and the accuracy and completeness of case registration. This objective was addressed through record linking of the TB treatment register and laboratory data sources. The second objective was to assess the contribution of capture-recapture analysis in estimating the completeness of recording and case ascertainment of bacteriologically confirmed TB. Chapter 3 then concludes with a brief overview of the study setting and previous research studies conducted in the setting.

Chapter 4 is represented by a published article in the traditional format. This chapter focuses on the first objective of this study, namely record linking. A brief introduction on record linking and the methods used are given. Detailed results of the record linking process and the findings from record linking were covered in chapter 4 as well as a detailed discussion and conclusion of the results and finding of record linking.

Chapter 5 is also represented by a published article in the traditional format. This chapter focuses on the second objective of this thesis, namely capture-recapture. Detailed results of the capture-recapture analysis were then given. Chapter 5 concludes with a discussion of the capture-recapture results, the interpretation of these results and the limitation of these results considering the available data and the setting.

Chapter 6 is a general overview of the conclusions reached in this study. The accuracy and completeness of TB recording and reporting are discussed in relation to the findings of the record linking and capture-recapture. The underreporting of TB cases is then discussed as well as the implications for the epidemiology of TB and the National TB Programme, the individual and the community. The findings from the record linking and capture-recapture analysis are then discussed in details. The reasons for the breakdown in the National TB programme as a system are discussed. This is an integral part of putting the finding of this thesis into perspective as the reasons for any health system breaking down are often complicated. The breakdown points are also the best places to institute remedial action. Chapter 6 then concludes with an overview of what the research in this study accomplished and overall recommendations.

The two annexures provided an illustration of the procedures followed by the record linking program, Link Plus, as it was used in this study. Annexure 1 describes the overall functionality of Link Plus, while annexure 2 describes the operation of Link Plus by using some fictitional names.

This fictitional sample data are used in to give a step-by-step explanation of the record linking process.

CHAPTER 1: INTRODUCTION

1.1 EPIDEMIOLOGY OF TUBERCULOSIS

The current epidemic of tuberculosis (TB) in South Africa and specifically in Cape Town is placing a huge burden of disease on the health services in areas of high prevalence.[1,2] Successful control depends on efficient, well-managed health services with the necessary capacity and the rapid detection and treatment of infectious TB patients.[3,4] The National TB Programme (NTP) tries to improve case detection rates by carefully monitoring treatment outcomes. However, little is known about the proportion of TB cases diagnosed that never start treatment and how accurate and complete TB registration in the NTP is. In order to understand and determine how accurate and complete TB case registration is, the epidemiology of TB should be understood.

Epidemiology is considered a basic science of public health and as a public health discipline should be used to promote and protect the public's health. Epidemiology is the study of patterns of health events, health characteristics, or health determinants in a society. It is the cornerstone method of public health research, and helps inform policy decisions and evidence-based medicine by identifying risk factors for disease and targets for preventive medicine. Epidemiology could therefore be used to address public health issues by identifying communities with high burden of disease and help decision makers to develop and implement targeted interventions.[5,6]

The epidemiologic triad or triangle is a model that has been developed to explain disease causation.[5,7] There are a number of such models of which one is represented in Figure 1-1. This particular model is commonly used to explain the causation of infectious diseases. The model consists of three components namely:

- Agent factors: the infectious microorganism which is generally necessary but not sufficient to cause disease. Examples are bacteria, viruses or parasites.
- Factors affecting the host are those that influence an individual's exposure, susceptibility or response to a causative factor. These factors commonly include age, gender, race (where appropriate), socioeconomic circumstances and behaviour (for instance smoking or lifestyle).
- Environmental factors are extrinsic factors which affect the agent and/or influence the opportunity for exposure. These factors include climate, water and/or soil composition, pollution levels, and type of dwelling.

Models such as these allow the identification of points for intervention. For instance, one can eliminate the agent, improve the host's resistance to the agent, or alter the environment so that risk of exposure is reduced.

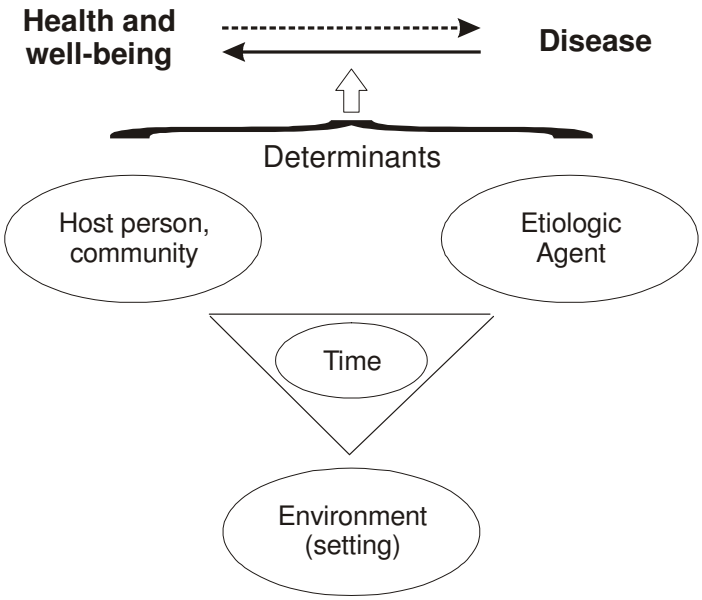

Figure 1-1: Example of an epidemiological triangle.[7]

A similar model has been developed to explain the pathogenesis of TB (Figure 1-2) which consists of four steps namely exposure, infection, disease and death. In this model the exposure to a potential infectious case is necessary in order to become infected.[8-11] This model represents the lifetime risk of exposure to TB and therefore ends in death. After death a person cannot be exposed to TB and/or transmit TB and therefore is the only outcome which eliminates a person from the model. There are other outcomes in TB programmes used for the management of TB e.g. cured, transferred out, defaulted etc. but none of these epidemiologically totally omits an individual from transmitting TB again.

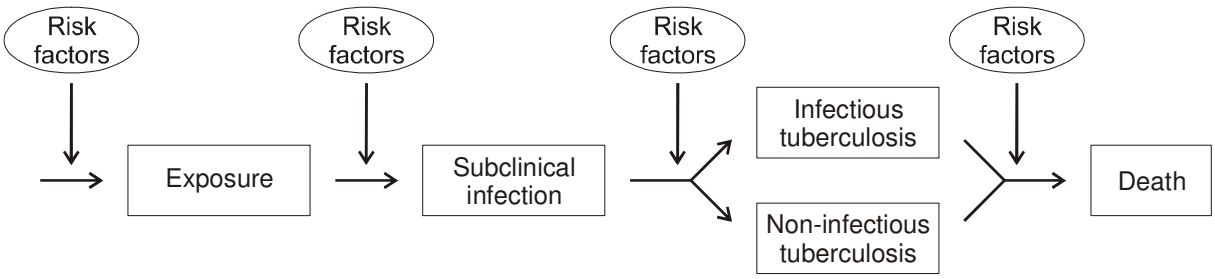

Figure 1-2: A model for tuberculosis epidemiology, following the pathogenesis of tuberculosis.[10,11]

1.1.1 Exposure to TB

The main factors influencing exposure to TB are the number of incident infectious cases in a community, the duration of infectiousness of these cases and the duration of exposure to these infectious cases. An infectious TB case may therefore be more likely to expose more people to TB in a densely populated urban area than in a sparsely populated rural area. Due to the fact that TB is communicated by airborne means; reducing distance between people reduces the chance of becoming infected. Overcrowded households also carry more risk than small households with few individuals living in spacious dwellings. It is also more likely for a TB case to expose people they socialise with more frequently, because there is more time for the exposure to occur. Climatic conditions may also affect exposure as people tend to frequent in indoor environments more often in colder climates than in warmer climates, again increasing the risk for exposure by airborne means.

One of the factors which most increase the risk of TB exposure is the complex factor of poverty.[12,13] Poverty is strongly associated with an increase in household crowding and poor nutrition; crowding increases risk of infection by airborne means, while poor nutrition reduces the ability of the host to resist infection.[14-16] It is also more common for those in socioeconomically disadvantaged situations to smoke tobacco,[17] misuse alcohol, be less educated regarding healthy living, and have poor working conditions.[15] All of these factors reduce the ability to resist infection. Health care workers are also at an increased risk of being exposed to TB.[18,19] TB is a significant occupational problem for health care workers especially in resource limited countries where effective infection-control programmes are lacking or in-adequate.

Duration of infectiousness

It is of utmost importance to limit the duration of infectiousness of incident cases of TB in order to limit the risk of exposure to the general population. The best way of shortening the duration of infectiousness is to diagnose the infectious cases as quickly as possible and to place them on appropriate treatment. In general, between 30% and 40% of contacts have already been infected by the time an infectious TB case is diagnosed.[20]

Duration and number of interactions

The factors that influence the interaction between infectious cases and their contacts vary considerably in time and geographical location.

Population density

Population density varies across even small distances in the same country and especially between countries.[21] Urban areas by definition have a much higher population density than almost all locations in rural areas. The risk for an infectious case to come into contact with many people would therefore be much higher in urban areas than in rural areas[12,13,15] such as in the case in the crowded conditions of urban India, versus the more sparsely populated rural areas.[22]

Urbanisation (in-migration into the cities) in developing countries has played a major role in the dissemination of TB in urban areas as poor people move to urban areas in search of work.[23] When these impoverished in-migrants arrive in the cities, they almost always end up living in informal settlements and therefore suffer the consequences of their poverty, namely poor, flimsy crowded housing, inadequate diets and living areas close to polluted waterways.[24] In developed countries immigration of young foreign-born adults creates a similar problem to that of internal migration in developing countries. These migrants and immigrants usually live in poor communities within the urban area and have poor living conditions and inadequate access to health care. The consequence of this is that even in developed countries, where there are large population clusters of foreign-born immigrants, there is a slower decline in TB incidence than would otherwise be expected.[25]

Family/Household size

The size of a family and the social arrangements within a family (e.g. sleeping arrangements and responsibility for looking after children) influence the risk of exposure to TB. In developed countries the average family size has decreased[26], but this is not the case in developing countries. It has been shown that the risk of infection of children by an infectious TB case in the home is positively associated with family size, because of increased crowding in the home.[27]

Climatic conditions

It is more common for people to spend longer periods of time indoors in countries with cold climates than in countries with warmer climates. In cold climates people are more likely to congregate indoors in closed-up houses for longer periods of time which increases the risk of being exposed if an infectious case is also living in the household – there is less circulation of fresh air indoors than outdoors, and so more opportunities for exposure to the pathogen.

Gender

The degree of social interactions differs by gender to a great extent in different societies. In some countries, women and men take part almost equally in public activities, while in other countries

women's lives are very secluded. The opportunity of becoming exposed to an infectious case both inside and outside of the home will thus differ for men and women across different cultures.[28]

1.1.2 TB infection

TB is a common infectious disease caused by various species of *Mycobacteria*.[10,29,30] In humans infection is caused mostly by *Mycobacterium tuberculosis*. TB usually affects the lungs (pulmonary TB) but can also spread to any other part of the body (extra-pulmonary TB). Pulmonary TB (PTB) is the most infectious and common form of TB disease and occurs in more than 80% of TB cases. Most people infected with TB are asymptomatic.[10] Of these asymptomatic people 1 in 10 develops active disease. Approximately 5% of exposed individuals will develop active TB disease within two years of being infected and another 5% will develop TB disease after reactivation many years after the initial infection.[10] Results from the pre-treatment era show that if active cases are left untreated, more than 50% die from the disease.[10,29-31]

An experiment conducted by Loudon and Roberts[32] greatly contributed to the understanding of the transmission of *Mycobacterium tuberculosis*. TB is almost exclusively transmitted via the airborne route, except in the rare occasion with accidental direct inoculation of infectious material into the skin.[33-36] TB is transmitted by a person with active PTB when they cough, sneeze or spit, thereby producing tiny droplets containing TB bacteria. The smaller of these airborne droplets can be carried in air currents for long periods of time when indoors. When a person inhales these droplets he/she can get infected with TB.[10,37,38] The successful transmission of TB therefore requires that a person produce airborne infectious droplets. This therefore limits the potential for transmission to those with PTB except in rare instances, such as when a person with laryngeal extra pulmonary TB transmits TB.

The potential for infecting a susceptible person with TB is strongly associated with the number of TB bacilli found in the infected person's sputum. A number of studies have shown the difference in infectiousness between sputum smear-positive, sputum smear-negative, culture-only positive and culture-negative TB cases.[39-41]

The factors that determine the likelihood of TB transmission are.[10,37,38]

- The number of *Mycobacterium tuberculosis* organisms expelled into the air within the droplets;

- The concentration of organisms in the air, which is influenced by the volume of the space and its ventilation;
- The length of time the person is exposed to the contaminated air.

Susceptibility to TB infection and disease is also determined by a combination of genetic, environmental and host related factors.[42-45] Factors directly associated to the bacteria such as strain virulence also plays a role in the probability of infection.

1.1.3 TB disease

TB can affect everyone but those most at risk for developing TB disease are the elderly, young children (those under 5 years of age) and those with weakened immune systems, including people who are infected with the human immunodeficiency virus (HIV) or those with overt acquired immune deficiency syndrome (AIDS). Other immunosuppressive diseases such as silicosis and diabetes mellitus also increase the risk of developing active TB.[9,10,46,47]

The HIV epidemic is fuelling the TB epidemic worldwide.[48] HIV infection increases the risk of progression to TB disease for those with a pre-existing infection with TB. In HIV-negative persons infected with TB the lifetime risk of developing TB disease is between 2% and 23%.[49] In HIV-positive persons infected with TB the annual risk of developing TB disease is around 10%.[38,48,49] The higher burden of TB due to HIV-infection increases risk of TB transmission to the general population, but this risk is mitigated by the fact that HIV-infected TB patients are generally less infectious.[50]

Other host related risk factors for overt disease given TB infection are male gender, age (adolescents and young children), alcohol misuse, malnutrition and smoking.[51-53] Smoking is a well described risk factor for developing TB disease as well as for death from TB.[54-57] A number of studies among high risk groups have shown that smoking also increases the risk of TB infection.[58-61] Only a limited number of studies investigated the association between passive smoking and TB disease.[54,62-64] A meta-analysis on these studies indicated that the risk of TB disease among children exposed to passive smoking was significantly higher than among adults.[56]

Malnutrition has been shown to be directly associated with developing TB disease from primary or latent TB infection.[65] Malnourished individuals usually have low immune defences, and therefore these individuals have a higher likelihood of developing TB disease after infection.

1.2 DESCRIPTIVE EPIDEMIOLOGY OF TB

The terms "notification rate", "incidence" and "recorded and reported cases (expressed per population)" are often erroneously used interchangeably to indicate the number of new TB cases per population per year. The "notification rate" refers to the number of cases that are reported to the Department of Health – "notified" – and therefore underestimates the actual number of TB cases (since some cases are undiagnosed or even unknown to the patients themselves and some diagnosed cases are not notified) and thus does not provide a reliable estimation of the prevalence or true incidence of TB. The definition of "notified tuberculosis cases" is different from country to country, for instance, in some countries notified tuberculosis cases include pulmonary cases but not extra-pulmonary. The number of notified cases might be influenced, for example, by quality of diagnosis and reporting activities. For instance, cases diagnosed and treated in the private sector are not included in official statistics in most developing countries.

Incidence *rates* refer to the number of new cases of illness occurring during a specific time period in a specific area. The same reservations about representativeness apply to recorded and reported cases (expressed per population). It is important to use these terms as they are defined epidemiologically and in the right context. The use of TB notification as a proxy for TB incidence should therefore be considered in the context of its limitations. TB statistics based on TB notification do not always directly reflect the epidemiological situation of TB in most countries. This is due to the fact that increases in the trend in the number of TB cases do not necessarily indicate a worsening in the TB situation. The increase in the number of cases could be due to improved diagnosis and reporting activities.

Recording of TB cases constitutes the act of capturing the details of a TB case in the country's official TB recording system. In some countries this recording system could be a paper based register or an electronic register. Once a TB case is recorded in the official TB register, the TB case is notified to the health services. The recording of TB cases is therefore analogous to notification and is now the mechanism used world-wide to report recorded cases to the World Health Organisation (WHO) annually.

Reporting on the other hand constitutes producing statistics on the specifics of the recorded cases. These reports could be based on different levels in an organisational structure (facility, sub/district etc.) and for various frequencies (monthly, quarterly, annually etc.).

1.2.1 Measuring TB infection

Tuberculin is produced from killed tubercle bacilli and was first introduced by Koch as a proposed treatment for TB and a diagnostic tool for TB. The idea of tuberculin being a TB treatment was found to be erroneous but has remained an important diagnostic tool for identifying TB infection. Tuberculin skin tests (TST) are still today used to identify TB infection. Tuberculin surveys are used to provide information on the annual risk of tuberculosis infection (ARTI) in communities.[66,67] TST surveys are a measure of current or recent TB infection.[68]

TST surveys have however recently been shown to have serious limitations due to cross-reaction caused by Bacillus Calmette-Guerin (BCG) vaccination and environmental mycobacterium.[69,70] Other limitations of TST are that the test would be less likely to be positive in children with severe malnutrition, HIV infection, disseminated TB such as military disease or TB meningitis, or children who are on immunosuppressive drugs. TST surveys are usually also only done in school children and are therefore not representative of the total population.[71]

1.2.2 TB prevalence

Prevalence surveys are the most direct epidemiological method for measuring the burden of TB. A TB prevalence survey aims to obtain an accurate estimate of the TB prevalence in a country or specific area. In order to obtain these estimates a pre-determined sample of the population in a country or research area undergoes diagnostic testing for TB. One of the major limitations of prevalence surveys is that they are extremely expensive and therefore not always a viable option in resource limited countries.[72] Other limitations of TB prevalence surveys are their inability to estimate the burden of childhood TB or to estimate the prevalence of extra-pulmonary TB. Prevalence surveys are usually only conducted on adults as it is very difficult to collect sputum from children. The same reason applies to the fact that TB prevalence surveys are usually not conducted for extra-pulmonary TB as extra-pulmonary TB is difficult and expensive to diagnose. Even though TB prevalence surveys are expensive they could be justified in high-burden countries where many TB cases and deaths are missed by routine reporting.[68]

1.2.3 TB incidence

TB incidence is measured as the number of new TB cases in a given time period. This is usually reported as the total number of incident cases per year for a specific population.[73] A limitation of incidence as an epidemiological measure is that incidence changes slowly in response to control

efforts compared to prevalence or mortality. TB incidence is also very difficult to measure directly and is therefore rather measured indirectly through the assessment of the completeness of TB case recording, measures of the prevalence of TB infection through TST surveys, and through estimates from TB mortality data.

Direct measure

Longitudinal cohort studies are very seldom conducted due to low TB incidence rates, even in high burden areas. It is also difficult to distinguish between active and incident TB with available clinical and diagnostic methods. These studies are also generally costly, logistically demanding and cases may be missed if follow-up periods are too long. This method could however be more feasible in cohorts of individuals at high risk of developing active TB, such as those infected with HIV.[68]

Prospective studies to measure TB incidence have been conducted in the Republic of Korea[74,75] and in India.[76] The high cost of these studies limits the viability of undertaking such longitudinal studies. The follow-up of participants could also be a limitation especially where patients seek treatment in unmonitored health sectors, for instance the private sector. These studies however have an important role to play in vaccine and clinical trials.[77-80]

Indirect measure

TB recording and reporting data are used as an indirect measure of incidence. This method is only valid if all or almost all TB cases are diagnosed and that all or almost all diagnosed cases are entered into the TB recording system. It is therefore imperative that a thorough assessment of the recording and reporting system is made before this data can be used as an accurate measure of TB incidence.[68]

Surveys of ARTI are used as an indirect measure for TB incidence.[2,67,81] ARTI can be estimated from the prevalence of infection in a TST survey. In principle TST surveys are reasonably easy to conduct, but making a valid estimate of TB incidence from ARTI is very complicated, making this a less feasible method for estimating incidence.[82,83]

Using studies of the prevalence of TB disease, incidence may be estimated as the prevalence of TB disease divided by the estimated average duration of disease in years.[84,85] In order to estimate incidence from a prevalence survey, the sample size has to be increased considerably, even to the extent of doubling it. This greatly increases the cost of such surveys, and is therefore not always a viable method for estimating incidence, especially in low and middle income countries.[86]

Using mortality data recorded in vital registration systems is another possible method for estimating incidence. In this case, TB incidence is estimated as the number of TB deaths divided by the estimated case-fatality rate.[2] However, accurate and complete vital registration systems that include information on deaths from TB are not widely available. A review of vital registration data in 2003 indicated that only 23 of the 115 countries, which reported deaths and their causes, had data of high quality.[87] None of 22 high burden TB countries were included in the 23 countries with high quality vital registration data.

1.3 TB GLOBALLY

In 2007 there were an estimated 9.27 million incident cases of TB which was an increase from the 2006 (9.24 million), 2000 (8.3 million) and 1990 (6.6 million) estimates. Most of these estimated cases were in Asia (55%) and Africa (31%). The countries ranked the highest with total number of TB cases in 2007 are India (2 million), China (1.3 million), Indonesia (0.53 million), Nigeria (0.46 million) and South Africa (0.46 million). An estimated 1.37 million (15%) of the 9.27 million estimated incident TB cases were HIV-positive. Of these 79% were in the African region.[73]

Table 1-1: Countries with TB ranked the highest by estimated epidemiological burden of TB in 2007.[73]

Rank	***Country***	***Population***	***Incidence***	
		(1000s)	*All forms of TB Number (per 1000)*	*Smear positive Number (per1000)*
1	India	1 169 016	1 962	873
2	China	1 328 630	1 306	585
3	Indonesia	231 627	528	236
4	Nigeria	148 093	460	195
5	South Africa	48 577	461	174

Global deaths among HIV negative incident cases of TB were estimated at 1.3 million in 2007 and an additional 456 000 among those that are HIV positive.[73] The estimated numbers of HIV-positive TB cases and deaths in 2007 are approximately double the numbers published by the World Health Organisation (WHO) in previous years. These increases are not due to the actual

number of deaths or cases doubling but due to better reporting from certain countries, especially those in the African region. [73]

In 2007, 5.5 million TB cases were notified globally by the Directly Observed Treatment Short-course (DOTS) programmes of which 2.6 million cases were smear-positive. The case detection rate of new smear-positive cases under DOTS was 63% which is still under the target of ≥70%.[73]

1.3.1 TB in South Africa

In South Africa the number of registered cases of all forms of TB (i.e., recorded in the TB treatment registers and reported to the National Department of Health) increased from 224 420 in 2002 to 405 699 in 2009.[88] The incidence rate for all forms of TB increased during the same period from 424 cases per 100 000 population to 823 per 100 000 population.[88] The incidence rate of TB varied between provinces with the Western Cape (WC) ranked 4th between provinces (Table 1-2).

Table 1-2: South African case findings per province and for the country (2007).[1]

Provinces	***Population***	***All TB***	***All TB per 100 000***	***Incidence of New Smear positive pulmonary TB cases per 100 000***
Eastern Cape	6 648 600	63 807	960	373
Free State	2 902 400	24 940	859	331
Gauteng	10 531 300	51 660	491	204
KwaZulu-Natal	10 449 300	122 642	1174	303
Limpopo	5 227 200	22 836	437	181
Mpumalanga	3 606 800	27 511	763	311
North West	1 147 600	31 682	2761	980
Northern Cape	3 450 400	10 503	304	119
Western Cape	5 356 900	50 118	936	293
Total	**49 320 500**	**405 699**	**823**	**282**

The Western Cape has one of the lowest HIV prevalence rates in South Africa. The Western Cape prevalence rate can be deduced from the estimated prevalence of 16.9% in 2009 among antenatal clinic attendees, which was just above half of the national estimate of 29.4%.[89] Even though these estimates are based on pregnant woman who access health care, antenatal surveillance is internationally recognised as one of the most useful ways of assessing HIV prevalence in countries with generalised epidemics.[90]

The low HIV prevalence and high TB incidence in the Western Cape compared to the other provinces would imply that the TB epidemic in the Western Cape is not solely related to the HIV epidemic. The number of all forms of TB registered in the Western Cape increased from 39 650 in 2002 to 50 118 in 2009. The TB incidence rate for all forms of TB increased from 917 per 100 000 population in 2002 to 936 per 100 000 population in 2009.[88]

1.4 TB NOTIFICATION

A notifiable disease is any disease that is required by law to be reported to government authorities. This collection of disease information allows the authorities to monitor the disease, and to provide early warning of possible outbreaks. Many governments have implemented regulations for reporting of diseases. The International Health Regulations 1969 of the WHO require disease reporting to the organisation in order to help with its global surveillance and advisory role. These regulations identify a number of specific diseases and define a limited set of criteria to assist in deciding whether an event is notifiable to WHO.[91]

The WHO has been collecting TB control data from all countries since 1995. A standard collection form[92] is proposed and NTP are expected to send data on these forms to WHO. The collection form consists of three components: case notification and treatment outcomes; data related to implementation of the Stop TB Strategy[73] and financing. All collection forms are systematically reviewed by WHO personnel located in each country, at regional offices and at headquarters. If data are deemed to be incorrect, the appropriate NTP correspondent is contacted to correct or clarify the errors. If the data appear to be inconsistent with data from previous years the correspondent is also contacted. The completed data are then used as the basis for producing a final dataset from which country profiles, summary analyses and regional and country-specific data are produced for the annual WHO Global Tuberculosis Control report.[73]

The aim of the annual WHO Global Tuberculosis Control report is to provide an up-to-date assessment of the TB epidemic and progress in controlling the disease at global, regional and country levels, in the context of global targets set for 2015.

1.4.1 TB notification in South Africa

TB is a notifiable disease in South Africa as in most other countries. South Africa has a routine notification system for reporting notifiable medical conditions. The South African Disease Notification System is a passive surveillance system that collects information on Notifiable

Medical Conditions. The National Department of Health (NDoH) manages the Disease Notification System. The notification of certain medical conditions in South Africa is based on the government's Health Act, Act No. 61 of 2003, coupled with regulations on the reporting of specific diseases to the Local, Provincial and/or National Health Department. There are currently 33 medical conditions that are notifiable. Some medical conditions have been sub-divided resulting in more than 40 different conditions which are notifiable.

The notification system has several objectives:

- At the national level, it helps the NDoH to plan and implement health promotional and intervention strategies.
- It helps the NDoH to monitor disease trends over time. In time this will permit an evaluation of the effectiveness of promotional and intervention strategies.
- At provincial level it helps to implement immediate interventions.

1.5 TB CONTROL

Dr Karel Styblo of the International Union Against Tuberculosis and Lung Disease (IUATLD) pioneered the development of a model of TB control in 1970.[93] This TB control model was based on a managerial approach of case-finding and treatment. The IUATLD supported nine countries between 1978 and 1991 based on the model proposed by Dr Styblo. The WHO Global Tuberculosis Programme declared TB as a global emerging disease in 1993. As a result the WHO introduced Dr. Styblo's strategy as a technical and management package known as Directly Observed Treatment Short-course (DOTS). DOTS is a community-based tuberculosis treatment and care strategy focused on a public health approach. The most important aspect of effective TB control is to interrupt the chain of transmission in a community by reducing the source of TB infection through early detection and treatment of infectious TB cases.[3,94] In 1995 this strategy was first introduced as a global TB control strategy and since 2005 has been known as the Stop TB strategy.[4] In order to have an impact on the TB epidemic, a TB programme should achieve a case detection rate of at least 70% of new smear-positive TB cases and successfully treat 85% of these cases. Reaching these targets should have an impact on TB transmission and reduce the incidence of TB.[95,96]

The five components of the DOTS strategy are:[4,93]

- Government commitment to sustain TB control.

- Case detection by sputum smear microscopy among symptomatic patients who self-report to health services.
- Standardised treatment regimen of 6 to 8 months for at least all confirmed sputum smear-positive cases, with DOTS for at least the first 2 months.
- A regular and uninterrupted supply of all essential anti-TB drugs.
- A standardised recording and reporting system that allows assessment of treatment results for each patient and of the TB control programme overall.

Even though the DOTS strategy had a major effect on global TB control, global statistics indicated that DOTS alone would not be sufficient to achieve global TB control and elimination. Recognizing this limitation of the DOTS strategy the World Health Assembly (WHA) developed a new strategy, The Stop TB Strategy, in 2005 which built on and enhanced the achievement of DOTS. The Stop TB Strategy was launched in 2006 on World TB Day. This strategy was designed in order to meet the TB-specific Millennium Development Goals (MDGs) developed by the WHA and the Stop TB Partnership.[2,97-99] The MDGs were set to be reached by 2015 with specific indicators identified in order to monitor the progress toward these targets. The Stop TB Strategy was specifically targeted at the Global Plan to Stop TB by 2015. The Stop TB Strategy has a vision of a world free of TB. The goal is therefore to dramatically reduce the global burden of TB by 2015. The Stop TB Strategy has the following objectives:[2,100]

- To achieve universal access to high-quality diagnosis and patient-centered treatment.
- To reduce the suffering and socioeconomic burden associated with TB.
- To protect poor and vulnerable populations from TB, TB/HIV and multidrug-resistant TB.
- To support the development of new tools and enable their timely and effective use.

The goal and target set by the MDG specific to TB are:[2,100]

- MDG 6, target 8 – to have halted and begun to reverse the incidence of TB by 2015.
- Targets linked to the MDGs and endorsed by the Stop TB Partnership:

    - By 2005, to have detected at least 70% of new sputum smear-positive TB cases and cured at least 85% of these cases.
    - By 2015, to have reduced TB prevalence and death rates by 50% relative to 1990 levels.
    - By 2050, to have eliminated TB as a public health problem (<1 case per million population).

1.6 TB CONTROL IN SOUTH AFRICA

The South African NTP adopted the WHO DOTS strategy in 1997 and this was updated to the WHO Stop TB strategy in 2006. At present, the South African NTP consists of four levels within the general health service. At national level the NDoH co-ordinates, facilitates and evaluates TB services countrywide. Implementation and budgeting of the TB programme is managed at provincial level. The district level manages primary health care and heads the administration of health services. Within each district Primary Health Care (PHC) facilities (clinics) provide primary health care to the community. The PHC consists of rural hospitals, health centres, dispensaries and clinics. This structure varies in some provinces: for instance, in the Western Cape, with a regional level functioning between provincial and district level and some districts are further subdivided into sub-districts. The City of Cape Town is one such an intermediate level.[1,101]

The TB Control Programme in the City of Cape Town is jointly administrated by the Provincial Government of the Western Cape (PGWC) and the Local Authority (LA). The Province of the Western Cape funds the NTP specific activities. The 35 Provincial Community Health Centres provide TB diagnostic services and this is the first point of entry for many TB patients. Other points of entry include hospitals, private practitioners and Local Authority Clinics. TB patients are registered in TB treatment registers at 99 reporting units (LA Clinics, Brooklyn Chest Hospital and two prisons) in the City of Cape Town. TB treatment is rendered at 121 treatment units. Coordinators at district level support the NTP in Cape Town through training and supporting clinic staff and by co-ordinating, monitoring and evaluating services delivered. The PAWC Metropole Regional Office has overall responsibility for monitoring and evaluation.[101]

The South African Department of Health has set the following TB priorities:[102]

- Building political commitment in order to raise the profile of TB and to secure sufficient resources to achieve the international targets.
- Providing good access to laboratory testing as a precondition for early detection of TB.
- Ensuring an uninterrupted supply of quality drugs through reliable suppliers and distribution systems.
- Ensuring the technical soundness of Directly Observed Treatment, using standard short-course chemotherapy, and the availability of social support.
- Implementing regular recording and reporting systems in order to assess the treatment outcome of each TB patient.

- Building partnerships between all levels of government, non-government and private sectors.
- Developing and implementing a policy on multi-drug resistant TB.
- Developing and implementing an advocacy and social mobilisation plan.
- Ensuring easy access to voluntary counselling and HIV-testing for all TB patients.

1.7 TB CASE DEFINITION

The use of a clear case definition is extremely important in epidemiology in order to standardise criteria for the identification of a case. In epidemiology all case definitions should include the three epidemiological variables: time, place and person. It is therefore imperative that a precise definition of a case is formulated. This is in order to accurately monitor the trends of reported cases, to detect any unusual occurrence (outbreaks) and to evaluate the effectiveness of interventions. The usefulness of surveillance data therefore depends on the uniformity, simplicity and timeliness of case definitions.[6]

The following definitions are common for case definitions in public health surveillance and are defined at different levels of certainty:

- Confirmed case: a case that is classified as confirmed for reporting purposes.
- Laboratory confirmed case: a case that is confirmed by one or more laboratory methods listed in the laboratory requirement of the case definition.
- Probable case: a case that is classified as probable for reporting purposes.
- Suspected case: a case that is classified as suspected for reporting purposes.
- Epidemiologically linked case: a case in which the person has had contact with a person or persons who have the disease and where the transmission of the agent by the usual modes of transmission is plausible.

The WHO case definition for TB includes the anatomical site of disease, the bacteriological result, the severity of disease and the history of previous TB treatment.[2,100,103]

Anatomical site of disease: This includes (i) pulmonary TB and disease affecting the lung parenchyma and (ii) extra-pulmonary TB (which may affect lymph nodes, pleura, meninges, pericardia, peritoneum, spine, intestine, genitourinary tract, larynx, bone and joints, or skin).

Bacteriological results: Smear-positive and smear-negative are defined for pulmonary cases and also correlate with infectiousness. In some settings where culture facilities are available, the result

of the culture is also included in the classification. In most settings, especially low and middle income countries, only microscopy laboratory services are available. TB cases diagnosed with both pulmonary and extra-pulmonary TB are classified as pulmonary TB.

Severity of disease: Bacillary load, extent of disease and anatomical site are factors that determine the severity of TB disease, and consequently its appropriate treatment.

The history of previous TB treatment: This provides information as to whether a TB cases had any previous anti-TB treatment. Previously treated patients are at higher risk of developing multi-drug resistant TB. These TB cases should therefore be investigated for drug susceptibility. TB cases with a history of previous TB treatment require a treatment regimen that differs from patients who were never previously treated.

1.8 TB CASE DEFINITION IN SOUTH AFRICA

The NTP in South Africa adopted the WHO standard case definition for defining a TB case which includes site of TB disease, severity of TB disease, bacteriology (sputum smear result), and history of previous treatment of TB.[1,30,101]

1.8.1 Site of disease

- Pulmonary TB refers to disease involving the lung parenchyma.
- Extra-pulmonary TB refers to TB of the organs other than the lungs: e.g. pleura, lymph nodes, abdomen, genito-urinary tract, skin, joints and bones, meninges.

A patient with both pulmonary and extra-pulmonary TB constitutes a case of pulmonary TB. The case definition of an extra-pulmonary case with several sites affected depends on the site representing the most severe form of disease.

The standard case definition is used for proper patient registration and case notification, to evaluate the trend in the proportions of new smear-positive cases and smear-positive relapse and other treatment cases, to allocate cases to standardised treatment categories and for cohort analysis. These case definitions are also used to give priority to the most infectious cases, and to increase cost-effective use of resources and to minimise side-effects for patients.

1.8.2 Severity of disease

The extent of disease and anatomical site determine the severity of disease and appropriate treatment. TB disease is considered severe if there is a significant threat to life or risk of long term consequences.

1.8.3 Bacteriology or sputum smear result

Cases are declared smear-positive PTB when:

- There are at least 2 sputum smears positive for acid-fast bacillus (AFBs) or
- 1 sputum smear positive for AFBs and chest X-ray abnormalities consistent with active TB or
- culture positive TB, or 1 sputum smear and clinically ill.

It is advisable that even if the first specimen is positive pre-treatment, another specimen should be taken. This will reduce the chances of a false-positive result as administrative errors may occur. The case definition for smear-positive TB cases was reviewed in 2009 to only one positive sputum required. This change was not uniformly implemented throughout South Africa and will make the comparison of TB incidence with previous years and between different areas within South Africa difficult.[1]

Cases are declared smear-negative PTB when:

- At least 2 sputum smears are negative for AFBs.
- Chest X-ray abnormalities are consistent with active TB.

1.8.4 History of previous treatment

It is important to define a case according to whether or not the patient has previously received TB treatment in order to identify those patients at increased risk of acquired drug resistance and prescribe appropriate treatment.

The important terms for this aspect are defined as follows:

- New case: A patient who has never had treatment for TB or who has taken anti-tuberculosis drugs for less than four weeks.
- Re-treatment case: A patient who has taken treatment for TB before and either relapsed, defaulted or had treatment failure.

- Relapse: A sputum smear positive pulmonary TB patient who received treatment and was declared cured (sputum smear negative) at the end of the treatment period and now developed sputum smear positive pulmonary TB again.
- Treatment after failure: A pulmonary TB patient who is still sputum smear positive at the end of the treatment period.
- Treatment after default: A patient who completed at least one month of treatment and returns after having interrupted treatment for two months or more, and still smear-positive (sometimes smear-negative but still with active TB as judged on clinical and radiological assessment).
- Transfer out: A patient already registered for treatment in one district who has been transferred to another to continue treatment.
- Chronic case: Patient who remains sputum smear positive after completing a supervised re-treatment regimen.

1.9 TB DIAGNOSIS

The effective detection of TB cases require that individuals know what the symptoms of TB are, that access to PHC is adequate and health care workers will maintain a high index of suspicion of TB when presented with an individual with TB symptoms. PHC should have the necessary access to reliable laboratories. The process of diagnosing a TB case from the identification of a TB suspect to a confirmed laboratory result is a very complex set of activities and the failure at any stage can cause delays or misdiagnosis.[2,73,100,103]

1.9.1 TB symptoms

The symptoms of pulmonary TB are:[104-107]

- A persistent cough for more the 2-3 weeks.
- A productive cough producing sputum.

The symptoms are often accompanied by other, nonspecific symptoms:

- respiratory symptoms: shortness of breath, chest and back pains, haemoptysis;
- constitutional symptoms: loss of appetite, weight loss, fever, night sweats, fatigue.

The symptoms for extra-pulmonary TB are directly related to the specific extra-pulmonary sites which include lymph nodes, pleura, larynx, meninges, genitourinary and intestinal tracts, bone, spinal cord, eye and skin.

1.9.2 Sputum smear microscopy

All TB suspects should have a sputum specimen collected for microscopic examination especially those suspected of having pulmonary TB. Microbiological diagnosis is confirmed by culturing mycobacterium tuberculosis from any suspected site of disease. In many resource-limited countries, neither culture nor rapid amplification methods are available or feasible. In these countries the diagnosis of TB may also be confirmed by the presence of AFB in sputum smear examination. Almost two-thirds of pulmonary TB cases will be diagnosed by repeated sputum smear microscopy. The identification of AFB by microscopy in high incident TB areas is very specific for mycobacterium TB. Sputum smear microscopy is the fastest method for identifying suspect cases with TB and identifies those who are most at risk of dying or transmitting TB. The WHO recommends the microscopic examination of two sputum specimens (formerly three) based on the evaluation that in the first specimen 83–87% of all patients AFB are detected, a further 10-12% in the second specimen and a further 3-5% in the third specimen.[100,103,108]

1.9.3 Diagnosis of smear-negative TB

For smear-negative and extra-pulmonary TB, a diagnosis by a clinician specially trained in TB may be required as well as X-ray examination. The diagnosis of smear-negative TB is always presumptive and should be based on other clinical and epidemiological information, including failure to respond to a course of appropriate broad-spectrum antibiotics and exclusion of other pathology.

1.9.4 Diagnosis of culture confirmed TB

The only definitive diagnosis of TB is made by means of an examination by mycobacterial culture. However the diagnosis of TB by means of culture has some limitations. These limitations include a longer turn-around time before a result is available, much higher expense than for smear microscopy and the requirement of specialised laboratory facilities with skilled staff.

1.10 TB DIAGNOSIS IN SOUTH AFRICA

The first point of entry for many TB cases into the health system is their nearest community clinic which acts as a diagnostic centre (collection points for samples for TB diagnostic purposes) as well as a treatment centre. Other points of entry include public and private hospitals, correctional facilities (prisons) and private practitioners. Individuals suspected of having TB and TB cases

presenting to private practitioners or other non-NTP health facilities should be referred to their nearest community clinic for TB diagnosis and treatment. A study conducted at Chris Hani Baragwanath Hospital in Johannesburg indicated that 21% of the patients referred to a clinic from the hospital did not attend any clinic in Soweto and were also not recorded in the TB treatment register.[109,110]

TB services are free at all NTP PHC facilities but this may not be the case at most non-NTP and private practitioners.[111] TB treatment may not be available to clients without private medical insurance when they seek care at these non-NTP practitioners.[112] TB suspects who present to non-NTP practitioners may not be diagnosed or diagnosis could be delayed as there practitioners have a lower index of suspicion for TB compared to public services. Currently there is very little known about the impact private practitioners and other non-NTP PHC have on the TB burden in South Africa.

TB case finding depends on individuals with TB symptoms presenting to a health facility (passive case finding). The problem of stigma against those with TB or individuals with asymptomatic TB not knowing or realising that they have TB may prevent those with TB from presenting at PHC clinics. There is the possibility that TB suspects presenting with TB symptoms are not diagnosed for TB. These infectious cases will stay untreated and continue to spread TB in the community.[113] In South Africa this problem is worsened by the following factors:

- The lack of trained health staff to identify those with TB.
- The shortage of staff specifically dealing with TB, high staff turnover and poor supervision of staff.
- Unmotivated staff due to difficult work conditions and overload due to understaffing and large numbers of patients.

Individuals suspected of having TB should have a sputum sample collected for bacteriological confirmation of TB. Once sputum samples are collected from TB suspects, the samples are sent to the National Health Laboratory Service (NHLS).

1.11 TB Laboratory Services for TB diagnosis

According to the WHO recommendations, TB laboratory services should be organised taking account of accessibility to the whole population and provision of all the necessary services for efficient TB case-management.[114] The NTP of some countries has built-in or fully integrated

laboratory networks, while in some countries TB laboratory services are integrated into the general health system or provided by completely independent organisations at all or certain levels. When the laboratory network has been dissociated from the NTP, it must be resolved by establishing good coordination to ensure functional integration of the network into NTP to obtain dependable TB laboratory services.

The WHO recommends TB laboratory services to be integrated within TB control programmes.[114] The TB control programme should in turn be integrated as part of the overall primary health care programme of a country. TB laboratory services should therefore follow the following recommended structure:

- Peripheral or district laboratory.
- Intermediate or regional laboratory.
- Central or national laboratory.

Peripheral laboratories should be capable of performing sputum smear microscopy. These laboratories should be integrated with primary health care services and could be located at primary health care centres or district hospitals. Intermediate laboratories should be capable of providing supervision, monitoring, training and quality assurance to peripheral laboratories. Mycobacterial culture and differentiation between mycobacterial TB and other mycobacterial species could be performed at intermediate laboratories. These intermediate laboratories could be integrated within existing public health laboratories.

Central laboratories should be at the highest level of the laboratory structure and should be capable of doing microscopy, mycobacterial culture, drug susceptibility testing and species identification. Central laboratories could be separate from the public health laboratories and could reside at research institutes or in a country's principal TB or public health institute. These central laboratories should also provide training for laboratory staff, perform quality assurance and proficiency testing, exercise surveillance of primary and acquired tuberculosis drug resistance and participate in epidemiological and operational research.

1.12 Laboratory Services for TB diagnosis in South Africa

In South Africa microscopy examination is the standard method for TB diagnosis with culture only done on request, or for re-treatment cases and high risk cases (e.g. HIV positive).

The NHLS provides a centralised laboratory service with laboratories situated across South Africa at major health service points. NHLS does not form part of the NTP but is part of a para-statal which is the sole provider of laboratory services to people in the public health services. Public hospitals also make use of NHLS services but these facilities have laboratories on-site. Even though the clinics serve as diagnostic collection centres for samples, no diagnostic analysis services are offered at these clinics. All clinics and other service points acting as diagnostic centres in reality only offer a sputum sample collection service to TB suspects, while TB diagnostic services are provided only at NHLS laboratories or, in the case of private patients, private pathology services. It is often said that private patients make up only a small fraction of the TB case load in the country, but no data is available to corroborate this.

Currently there is no strict control and tracking of samples sent to the laboratory at clinic level. There is no indication of how many people presented to clinics as TB suspects but for whom no diagnostic answer was received because of lost samples.[115,116] The number of samples sent for analysis is not reconciled with the number of suspected TB cases seen at the clinic and the latter number is unknown at most clinics. If there is a significant loss of samples somewhere in this diagnostic chain, it represents patients with possible TB who slip through the net. There are many steps to complete for preparing and sending sputum samples to the laboratory. If any of these steps are not completed properly it may delay or prevent a TB suspect from being diagnosed.

The following steps need to be completed for sputum samples:[1]

- Sputum labelling: All sputum containers should clearly be labelled with the name of the clinic, name of the client, client's clinic or hospital number, the date the specimen was collected and an indication if it is a diagnostic or follow-up specimen. If any of this information is incorrect the specimen may not be able to be traced back to the original client or facility.
- Sputum collection: The correct method for collecting sputum should be followed under supervision of the healthcare worker. Care should be taken though that this should occur in a well-ventilated and private area as to not infect other clients or healthcare workers. The healthcare worker should make sure that the lid of the container is securely fastened in order to prevent leakage during transit.
- Completion of the laboratory form: The laboratory form should be accurately completed with clear instructions on which laboratory investigation is required. Incorrect or incomplete forms may delay the diagnosis of TB.

- Storage of sputum: The sputum containers should be stored in the plastic bag provided by NHLS together with the laboratory form. Care should be taken to place the correct sputum container and laboratory form together. If a mix-up does occur the incorrect result may be returned to a client who may therefore be misdiagnosed.
- Transportation of sputum specimens: The specimens should be transferred to the laboratory as soon as possible. In order to ensure the viability of the specimens, the specimens should be transferred in a cooler bag. High temperature and direct sunlight will kill the bacilli.

Results from NHLS are regularly faxed back to the PHC from which the sputum sample was sent. The results from the laboratory are subject to various errors which include poor quality of specimens, clerical errors, handling errors, laboratory processing problems and poor quality control. If a laboratory report does not match a patient's clinical information the reports should be interpreted with extreme care. If a result from a TB suspect is expected by the TB staff at the PHC and not received, the TB staff are required to follow-up the result by phone. It is however possible those results are never received or followed up, which results in a TB case that could be missed. It is the responsibility of the TB suspects to return to the PHC clinic in order to receive their results. However if a TB case does not return for his or her results, the TB staff are required to locate the individual and request them to return to the PHC clinic to commence treatment. It does however happen that TB cases are never located and never placed on treatment (initial defaulter), die before commencing treatment or commence treatment at a different PHC. A study conducted in the Stellenbosch district of the Western Cape indicated that 17% of TB cases with two smear positive results were not recorded in the TB treatment register and were therefore initial defaulters.[115]

1.13 TB RECORDING AND REPORTING

Good recording of TB cases is essential for the effective management of TB.[100] The evaluation of the performance of a programme as well as the epidemiological trends are the basis for programmatic and policy development. Appropriate recording and reporting systems are therefore required for effective monitoring. High-quality TB patient care and information-sharing rely on the accuracy and completeness of these systems.

In recent years these recording and reporting systems have become more complex and with the introduction of computers in health facilities as standard equipment, many NTPs have adopted or are adopting electronic TB registration.[100] These electronic TB registers (ETR) could potentially

provide more accurate and complete data entry, improve the transfer of data to other levels of the health system, and provide better and timelier analysis for the evaluation of programme performance. It should be remembered though that any electronic system is only as good as the data being recorded into the system and the manual process involved. These ETR systems mirror the paper-based recording and reporting systems. The ETR system should give a facility capacity to produce regular quarterly reports and registers to facilitate the analysis of data and data verification.

The WHO recommends that the recording and reporting systems be part of the general health information system.[100] This should include detailed patient forms that are completed at the treatment facility and with the patient details summarised in the laboratory register and medical register. This information is aggregated to produce quarterly reports on TB case reporting and recording activities and results. District level annual reports should also be produced to be sent to central level. The recording and reporting system should be used to systematically evaluate the progress and treatment outcome of a TB case. The system should also, through cohort analysis, monitor overall NTP performance.

The recording system should comprise: (i) laboratory registers, which record all symptomatic patients who have had a sputum smear examination; (ii) patient treatment cards, detailing the regular intake of medication and follow-up sputum examinations; (iii) identity cards, which are kept by the patient; and (iv) basic management unit (BMU) level (usually district level) TB registers, which list each patient starting treatment and monitor progress towards cure. Some facilities have additional registers, e.g. for TB suspects, culture results, contacts, referrals and transfers, which are adapted to country needs.

The reporting system consists of: (i) quarterly reports on TB case registration, which summarize the numbers of TB patients started on treatment, laboratory tests performed and HIV tests and results obtained; (ii) quarterly reports, which detail treatment outcomes and TB/HIV activities after all patients in the cohort have completed their course of treatment; (iii) quarterly order forms, which specify the required anti-TB drugs; (iv) quarterly order forms, which detail the required laboratory supplies; (v) annual reports on programme management, which describe the human resource and TB service delivery facilities as well as the contribution of the private sector and community to referral, diagnosis and treatment.

The WHO recommended structure of a recording system consist of the following:[100]

- At BMU level

  - TB laboratory register
  - BMU TB register
  - Quarterly report on TB case registration in BMU
  - Quarterly report on TB Treatment outcomes and TB/HIV activity in BMU
  - Quarterly order form for TB drugs
  - Quarterly order form for laboratory supplies in BMU
  - Yearly report on programme management in BMU
  - Quarterly report on sputum conversion, which is optional.

- At BMU level using routine culture and drug susceptibility testing (DST)

  - TB laboratory register for culture
  - TB registration in BMU using routine culture and DST
  - Quarterly report on TB case registration in BMU routine culture
  - Quarterly report on TB Treatment outcomes and TB/HIV activities in BMU using routine culture
  - Quarterly order form for culture and DST laboratory supplies in BMU

- At health facility in central, regional or peripheral level

  - Request for sputum smear microscopy examination
  - Request for sputum smear microscopy examination, culture, DST
  - TB Treatment card
  - TB identity card
  - TB treatment referral/transfer
  - Register of TB suspects (optional)
  - Register of TB contacts (optional)
  - Register of referred TB cases (optional)

This recommended recording and reporting system are implemented and adapted as needed in each country.

1.14 TB RECORDING AND REPORTING IN SOUTH AFRICA

Before the introduction of the DOTS strategy in 1997, the NTP in South Africa relied on the clinician diagnosing the TB case to complete a notification form and to send this form containing the patient information to a centralised notification system. In 1996 the South African NTP adopted the WHO recording and reporting to a certain extent. The South African recording and reporting system is a paper based system consisting of the following elements:[1]

- Case identification and follow-up form: Used at facility level to record symptomatic clients reporting to the facility, to assist with the follow-up of results and initiation of treatment.
- Laboratory request form for sputum examination: A laboratory request form completed at the facility where sputum is collected. Should be completed with personal information of client and specifics regarding the tests requested to be completed at the laboratory.
- Clinic/Hospital card: The card is used at all facilities to collect all the information about the client, treatment and outcome. This is the source document used to complete the treatment register.
- Client treatment card: The card is held by the client and is used to record treatment details including daily doses taken for anti-TB medication.
- Tuberculosis treatment register: Used in all facilities to summarise key information from the clinic/hospital card on each registered client (demographic, disease classification, treatment regimen, treatment monitoring and outcome).
- Transfer form: Used in all facilities to report on the key client information from the register when the client is transferred or moved from one district or facility to another.
- Electronic tuberculosis register: This is an electronic recording and reporting system located at sub/district level. The source of data is the Tuberculosis treatment registers sent from the facilities to sub/district level. This system also provides the functionality to produce reports and cohort analysis. The follow standard reports are available:
- Quarterly reports on case detection and case finding: Completed at sub/district level and reports on the complete quarters cohort.
- Quarterly reports on smear conversion: Completed at sub/district level and reports on the previous quarter's cohort.
- Quarterly reports on treatment outcome for new and retreatment smear positive cases of pulmonary TB: Completed at sub/district level for the cohort 9 months earlier.

- Quarterly report on HIV indicators: Completed at sub/district level for the cohort registered 3 months earlier.
- Quarterly report on programme management: Completed at sub/district level.

The TB treatment registers are used to register all TB cases and to manage the treatment of TB cases. Only community clinics, TB hospitals and correctional facilities have formal TB registers where TB patients are recorded and their treatment and outcome tracked.

These TB treatment registers are updated on a daily basis by TB staff at the PHC facilities with new TB cases and on-going progress of TB treatment. The treatment registers should be monitored on a regular basis to assure accuracy and completeness. As pages from the TB treatment registers are completed they are sent to the sub-district office. In South Africa the sub-district serves as the BMU in the NTP. It is the responsibility of the TB co-ordinator or health information officer at the sub-district office to ensure that all outstanding pages are collected. The data capture person or health information officer then captures the data into the electronic TB register (ETR).

The ETR system was introduced in January 2003, replacing previous manual reports.[101] The data from the ETR are sent electronically from the sub-district or district office to the provincial level where further analyses are conducted and the data is aggregated before being sent to national level. Aggregate data from the ETR is also sent to the district health information system and the Notifiable Medical Disease System, and then incorporated into national disease notification data. Data sent to the national level are then used as a source for completing the WHO standard collection form.

1.15 ACCURACY AND COMPLETENESS OF TB RECORDING AND REPORTING

In order for NTPs to control TB effectively, routine data are required to assess the scope of the TB programme and implement appropriate intervention strategies.[117] The data are also needed to determine the progress towards the 70% case detection rate and 85% cure rate set by the Stop TB strategy.[118] Routinely collected cohort data on TB case finding and treatment outcome are the main source for these assessments. The main advantage of using routine data is that it is collected systematically and the recorded and reported cases give a reasonable indication of the incidence of TB. A major limitation of using routine data for this purpose is the uncertainty of the proportion of incident cases not diagnosed or reported.[119] The following are some limitation of routinely collected data:[6]

- Registers typically have uneven coverage.
- Technology to reach diagnosis changes over time.
- New knowledge changes inclusion and exclusion criteria.
- Hospital based or practice based registers can be very inaccurate (especially in developing countries).
- Records usually contain large gaps (information not supplied at time of diagnosis or treatment).
- Exposure to risk factors for TB is very infrequently noted in case files at the time.
- Chronic diseases or diseases with long lead times that develop a long time after exposure such as TB have incidence rates that are very difficult to determine without meticulously collected data on risk factors.
- Increases in rates may partly be due to increase in efficiency of surveillance or detection and not increases in disease occurrence. The degree of the increase attributable to better detection is usually not quantifiable with routine data.

"Accurate rates are the hallmark for effective epidemiology because they form the basis of comparison between populations, risk groups, etc." [6]

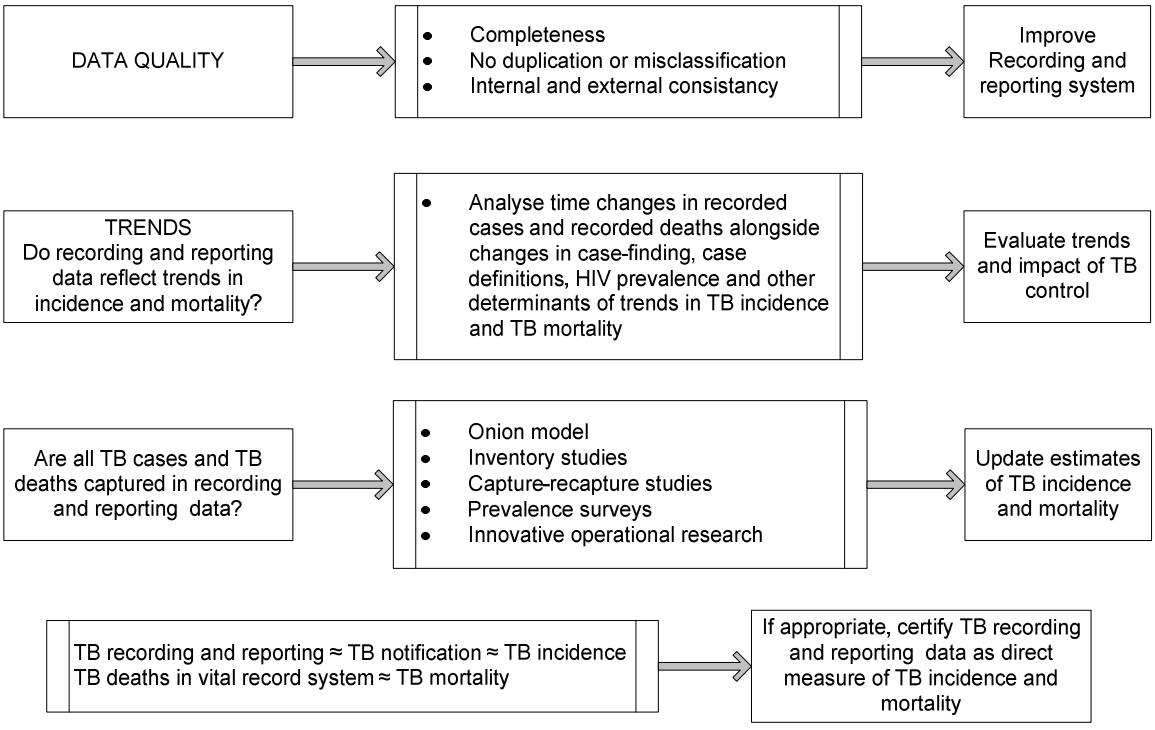

Figure 1-3: Adapted from the WHO Framework for assessment of TB surveillance notification data.[2]

In order to assess if the MDG and specifically the Stop TB goals and targets are met, good quality and a wide coverage of TB related data are required. Good quality and wide coverage of data are analogous to the accuracy and completeness of data. Accurate and complete TB recording will ensure the estimate of TB incidence from routinely collected TB data is as accurate as possible.

The approach in assessing the accuracy and completeness of routinely collected TB data consists of three components (Figure 1-3):[73]

- Assessing the quality and availability of TB recording and reporting includes:
- Checking the completeness of TB case reporting.
- Finding duplicated or misclassified records.
- Analysing the internal and external consistency of data using national and sub national data.

Internal consistency means that data are consistent over time and space while external consistency means that data are consistent with existing evidence about the epidemiology of TB.

- Analysis of the trends in TB recording and reporting data.
- These actions include

    - Assessing how the trends in TB recording and reporting reflect trends in rates of TB incidence.
    - Assessing how the trends in TB recording and reporting reflect changes in other factors. These changes include programmatic efforts to find and treat more cases, laboratory diagnostic methods etc.

- Analysis of whether all TB cases are being captured in official recording and reporting systems. These actions include:

    - Conducting operational research (capture–recapture studies) and collection of supporting evidence. The supporting evidence can include: (i) the knowledge and practices of health-care staff related to definition of TB suspects; (ii) the extent to which regulations about recording of cases are observed and (iii) population access to health services.
    - Assessment of the coverage of recording and reporting data and cross-validation estimates of TB incidence produced by using other methods such as analysing the number of TB deaths recorded in vital registration systems.

Assessing the quality and availability of TB recording and reporting data can be used to identify where and how recording and reporting systems need to be strengthened. It is also crucial to distinguish between changes in TB recording and reporting data that are due to incidence, and changes that are due to other factors, before considering using recording and reporting data to estimate trends in the rates of TB incidence and case detection. By determining the extent to which the change in TB recording and reporting data are influenced by incidence or other factors it is possible to assess if time series TB recording and reporting data is a good proxy for trends in TB incidence, or if the data needs to be adjusted for other factors. Other factors that could influence TB recording and reporting data are programmatic efforts to find and treat more cases and new diagnostic tests and policies.[73]

After determining if the data is a good proxy for trends in TB incidence and if TB recording and reporting data is accurate and complete, it is still not adequate to use TB recording and reporting data to get an absolute estimation of TB incidence. In order to estimate TB incidence it is necessary to determine if all TB cases are being captured in official recording and reporting systems. A framework that was developed to explain the reason for TB cases not recorded in the recording and reporting system is called the onion model (Figure 1-4).[97]

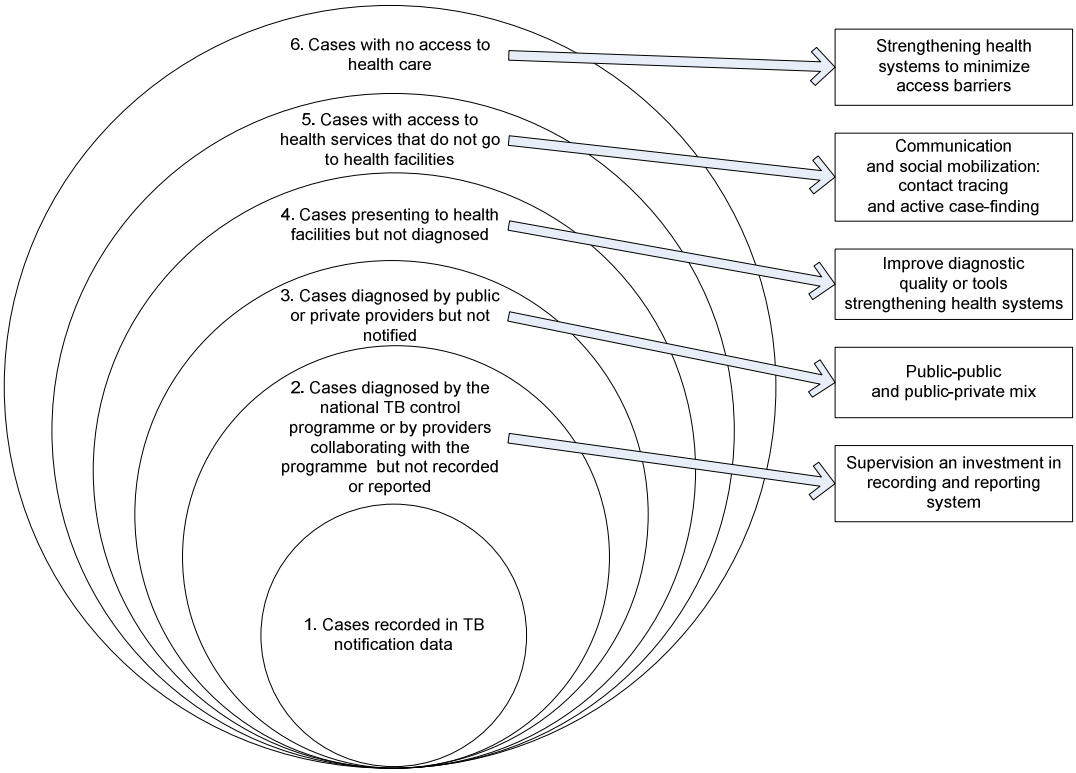

Figure 1-4: The "onion model": A framework for assessing the fraction of TB cases accounted for in TB notification data and how to increase it.[2]

The onion model can be used to explain why TB cases are not accounted for in the recording and reporting system and an assessment can be made to determine how and where the TB programme or health system can be improved to limit the number of missing TB cases.

The onion model was first presented to the international TB community in 2002.[2] In the onion model the cases that are recorded and most likely been initiated on TB treatment can be found in layer 1. Layers 2 to 6 constitutes the overall proportion of TB incident cases being accounted for in TB recording and reporting data.

The reason for TB cases being missed in national recording and reporting data for each layer of the onion model are:

- Layer 2: Cases diagnosed by national TB control programme but not recorded in recording and reporting data.[116,120-126]
- Layer 3: Cases diagnosed by providers not linked to the national TB control programme that are not recorded.[116,120-126]
- Layer 4: Cases presenting to health facilities that are not diagnosed.[127-129]
- Layer 5: Cases that have access to health services but do not seek care.[127-129]
- Layer 6: Cases that do not have access to health services.[130-132]

One method to determine how many cases are missing from the recording and reporting system is to perform an inventory study (record linking), layer 2 of onion model. This involves using a number of registers or data sources to compare with the TB recording and reporting system and to identify TB cases not recorded in the recording and reporting system. These registers or data sources could include hospital registers, HIV notification records with information on TB co-morbidity, laboratory registers and prescriptions from pharmacies. If there are three or more of these registers or data sources available, a capture-recapture study could be performed to estimate the number of cases not captured in any of the data sources.

The following data could also be used to determine the number of TB cases in layers 2 and 3:

- Drug sales in the private sector.
- Health expenditures in private and/or nongovernmental organisations.
- The number and proportion of health facilities and private practitioners that are not collaborating with the NTP.
- The number of TB prescriptions from pharmacies.
- Regulations on prescribing and availability of drugs.

- Knowledge and use of the international standards for TB care.

Layers 4 and 5 are more challenging to determine the number of TB cases. These layers require studies to determine and understand the reasons why TB cases are being missed at health care facilities and why cases are experiencing symptoms, but not seeking care. These studies should include:

- Data on population access to health care.
- Data from surveys of the knowledge, attitudes and practices of staff and the general population.
- Practices for managing people suspected of having TB.
- Data on the number of people who had samples sent for microscopy who are suspected of having TB.

A challenging limitation of the above studies could be to convince providers outside the NTP to participate in such studies.

The South African TB recording and reporting system relates to the onion model as follows. The TB treatment register and ETR fall within layer 1 of the onion model. During the transfer of these completed pages to the district office, pages may go missing. The cases recorded on these pages will therefore not be reported in the ETR. These will fall in layer 2 of the onion model. It is also possible for cases to be missed or incorrectly captured when the data are captured into the ETR.[133,134] During this aggregation TB cases can also be excluded as a result of the missing data or incorrect data entry (layer 2 onion model).[133,134]

The current study therefore proposed to evaluate the accuracy and completeness of the South African reporting and recording system based in part on the WHO Framework for assessment of TB surveillance notification data[2] and the onion model[2]. The accuracy and completeness will be assessed through record linking and capture-recapture methods. Record linking and capture-recapture methods will be discussed in the following chapter.

1.16 REFERENCES TO CHAPTER 1

1. National Department of Health. The South African National Tuberculosis Control Programme: Practical Guidelines. Pretoria: National Department of Health; 2009.
2. World Health Organization. Stop TB Dept. Stop TB policy paper : 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.WHO/HTM/TB/2009.416, 58 pp.
3. Styblo K, Bumgarner JR. Tuberculosis can be controlled with existing technologies: evidence. The Hague: Tuberculosis Surveillance Research Unit.; 1991.
4. Raviglione MC, Uplekar MW. WHO's new Stop TB Strategy. Lancet 2006;367:952-5.
5. US Department of Health and Human Services. Principles of epidemiology. An introduction to applied epidemiology and biostatistics. In: Self-study course 3030-G. 2nd ed. Atlanta: Centre for Disease Control; 1992.
6. Greenberg RS, Daniels SR, Flanders WD, Eley JW, Boring JR. Medical epidemiology. 2nd ed. New York: Lange Medical Books/McGraw-Hill; 2001.
7. Enarson DA, Kennedy SM, Miller DL. Research methods for promotion of lung health. Int J Tuberc Lung Dis 2004;8:915-9.
8. American Thoracic Society. Diagnostic standards and classification of tuberculosis and other mycobacterial diseases (14th edition). Am Rev Respir Dis 1981;123:343-58.
9. Rieder HL. Opportunity for exposure and risk of infection: the fuel for the tuberculosis pandemic. Infection 1995;23:1-3.
10. Rieder HL, ed. International Union against Tuberculosis and Lung Disease. Epidemiologic basis of tuberculosis control. Paris: International Union Against Tuberculosis and Lung Disease; 1999.
11. Rieder HL, ed. International Union against Tuberculosis and Lung Disease. Interventions for tuberculosis control and elimination. Paris: International Union Against Tuberculosis and Lung Disease; 2002.
12. Frieden TR. Lessons from tuberculosis control for public health. Int J Tuberc Lung Dis 2009;13:421-8.
13. Lonnroth K, Holtz TH, Cobelens F, et al. Inclusion of information on risk factors, socio-economic status and health seeking in a tuberculosis prevalence survey. Int J Tuberc Lung Dis 2009;13:171-6.

14. Schaaf HS, Michaelis IA, Richardson M, et al. Adult-to-child transmission of tuberculosis: household or community contact? Int J Tuberc Lung Dis 2003;7:426-31.
15. Commission on Social Determinants of Health. Commission on Social Determinants of Health. Geneva: World Health Organization; 2006.WHO/EIP/EQH/01/2006, 16 pp.
16. Classen CN, Warren R, Richardson M, et al. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. Thorax 1999;54:136-40.
17. Brands A OS, Lonroth K, et al. Reply to Addressing smoking cessation in tuberculosis control. Bull World Health Organ 2007;85:647-8.
18. Claassens MM, Sismanidis C, Lawrence KA, et al. Tuberculosis among community-based health care researchers. Int J Tuberc Lung Dis 2010;14:1576-81.
19. Joshi R, Reingold AL, Menzies D, Pai M. Tuberculosis among health-care workers in low- and middle-income countries: a systematic review. PLoS Med 2006;3:e494.
20. Rouillon A, Perdrizet S, Parrot R. Transmission of tubercle bacilli: The effects of chemotherapy. Tubercle 1976;57:275-99.
21. Dye C, Williams BG. The population dynamics and control of tuberculosis. Science 2010;328:856-61.
22. Chadha VK, Agarwal SP, Kumar P, et al. Annual risk of tuberculous infection in four defined zones of India: a comparative picture. Int J Tuberc Lung Dis 2005;9:569-75.
23. Greenhalgh S, Montgomery M, Segal SJ, Todaro MP, UN-Population-Fund. State of world population 2007: Unleashing the potential of urban growth. Popul Dev Rev 2007;33:639-40.
24. United Nations Development Programme. Chapter 4: Unequal human impacts of environmental damage. In: Human Development Report 1998. New York: Oxford University Press; 1998:66-85.
25. Dye C, Lonnroth K, Jaramillo E, Williams BG, Raviglione M. Trends in tuberculosis incidence and their determinants in 134 countries. Bull World Health Organ 2009;87:683-91.
26. Toward 7 Billion: Why World Population is Still Growing. Population Action International, 2005. (Accessed 29/08/2011, at http://www.populationaction.org/Publications/Fact_Sheets/FS7/Summary.shtml.)
27. Lienhardt C, Sillah J, Fielding K, et al. Risk factors for tuberculosis infection in children in contact with infectious tuberculosis cases in the Gambia, West Africa. Pediatrics 2003;111:e608-14.
28. Lienhardt C, Fielding K, Sillah JS, et al. Investigation of the risk factors for tuberculosis: a case-control study in three countries in West Africa. Int J Epidemiol 2005;34:914-23.

29. Styblo K, Meijer J, Sutherland I. Tuberculosis Surveillance Research Unit Report No. 1: the transmission of tubercle bacilli; its trend in a human population. Bull Int Union Tuberc 1969;42:1-104.
30. National Department of Health. The South African National Tuberculosis Control Programme: Practical Guidelines. Pretoria: National Department of Health; 2004.
31. Marais BJ, Gie RP, Schaaf HS, et al. The natural history of childhood intra-thoracic tuberculosis: a critical review of literature from the pre-chemotherapy era. Int J Tuberc Lung Dis 2004;8:392-402.
32. Loudon RG, Roberts RM. Droplet expulsion from the respiratory tract. Am Rev Respir Dis 1967;95:435-42.
33. Beyt BE, Jr., Ortbals DW, Santa Cruz DJ, Kobayashi GS, Eisen AZ, Medoff G. Cutaneous mycobacteriosis: analysis of 34 cases with a new classification of the disease. Medicine (Baltimore) 1981;60:95-109.
34. Horney DA, Gaither JM, Lauer R, Norins AL, Mathur PN. Cutaneous inoculation tuberculosis secondary to 'jailhouse tattooing'. Arch Dermatol 1985;121:648-50.
35. MacGregor RR. Cutaneous tuberculosis. Clin Dermatol 1995;13:245-55.
36. Sehgal VN. Cutaneous tuberculosis. Dermatol Clin 1994;12:645-53.
37. Frieden TR, Sterling TR, Munsiff SS, Watt CJ, Dye C. Tuberculosis. Lancet 2003;362:887-99.
38. Harries AD, Dye C. Tuberculosis. Ann Trop Med Parasitol 2006;100:415-31.
39. Grzybowski S, Barnett GD, Styblo K. Contacts of cases of active pulmonary tuberculosis. Bull Int Union Tuberc 1975;50:90-106.
40. Shaw JB, Wynn-Williams N. Infectivity of pulmonary tuberculosis in relation to sputum status. Am Rev Tuberc 1954;69:724-32.
41. van Geuns HA, Meijer J, Styblo K. Results of contact examination in Rotterdam, 1967-1969. Bull Int Union Tuberc 1975;50:107-21.
42. Altare F, Durandy A, Lammas D, et al. Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency. Science 1998;280:1432-5.
43. Bellamy R. Interferon-gamma and host susceptibility to tuberculosis. Am J Respir Crit Care Med 2003;167:946-7.
44. Bellamy R. Susceptibility to mycobacterial infections: the importance of host genetics. Genes Immun 2003;4:4-11.

45. Singh SP, Mehra NK, Dingley HB, Pande JN, Vaidya MC. Human leukocyte antigen (HLA)-linked control of susceptibility to pulmonary tuberculosis and association with HLA-DR types. J Infect Dis 1983;148:676-81.
46. Aït-Khaled N, Alarcón E, Armengol R, et al., eds. Managment of tuberculosis. A guide to the essentials of good practice. Sixth ed. Paris: International Union Against Tuberculosis and Lung Disease; 2010.
47. Alisjahbana B, van Crevel R, Sahiratmadja E, et al. Diabetes mellitus is strongly associated with tuberculosis in Indonesia. Int J Tuberc Lung Dis 2006;10:696-700.
48. Corbett EL, Watt CJ, Walker N, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch Intern Med 2003;163:1009-21.
49. Parrish NM, Dick JD, Bishai WR. Mechanisms of latency in Mycobacterium tuberculosis. Trends Microbiol 1998;6:107-12.
50. Corbett EL, Marston B, Churchyard GJ, De Cock KM. Tuberculosis in sub-Saharan Africa: opportunities, challenges, and change in the era of antiretroviral treatment. Lancet 2006;367:926-37.
51. Comstock GW, Livesay VT, Woolpert SF. The prognosis of a positive tuberculin reaction in childhood and adolescence. Am J Epidemiol 1974;99:131-8.
52. Marais BJ, Gie RP, Schaaf HS, et al. The clinical epidemiology of childhood pulmonary tuberculosis: a critical review of literature from the pre-chemotherapy era. Int J Tuberc Lung Dis 2004;8:278-85.
53. Stead WW, Lofgren JP. Does the risk of tuberculosis increase in old age? J Infect Dis 1983;147:951-5.
54. Alcaide J, Altet MN, Plans P, et al. Cigarette smoking as a risk factor for tuberculosis in young adults: a case-control study. Tuber Lung Dis 1996;77:112-6.
55. Gajalakshmi V, Peto R, Kanaka TS, Jha P. Smoking and mortality from tuberculosis and other diseases in India: retrospective study of 43000 adult male deaths and 35000 controls. Lancet 2003;362:507-15.
56. Lin HH, Ezzati M, Murray M. Tobacco smoke, indoor air pollution and tuberculosis: a systematic review and meta-analysis. PLoS Med 2007;4:e20.
57. Maurya V, Vijayan VK, Shah A. Smoking and tuberculosis: an association overlooked. Int J Tuberc Lung Dis 2002;6:942-51.
58. Anderson RH, Sy FS, Thompson S, Addy C. Cigarette smoking and tuberculin skin test conversion among incarcerated adults. Am J Prev Med 1997;13:175-81.

59. Hussain H, Akhtar S, Nanan D. Prevalence of and risk factors associated with Mycobacterium tuberculosis infection in prisoners, North West Frontier Province, Pakistan. Int J Epidemiol 2003;32:794-9.
60. McCurdy SA, Arretz DS, Bates RO. Tuberculin reactivity among California Hispanic migrant farm workers. Am J Ind Med 1997;32:600-5.
61. Plant AJ, Watkins RE, Gushulak B, et al. Predictors of tuberculin reactivity among prospective Vietnamese migrants: the effect of smoking. Epidemiol Infect 2002;128:37-45.
62. Altet MN, Alcaide J, Plans P, et al. Passive smoking and risk of pulmonary tuberculosis in children immediately following infection. A case-control study. Tuber Lung Dis 1996;77:537-44.
63. Ariyothai N, Podhipak A, Akarasewi P, Tornee S, Smithtikarn S, Thongprathum P. Cigarette smoking and its relation to pulmonary tuberculosis in adults. Southeast Asian J Trop Med Public Health 2004;35:219-27.
64. Tipayamongkholgul M, Podhipak A, Chearskul S, Sunakorn P. Factors associated with the development of tuberculosis in BCG immunized children. Southeast Asian J Trop Med Public Health 2005;36:145-50.
65. Cegielski JP, McMurray DN. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. Int J Tuberc Lung Dis 2004;8:286-98.
66. Arnadottir T, Rieder HL, Trebucq A, Waaler HT. Guidelines for conducting tuberculin skin test surveys in high prevalence countries. Tuber Lung Dis 1996;77 Suppl 1:1-19.
67. Rieder HL, Chadha VK, Nagelkerke NJ, van Leth F, van der Werf MJ. Guidelines for conducting tuberculin skin test surveys in high-prevalence countries. Int J Tuberc Lung Dis 2011;15 Suppl 1:S1-25.
68. Dye C, Bassili A, Bierrenbach AL, et al. Measuring tuberculosis burden, trends, and the impact of control programmes. Lancet Infect Dis 2008;8:233-43.
69. Fine PE, Bruce J, Ponnighaus JM, Nkhosa P, Harawa A, Vynnycky E. Tuberculin sensitivity: conversions and reversions in a rural African population. Int J Tuberc Lung Dis 1999;3:962-75.
70. Pai M, Zwerling A, Menzies D. Systematic review: T-cell-based assays for the diagnosis of latent tuberculosis infection: an update. Ann Intern Med 2008;149:177-84.

71. Shanaube K, Hargreaves J, Fielding K, et al. Risk factors associated with positive QuantiFERON-TB Gold In-Tube and tuberculin skin tests results in Zambia and South Africa. PLoS One 2011;6:e18206.
72. Glaziou P, van der Werf MJ, Onozaki I, et al. Tuberculosis prevalence surveys: rationale and cost. Int J Tuberc Lung Dis 2008;12:1003-8.
73. World Health Organization. Global tuberculosis control : epidemiology, planning, financing : WHO report 2009. Geneva: World Health Organization; 2009.WHO/CDS/TB/2003.313, 303 pp.
74. Bai GH, Kim SJ, Lee EK, Lew WJ. Incidence of pulmonary tuberculosis in Korean civil servants: second study, 1992-1994. Int J Tuberc Lung Dis 2001;5:346-53.
75. Kim SJ, Hong YP, Lew WJ, Yang SC, Lee EG. Incidence of pulmonary tuberculosis in Korean civil servants. Tuber Lung Dis 1995;76:534-9.
76. Trends in the prevalence and incidence of tuberculosis in south India. Int J Tuberc Lung Dis 2001;5:142-57.
77. Fifteen year follow up of trial of BCG vaccines in south India for tuberculosis prevention. Tuberculosis Research Centre (ICMR), Chennai. Indian J Med Res 1999;110:56-69.
78. Effectiveness of BCG vaccination in Great Britain in 1978. A report from the Research Committee of the British Thoracic Association. Br J Dis Chest 1980;74:215-27.
79. Randomised controlled trial of single BCG, repeated BCG, or combined BCG and killed Mycobacterium leprae vaccine for prevention of leprosy and tuberculosis in Malawi. Karonga Prevention Trial Group. Lancet 1996;348:17-24.
80. Narayanan PR. Influence of sex, age & nontuberculous infection at intake on the efficacy of BCG: re-analysis of 15-year data from a double-blind randomized control trial in South India. Indian J Med Res 2006;123:119-24.
81. Styblo K, Dankova D, Drapela J, et al. Epidemiological and clinical study of tuberculosis in the district of Kolin, Czechoslovakia. Report for the first 4 years of the study (1961-64). Bull World Health Organ 1967;37:819-74.
82. Rieder H. Annual risk of infection with Mycobacterium tuberculosis. Eur Respir J 2005;25:181-5.
83. Menzies D. Tuberculin surveys--why? Int J Tuberc Lung Dis 1998;2:263-4.
84. Shimao T. [Tuberculosis prevalence survey in Japan]. Kekkaku 2009;84:713-20.
85. Shimao T. Measuring tuberculosis: the role of the Tuberculosis Prevalence Survey as developed in Eastern countries. Tuber Lung Dis 1993;74:293-4.

86. Rieder HL. Methodological issues in the estimation of the tuberculosis problem from tuberculin surveys. Tuber Lung Dis 1995;76:114-21.
87. Mathers CD, Fat DM, Inoue M, Rao C, Lopez AD. Counting the dead and what they died from: an assessment of the global status of cause of death data. Bull World Health Organ 2005;83:171-7.
88. Case findings per province and for the country. Health Systems Trust, 2009. (Accessed 29/08/2011, at http://www.hst.org.za/indicators/TB/TB_2008August.)
89. National Department of Health. Report national HIV and syphilis prevalence survey South Africa 2009. Pretoria: National Department of Health; 2010.
90. UNAIDS/WHO Working Group on Global HIV/AIDS/STI Surveillance., World Health Organization. Dept. of HIV/AIDS., UNAIDS. Guidelines for measuring national HIV prevalence in population-based surveys. Geneva: World Health Organization; 2005.67 pp.
91. World Health Organization. International health regulations (2005). 2nd ed. Geneva: World Health Organization; 2008.74 pp.
92. "Data collection form" Global tuberculosis control - epidemiology, strategy, financing. (Accessed 29/08/2011, at http://www.who.int/entity/tb/publications/global_report/2009/dcf_2008.xls.)
93. World Health Organization. Communicable Diseases Cluster. What is DOTS? : a guide to understanding the WHO-recommended TB control strategy known as DOTS. Geneva: World Health Organization; 1999.WHO/CDS/CPC/TB/99.270, 30 pp.
94. Frieden TR. Can tuberculosis be controlled? Int J Epidemiol 2002;31:894-9.
95. Borgdorff MW, Floyd K, Broekmans JF. Interventions to reduce tuberculosis mortality and transmission in low- and middle-income countries. Bull World Health Organ 2002;80:217-27.
96. Dye C, Garnett GP, Sleeman K, Williams BG. Prospects for worldwide tuberculosis control under the WHO DOTS strategy. Directly observed short-course therapy. Lancet 1998;352:1886-91.
97. Dye C, Watt CJ, Bleed DM, Williams BG. What is the limit to case detection under the DOTS strategy for tuberculosis control? Tuberculosis (Edinb) 2003;83:35-43.
98. Meimanaliev A-S, Prochorskas R, Bullinger M, World Health Organization. Regional Office for Europe. Availability of data in CIS countries on the health-related indicators of the Millennium Development Goals. Copenhagen: WHO Regional Office for Europe; 2006.40 pp.

99. World Health Organization. Dept. of Health Financing and Stewardship. Millennium development goals : the health indicators : scope, definitions and measurement methods. Geneva: World Health Organization; 2003.WHO/EIP/HFS/03.2, 14 pp.
100. World Health Organization. Stop TB Dept. Implementing the WHO Stop TB Strategy : a handbook for national tuberculosis control programmes. Geneva: World Health Organization; 2008.WHO/HTM/TB/2008.401, 184 pp.
101. Cape Town TB Control: Progress Report 1997-2003. Provincial Administration of the Western Cape Metro Region and the City of Cape Town, 2004. (Accessed 29/08/2011, at http://www.capetown.gov.za/en/CityHealth/Documents/Guidelines,%20Specifications/TB%20Progress%20Report%201997%20-%202003.pdf.)
102. Department of Health. TB Indaba: Newsletter of The National TB Control Programme. Pretoria: South African National Department of Health; 2003.
103. Enarson DA, International Union against Tuberculosis and Lung Disease. Management of tuberculosis : a guide for low income countries. 5th ed. Paris: International Union against Tuberculosis and Lung Disease; 2000.53 pp.
104. Ottmani S-E, Scherpbier R, Chaulet P, et al. Respiratory care in primary care services : a survey in 9 countries / edited by: Salah-Eddine Ottmani, ... [et al.]. Geneva: World Health Organization; 2004.WHO/HTM/TB/2004.333, 107 pp.
105. WHO Global Tuberculosis Programme. Treatment of tuberculosis : guidelines for national programmes. 3rd ed. Geneva: World Health Organization; 2003.WHO/CDS/TB/2003.313, 108 pp.
106. Revised international definitions in tuberculosis control. Int J Tuberc Lung Dis 2001;5:213-5.
107. Toman K, Frieden TR, World Health Organization. Toman's tuberculosis : case detection, treatment, and monitoring : questions and answers. 2nd ed. Geneva: World Health Organization; 2004.WHO/HTM/TB/2004.334, 332 pp.
108. Wu ZL, Wang AQ. Diagnostic yield of repeated smear microscopy examinations among patients suspected of pulmonary TB in Shandong province of China. Int J Tuberc Lung Dis 2000;4:1086-7.
109. Edginton ME, Wong ML, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: an intervention to improve patient referrals to district clinics. Int J Tuberc Lung Dis 2006;10:1018-22.

110. Edginton ME, Wong ML, Phofa R, Mahlaba D, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: numbers of patients diagnosed and outcomes of referrals to district clinics. Int J Tuberc Lung Dis 2005;9:398-402.
111. Van Wyk SS, Enarson DA, Beyers N, Lombard C, Hesseling AC. Consulting private health care providers aggravates treatment delay in urban South African tuberculosis patients. Int J Tuberc Lung Dis 2011;15:1-8.
112. Malmborg R, Mann G, Thomson R, Squire SB. Can public-private collaboration promote tuberculosis case detection among the poor and vulnerable? Bull World Health Organ 2006;84:752-8.
113. Baral SC, Karki DK, Newell JN. Causes of stigma and discrimination associated with tuberculosis in Nepal: a qualitative study. BMC Public Health 2007;7:211.
114. Narvaiz de Kantor I, Kim SJ, Frieden TR, et al. Laboratory services in tuberculosis control. Geneva: World Health Organization; 1998.WHO/TB/98.258, 47 pp.
115. Botha E, den Boon S, Lawrence KA, et al. From suspect to patient: tuberculosis diagnosis and treatment initiation in health facilities in South Africa. Int J Tuberc Lung Dis 2008;12:936-41.
116. Botha E, Den Boon S, Verver S, et al. Initial default from tuberculosis treatment: how often does it happen and what are the reasons? Int J Tuberc Lung Dis 2008;12:820-3.
117. Castro KG. Tuberculosis surveillance: data for decision-making. Clin Infect Dis 2007;44:1268-70.
118. Attaran A. An immeasurable crisis? A criticism of the millennium development goals and why they cannot be measured. PLoS Med 2005;2:e318.
119. van der Werf MJ, Borgdorff MW. Targets for tuberculosis control: how confident can we be about the data? Bull World Health Organ 2007;85:370-6.
120. Migliori GB, Spanevello A, Ballardini L, et al. 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.
121. Maung M, Kluge H, Aye T, et al. Private GPs contribute to TB control in Myanmar: evaluation of a PPM initiative in Mandalay Division. Int J Tuberc Lung Dis 2006;10:982-7.
122. Lonnroth K, Lambregts K, Nhien DT, Quy HT, Diwan VK. Private pharmacies and tuberculosis control: a survey of case detection skills and reported anti-tuberculosis drug dispensing in private pharmacies in Ho Chi Minh City, Vietnam. Int J Tuberc Lung Dis 2000;4:1052-9.

123. Lonnroth K, Thuong LM, Lambregts K, Quy HT, Diwan VK. Private tuberculosis care provision associated with poor treatment outcome: comparative study of a semi-private lung clinic and the NTP in two urban districts in Ho Chi Minh City, Vietnam. National Tuberculosis Programme. Int J Tuberc Lung Dis 2003;7:165-71.
124. Ambe G, Lonnroth K, Dholakia Y, et al. Every provider counts: effect of a comprehensive public-private mix approach for TB control in a large metropolitan area in India. Int J Tuberc Lung Dis 2005;9:562-8.
125. Arora VK, Lonnroth K, Sarin R. Improved case detection of tuberculosis through a public-private partnership. Indian J Chest Dis Allied Sci 2004;46:133-6.
126. Dewan PK, Lal SS, Lonnroth K, et al. Improving tuberculosis control through public-private collaboration in India: literature review. BMJ 2006;332:574-8.
127. Gasana M, Vandebriel G, Kabanda G, et al. Integrating tuberculosis and HIV care in rural Rwanda. Int J Tuberc Lung Dis 2008;12:39-43.
128. Espinal MA, Reingold AL, Koenig E, Lavandera M, Sanchez S. Screening for active tuberculosis in HIV testing centre. Lancet 1995;345:890-3.
129. Lee MS, Leung CC, Kam KM, et al. Early and late tuberculosis risks among close contacts in Hong Kong. Int J Tuberc Lung Dis 2008;12:281-7.
130. van Hest NA, Smit F, Baars HW, et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? Epidemiol Infect 2007;135:1021-9.
131. Baussano I, Bugiani M, Gregori D, et al. Undetected burden of tuberculosis in a low-prevalence area. Int J Tuberc Lung Dis 2006;10:415-21.
132. Crofts JP, Pebody R, Grant A, Watson JM, Abubakar I. Estimating tuberculosis case mortality in England and Wales, 2001-2002. Int J Tuberc Lung Dis 2008;12:308-13.
133. Validation of the TB indicators in Umzinyathi District. KwaZulu-Natal Department of Health, Unit of Epidemiology, 2005. (Accessed 29/08/2011, at http://www.kznhealth.gov.za/epibulletin12.pdf.)
134. Naidoo S. Evaluation of the TB control programme in Umzinyathi District. In: Unit E, ed. Pietermaritzburg: KwaZulu-Natal Department of Health; 2006.

CHAPTER 2 : REVIEW OF METHODOLOGICAL ASPECTS

2.1 INTRODUCTION

The frequency of disease conditions as well as their causes among different groups of people are often constrained by uncertainty regarding the validity of data collected by population registers.[1,2] This uncertainty includes information on classification errors, i.e. the absence of true cases and the presence of false-positive cases in registrations.[3-5] Before considering the suitability of routine data for research purposes, the quality of the data and completeness of case recording need to be assessed.

Accuracy in the context of assessing data quality in information systems is defined as the proportion of data fields in the information system with a given characteristic, which truly have this attribute.[6] The individual record will often contain several data fields apart from the one by which the person is identified. For example, these data fields may contain results of certain diagnostic tests, diagnoses, age, gender and demographic data. The completeness of data relates to how many data fields have missing data and is thus a subcategory of accuracy of data.

The accuracy and completeness of case recording, or in other words, the number of individuals with a certain condition, can be determined by counting every single person or event. This approach is directly attempted in a census or indirectly through estimating prevalence or incidence by means of surveys of a sub-sample of the population (active case-finding) or by notification (passive case-finding). Other examples of indirect ascertainment of the number of cases are pharmaco-epidemiological studies and record-linking.[7,8] Another indirect technique that estimates completeness of case recording used in epidemiological studies is capture-recapture analysis as described in section 3.4.[9,10]

2.2 RECORD LINKING

Record linking, also called inventory studies, is a valuable tool for use in healthcare research, epidemiology, and “business intelligence”. Often research questions need to be answered by making use of data residing in different databases. Record linking provides an opportunity to effectively link these databases to answer a proposed research question.[11-15] Record linking can also be used to increase the accuracy and completeness of datasets. In theory any database can be

used for record linking as long as there is an adequate identifier to link records, but this is not always the case in practice.[14,15]

The idea of record linking is not new. It was first used by Jenner[16] in research on cow pox and smallpox vaccination. Jenner produced a record system in the late 18th century to link cows with human beings. Jenner proposed that material taken from cows and injected into humans produced protection against smallpox. The first time that record linking appeared in published literature was in 1946 in an article by Dunn.[17]

There are two approaches to record linking namely, (i) within-data-source linking and (ii) between-data-source linking.[13,18,19] Within-data-source linking will usually be done to identify duplicate records within a data source. Between-data-source linking requires two or more registers with a common unique identifier or a combination of identifiers to link on. Record linking can be accomplished through a manual process or through computer record linking systems. The manual process is very time consuming and error prone, especially if the number of records needing to be linked is very large. Computer-based linking systems are therefore preferred to the manual process but some insight in using these computer-based systems is however indispensable. Computer-based systems for record-linking could add extra cost to the study but the benefits in time and accuracy far exceed the cost of computer equipment and software when compared to the manual process.

The most effective method of linking records is by means of unique identification numbers such as a unique national identification number or unique patient number on which to base the record link. If such a unique identifier was available the records could be directly matched. Such unique identifiers in datasets are not widely available and, even if present, they have some limitations. These unique identification numbers suffer from the same limitations as any other routinely collected data, namely the lack of accuracy and completeness.[20] Clerical errors occur when identification numbers are not entered correctly or not be entered at all. Some individuals may provide the wrong identification numbers or use somebody else's identification number. Such records cannot be matched or will be incorrectly matched.

In the absence of a unique identification number, there are two commonly used alternative methods for linking records, namely deterministic and probabilistic record linking.[13,18,20-26] With these methods a combined identifier is used to match records. This combined identifier could consist of data fields in which first name, surname, date of birth, gender, etc. are recorded. The

selection of the data fields to be used as unique identifier is critical as data fields with many errors or which are incomplete will hamper the linking process.[20,21,25,27]

When determining the selection of data fields to use as unique identifiers, error rate and discrimination should be kept in mind.[28,29] These two concepts together influence the number of errors that could be made during the linking process. The discriminating power of a unique identifier directly relates to the probability that two different persons have the same unique identifier. A low discriminating power will result in a high number of false matches (false positives) whereas a high error rate within the data will lead to a high number of matches being missed (false negatives).

When data fields with first name and/or surname are used as unique identifiers, the misspelling of names could limit the effectiveness of the matching process. Phonetic indexes or coding systems such as Soundex or the New York State Identification and Intelligence System (NYSIIS)[30] can be used to limit the effect of misspelling.[14,27,31] The basic concept of phonetic coding is to code names based on the way they sound and not how they are spelled. This conversion reduces the possibility that slight misspelling in names or surnames cause records not to be matched.

Howard Newcombe's[32] early work in record linking approaches are considered to have led to computerised approaches in record linking in the absence of a unique identifier. He observed that the frequency of occurrence of a characteristic, such as surname, among matches and non-matches could be used to compute a score or matching weight associated with the linking of two records. Links are pairs formed between records and do not necessarily refer to the same individual or entity, while a match is a pair formed that refers to the same individual or entity. It is not always possible to consider a pair as a match or as a non-match, and therefore a third category, a possible match, was introduced. Newcombe also observed that matching weights over different variables, such as age, surname, and first name, should be computed and added to obtain an overall matching weight. Pairs with higher values would be considered matches. Fellegi & Sunter[33] built on Newcombe's intuition and introduced the mathematical and statistical foundation for probabilistic record linking, which is extensively used to the present.

With deterministic record linking records are matched between datasets on the agreement of specific identifiers.[13,33] In contrast, probabilistic record linking uses probabilities to match records. Deterministic record linking requires that the datasets to be matched contains a unique identifier or a combination of unique identifiers. The greatest limitation of the deterministic method is that

possible matches may be missed, therefore increasing the possibility of missing matches.[34] The first mention of probabilistic record linking was in 1959 by Newcombe et al.[32] During this study, records of individuals were matched in order to determine their cause of death and the impact of low level radiation on fertility and genetic defects.

The probabilistic record linking approach creates links between-data-sources based on statistical probability calculated on the combined common identifiers. The probability is used to determine the likelihood that a pair of records refers to the same individual. Probabilistic linking maximises matches but may result in uncertainty for some potential links. This will reduce the possibility of missing matched but may increase the necessity to perform manual review of large numbers of possible matches in order to decide if a match is valid or not.

Computer technology has evolved exponentially since the late 1960s. Low computer cost of high performance computer equipment has boosted the development of software suitable for record linking. This has made record linking with the use of computers available to more researchers. The software developed by Matthews Jaro in 1985[35], called Automatch©, has been used extensively in the United States. It has been used by individuals and researchers and also by the Census Bureau. The software GRLS (Generalised Record Linking System) was developed at Statistics Canada in Canada[36]. This system has been used extensively in Canada in the linking of national data systems, especially to monitor health[37]. For many researchers though, the use of the commercial software such as Automatch© is not viable due to the high cost of its private licensing. Some freely accessible software has also become available such as Registry Plus™ Link Plus[38], developed by the Centers for Disease Control and Prevention, Division of Cancer Prevention and Control. It has been developed specifically for linking cancer registries, but can be used in other settings as well.[39]

Previous studies have shown that the selection of a record linking method relies on the type of data fields available for the linking process. In many situations a combination of the record linking methods are required in order for the best record linking result. In Brazil probabilistic record linking on the Brazilian Information System for Tuberculosis Notification for the period 2000 to 2004 was used to identify duplicate records.[40] By removing duplicate records the TB incidence were reduced by 6.1% in the year 2000, 8.3% in 2001, 9.4% in 2002, 9.2% in 2003 and 8.4% in 2004.

A study was conducted in Cleveland, United States, on birth outcomes using data from the comprehensive maternity and infant care project (MIC).[41] The researchers first linked the MIC

data with a hospital database using deterministic record linking in order to add additional personal identification information to the MIC data. This new database was then linked to live birth and still birth certificates using a combination of deterministic and probabilistic record linking. The researchers however decided not to continue with the study as they considered the 85% linking success to be inadequate to have a successful research outcome. This study showed that the outcome of a record linking study is not always successful, and careful consideration should be made before taking on such an endeavour.

Three registers of tuberculosis cases in The Netherlands in 1998 were examined using manual record linking.[42] The aim of the study was to determine the completeness of TB case notification in The Netherlands in 1998. The study indicated that there was an initial under-reporting of 13.4% of TB cases. After correcting for false-positive cases, the under-reporting was 7.3%.

At least two publications have reported on the completeness of case registration in South Africa.[43,44] A study published in 2005 and conducted at Chris Hani Baragwanath Hospital in Johannesburg indicated that 21% of the patients referred to an outside clinic for treatment by the hospital did not attend any clinic in Soweto and were also not recorded in the TB treatment register.[43] Of the TB patients identified in this study 25% died, and of those who died 72% died in the hospital. The 21% of patients who were not recorded in the TB treatment register were followed-up in the community in order to determine the reason for non-registration. Of these patients only 37 (43%) were found for a follow-up interview. The majority of these individuals stated that they did not know that they had to take further treatment and 12 individuals were too sick to attend a clinic. Due to these findings a TB Care Centre were established on the hospital premises. The aim of this intervention was 1) to ensure adequate referral of TB patients to ambulatory treatment facilities (district clinics); 2) to improve patient education; and 3) to improve the registration of diagnosed patients, including those who died in the hospital. A follow-up study conducted in order to evaluate the intervention found that after the intervention, 93% of TB cases referred to a clinic actually attended a clinic.[45]

A study conducted between 2004 and 2005 at 13 primary health care facilities in the Stellenbosch district of the Western Cape indicated that 17% of TB cases with two smear positive results were not recorded in the TB treatment register.[44] These patients accessed health care facilities and were diagnosed but never started treatment and were referred to as ‘initial defaulters’. A follow-up study was conducted in 11 of the 13 primary health care facilities in order to determine what happened to these ‘initial defaulters’.[46] Of the 58 initial defaulters 14 (24%) had died. Of the 44 remaining

initial defaulters, 26 could not be found due to address-related reasons. The remaining 18 initial defaulters (41%) were found and interviewed. Reported reasons for not starting treatment were directly linked to services in 56% of cases and not directly linked to service in 44% of cases.

Record linking in its own right can provide a much clearer picture of the actual numbers of correctly classified patients in the database. The next question that should be answered is an estimation of the number of cases not captured by any of the databases in the study. For this a capture-recapture analysis is usually conducted. The capture-recapture analysis is usually conducted after record linking studies in which the accuracy and completeness of case registration have been assessed.

2.3 DESCRIPTION OF RECORD LINKING AS CARRIED OUT IN THIS STUDY

The approach followed in the present study was to use probabilistic record linking with manual review using Link Plus (Annex 1). Once records were matched within and between-data-sources using Link Plus, all matches, non-matches and probable matches were manually reviewed for correctness. Uncertain matches were reviewed with the help of experienced nurses who work in the clinics and know the patients. The data fields used for linking purposes were name, surname, age and address. Name and surnames were converted with the NYSIIS phonetic coding system.

The following steps were followed:

- Determine the accuracy and completeness of linking fields
- Detect duplicates and conduct within-data-source linking
- Do between-data-source linking

The accuracy and completeness of the linking fields were firstly assessed to determine if linking would be possible. This should always be done for any study where record linking will be used. Linking fields were then formatted to be consistent within and between-data-sources as follows.

- Name and surname were checked to eliminate any obvious mistakes such as name and surname been swapped or obvious spelling mistakes. Any unnecessary characters such as "-", "/" etc. were removed.
- Gender was formatted as F = Female, M = Male and U = Unknown throughout all data sources.

- For the NHLS data sources a new variable for age was generated. This was done in order to be better comparable with the TB register's age field. This field was calculated as the age of the individual at the time the specimen was collected; the difference in years between the date the specimen was received in the laboratory and the date of birth. Where date of birth was not available and the age was recorded, the age was copied to the new field. It would have been preferable to use date of birth instead of age, but date of birth was not available in the TB register.
- It was decided not to use address as a primary linking field but rather to confirm uncertain links through manual review. The reason for this was because the addresses were written very inconsistently and were often absent in the NHLS data source.

The following linking criteria were followed in order to identify links:

- Perfect match: all identifiers agree.
- Probable match: 3 out of 4 identifiers agree.
- Possible match: 2 out of 4 identifiers agree.

In all matches if a name or address can be logically identified as a match the match was accepted. A plus/minus two-year difference in age was considered a match. For probable and possible matches the match was accepted if the local nurse was able to testify that it was the same individual.

In order to detect duplicates in the TB treatment register, LinkPlus was configured for "de-duplication" on the selected linking fields. The strategy in Table 2-1 was used to identify if a treatment episode in the TB register was a duplicate.

Duplicates in the NHLS data sources were less of a concern as all specimens received in the laboratory is assigned a unique laboratory number. This is done automatically with the laboratory registration system. LinkPlus was configured for "de-duplicating" for each of the two laboratory data sources separately. This was done in order to link all specimens belonging to the same individual together. The same unique identifier was then assigned to all specimens belonging to the same person.

Table 2-1: Assignment of unique identifiers for multiple TB episodes.

Previous treatment outcome*	***Assignment of unique identifiers to subsequent treatment episode***
Cure: Patient who is smear-negative at, or one month prior, to the completion of treatment and also on at least one previous occasion.	New identifier
Treatment completed: Patient who has completed treatment but without proof of cure, smear results are not available on at least two occasions prior to the completion of treatment.	
Died: Patient who dies for any reason during the course of TB treatment.	Error in register
Treatment failure: Patient who remains or is again smear-positive at five months after starting treatment.	Link to previous episode
Treatment interrupted: Patient whose treatment was interrupted for more than two consecutive months before the end of the treatment period.	
Transfer out: Patient who has been transferred to another reporting unit (e.g. district) and for whom the treatment outcome is not known.	
Moved: Patient who is moved to another facility within the same district.	
* At the end of treatment, each patient is assigned one of the above mutually exclusive treatment outcomes.[47]	

A register was generated for each data source with a minimum set of fields as shown below:

Table 2-2: New structure of registers created after with-in data source linking.

TB register unique identifier	Centralised laboratory unique identifier	Hospital laboratory unique identifier	Name	Surname	Gender	Age	Address	Patient Type

Between-data-source linking was then performed between the newly created registers as follows: TB treatment register and centralised laboratory, TB treatment register and hospital laboratory and centralised laboratory and hospital laboratory. LinkPlus was configured for linking between-data-sources and each register was linked to the other. If a link between registers was identified the unique identifier of each linking register was copied to the corresponding unique identifier field for the other registers.

After linking the NHLS registers to the TB treatment register, each data point was explored to see if the results from the NHLS data sources belonged to the specific treatment episode in the TB treatment register. If the date that the specimen was registered in the laboratory fell within two months before the start of treatment date and two months after the end of treatment in the TB treatment register the result of the specimen was considered to belong to the treatment episode. These results in the NHLS data sources were then assigned the same TB treatment register unique identifier that was assigned to the corresponding NHLS register from the between-data-source linking.

The TB cases were then classified in all three registers according to the number of smear and culture results found, in their respective data sources, to be positive. The following classifications were used for Patient Type:

- One smear positive
- Two smears positive
- Only culture positive

- One or more smear positive and culture positive

From the linking process a Venn diagram was generated as represented in Figure 2-1, based on the following criteria:

- a = TB cases recorded in the TB treatment register
- b = TB cases recorded in the TB treatment register and Centralised laboratory register
- c = TB cases recorded in all three registers
- d = TB cases recorded in the TB treatment register and Hospital register
- e = TB cases recorded in the both laboratory registers
- f = TB cases recorded in the Centralised register
- g = TB cases recorded in the Hospital register
- N = overall number of TB cases in the two communities.

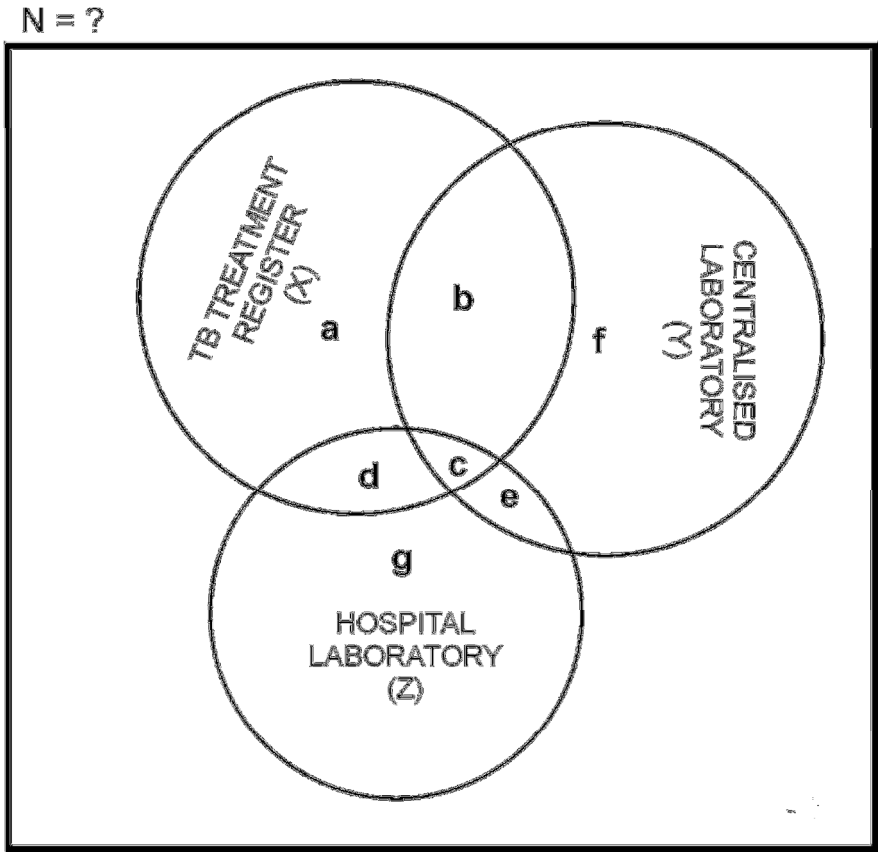

Figure 2-1: Venn diagrammes representing the distribution of TB cases in each data source after record linking.

2.3.1 Accuracy and completeness of data recorded

Before record linking

It was determined whether the data in each data field were accurately recorded and whether any data were missing.

After record linking

Accuracy and completeness of data were assessed after linking results from the laboratories to the TB treatment registers. All data fields available in all the TB treatment registers and the laboratory databases were compared. Names and surnames were compared between TB treatment registers and the laboratory database to determine how many differed to such extent that it would hamper tracing the result back to a TB case. Ages were compared to determine if any differed by more than two years between-data-sources.

For discrepancies regarding demographic information, timing of results (diagnostic or follow-up), and patient category (new or re-treatment), the TB treatment register was regarded as the most accurate.

For discrepancies regarding sample results, the laboratory data source was considered as the most accurate source. If there was a discrepancy in the sample result between the TB treatment register and the laboratory data source, the case recorded in the TB treatment register was re-evaluated to determine if any TB case already recorded in the TB treatment register as a non-bacteriological case, should be reclassified to a bacteriologically confirmed case, based on the corrected results.

For assessment of completeness of the TB treatment register, all laboratory sputum results that were collected within three months prior to and six weeks after the date of recording the TB case in the TB treatment register were included in the analysis. The case classification in the TB treatment register was re-assessed to determine if any case classified as extra-pulmonary TB or non-bacteriological TB could be reclassified as pulmonary TB or bacteriologically confirmed TB based on positive results in the laboratory database which were not transcribed into the TB treatment register.

2.3.2 Accuracy and completeness of case recording

Before record linking

Determination of the accuracy of case recording, before record linking, was possible only for the TB treatment registers. Accuracy of case recording could only be determined for pulmonary disease, and was conducted based on criteria from the National TB Guidelines.[47] All cases with smear-positive sputum results were assessed to determine if these were recorded as pulmonary TB in the TB treatment register.

After record linking

For estimation of accuracy of case recording, the results in the TB treatment registers were compared to the linked results in the laboratory data source to determine if results were recorded correctly in the TB treatment register. Completeness of case recording was determined by counting how many TB cases in the centralised and hospital laboratory data sources were not recorded in the TB treatment register within two months of receiving the specimen in the laboratory.

2.4 CAPTURE-RECAPTURE

Capture-recapture was originally developed during ecological studies for counting wildlife (e.g. birds, polar bears and wild salmon)[9] and later adapted for epidemiological purposes. In animal research the methods are commonly known by the usual name of capture-recapture, mark-recapture or tag-recapture. In human research and record linking studies capture-recapture is also commonly known as dual-system methods or dual-record system methods, Petersen estimators[48], multiple-record system methods, or Bernoulli census estimates and ascertainment corrected rates.[49,50]

The two-source capture-recapture model is the simplest model available for estimating the size of the portion of a population for whom information is unobtainable or missing. The first capture provides the individuals for marking or tagging and is returned to the general population. The second capture provides the individuals recaptured. The numbers of these captured and recaptured individuals can then be used to estimate the number of individuals captured in neither of the first or second capture and therefore estimating the total population size.[9]

The two-source capture-recapture method is generally called the Petersen method due to Petersen's work in 1894 with tagged fish. The first time the method was used in fisheries was in 1917 by

Dahl.[9] Other early ecological studies include those done by Lincoln in 1930 to estimate the size of a duck population[51]. The earliest capture-recapture studies in human populations were conducted by Sekar and Deming[52] in 1949 to estimate birth and death rates and the extent of registration. The two-source capture-recapture method has also been used to determine the completeness of census data by acquiring an additional sample in addition to the census. The capture-recapture method was then used to determine the census undercount.[53]

The capture-recapture method can be applied to any situation in which there are two incomplete lists. This could be accomplished by replacing the concept “caught in sample i” by “being present in list i”. In epidemiology many such 'lists' are available such as hospital records, medical practice files, medicine prescriptions, laboratory registers etc. Due to the incompleteness of such lists, the object would then be to estimate the number missing from both lists. Some of the earliest studies conducted in the field of epidemiology were by Wittes.[54,55] Wittes described the use of capture-recapture to determine the completeness of notifications in the source lists and the total population based on the capture-recapture estimation.

A two-source capture-recapture problem with registers A and B can be tabulated as follows:[9,10]

Table 2-3: The elements of a two-source capture-recapture application.

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

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

number of cases not observed in both registers, $\hat{n}_{00}$. The estimated total number of cases, $\hat{N}$, is the observed number of cases plus the estimated unobserved number of cases. When the basic requirements hold, $\hat{n}_{00}$ can be expressed as

$$\hat{n}_{00} = n_{10} \times n_{01}/n_{11} \tag{3.1}$$

and $\hat{N}$ as

$$\hat{N} = N_A \times N_B/n_{11} \tag{3.2}$$

Equation (3.2) is known as the Petersen estimator. Approximately unbiased estimates of $\hat{N}$ are expected when the registers are large. The correction for bias caused by small registers, the Nearly Unbiased Estimator proposed[56,57] by Chapman, can be expressed as

$$\hat{N} = \left[\frac{(N_A + 1)(N_B + 1)}{n_{11} + 1}\right] - 1 \tag{3.3}$$

The confidence interval is calculated as $\hat{N} \pm 1.96$ times the standard error.

The limitations of the two-sample model soon stimulated interest in an approach incorporating more than two databases or lists. In 1938 Schnabel[58] introduced the K-sample capture-recapture method. In his method more than two samples were used to estimate the population of fish in a lake. The K-sample method was further developed in 1950 by Chapman and Darroch.[59] Using essentially the same assumptions as in the two-sample case; however, it was recognised that some of the underlying assumptions may not hold, especially those related to dependence and heterogeneity.

Fienberg's[60] approached the interdependence among lists or captures through the use of the log-linear model. Many other models have since been developed. All these models however require that samples are independent.[61,62] As this is seldom the case with patient entries on lists, it is unlikely that these general models will be directly useful in epidemiology. A general log-linear model allows for the representation and incorporation of most of these models for K lists, as well as some extensions for the generalisation from closed to open populations. The log-linear model should still be used with caution as one still needs some requirements to hold for the model to be useful.

The requirements for a capture-recapture estimate to be valid can be spelled out in a number of ways, but the key ingredients are the following:

- There must not be any change to the population during the investigation (the population is closed).
- There should be no loss of tags (individuals can be matched from capture to recapture).
- For each sample, each individual should have the same chance of being included in the sample.
- The two samples must be independent.

Assumption 4 actually follows from 3 since the latter implies that marked and unmarked cases have the same probability of being caught in the second sample so that capture in the first sample does not affect capture in the second sample, i.e. samples are independent.[9,10]

Some of the requirements can reasonably easily be dealt with. In epidemiological studies the study design can be structured in such a way to limit the effect of requirement 1 on the capture-recapture estimate by selecting the time period and study population very carefully beforehand.[10,63,64] Requirement 2 depends to a large extent on the success of the record linking conducted on the lists. Record linking was discussed in detail in the Record Linking section (Chapter 2.2), but in principle it depends on the availability of adequate information to identify individuals between lists.[10,63,64] Requirement 3 and 4 are more difficult to adhere to in capture-recapture studies. With reference to requirement 3 subgroups may be present in the population. These subgroups may have a different probability of being captured and recaptured e.g. age, location of disease and infectiousness can cause different probabilities of being observed in a register. Independence between lists is also a problem due to the fact that, for example, if doctors refer their patients exclusively to certain hospitals, the doctor's records and hospital admissions will not provide two independent lists. Requirements 3 and 4 are discussed in detail by Hook and Regal[65] and Darroch et al.[66]

In two-source capture-recapture analysis the last requirement is crucial because it is impossible to check independence mathematically. Judging whether the requirement is met relies on the users to make the plausibility judgement. Dependencies can cause under-estimation (in case of positive dependence) or overestimation (in case of negative dependence).[67] Heterogeneity of the population and violation of the perfect record-linking and closed population requirements can also cause bias in either direction.[10,67,68] These biases are difficult to measure mathematically and can vary considerably between populations.[69] In addition to the requirements mentioned, it is important that

the various registers only contain individuals with the condition under study, i.e. the registers should not include false-positive records. In other words, the specificity and positive predictive value of the registers should ideally be 100%. A low positive predictive value results in overestimation of the true population size.[28] Finally, the individuals under study should be captured within the time and space defined by the investigation.

There are various strategies available to reduce the impact of violations of the requirements underlying human capture-recapture studies.[10] Complete and good quality information on the personal identifiers of individuals in the different registers will limit the impact of violation of the perfect record-linking requirement. Collection of data within a short period of time will minimise the effects of violation of the closed population requirement. Violation of the homogeneity requirement can be handled by stratification of the population into more homogeneous strata, performing capture-recapture analysis for each of the distinct subgroups and subsequently adding the results for the total estimate. An alternative is to include covariates with a strong relationship to the probability of capture in a log-linear covariate capture-recapture model.[65,70] Violation of the independency requirement can be partially identified and controlled when more than two sources are linked, allowing for sources to be examined pair-wise, i.e. two at a time.[55] In the absence of source dependence, the possible pair-wise capture-recapture estimates of the total number of cases should be reasonably similar. Positive dependence between two of the lists can be suspected when one of the pair-wise estimate is considerably lower than the other pair-wise estimates.

With more than two data sources the capture-recapture analysis is more complex due to the fact that multiple estimates are possible. In the three-source capture-recapture approach according to Fienberg[60], pair-wise dependencies can be incorporated in the log-linear model as interactions (models). In the log-linear approach, if k lists are used then there are at most $k - 1$ models. Therefore with three sources there will be eight possible models, excluding the three way interactions. With four sources there will be 113 models plus one four-way interaction In Table 2-4, there are four general types of models:[10]

- Independent model (assume all sources are independent).
- Models that are equal to two independent sources (full two source models); models 1-2, 1-3, and 2-3.
- The model that assumes all possible interactions (saturated model as this model has no degrees of freedom); model 8.
- Other models (there are no other models with three sources).

Additional types of estimates are also available (bottom of Table 2-4). If three sources are available, then there are three different two-source estimates if one of the sources is ignored with each estimation (restricted two-source estimates).

Table 2-4: Three source model.[10]

		Source 1			
		Yes		No	
		Source 2		Source 2	
Source 3		Yes	No	Yes	No
Yes		a	b	e	f
No		c	d	g	x

$N_{obs} = a + b + c + d + e + f + g$

$N_1 = a + b + c + d$

$N_2 = a + c + e + g$

$N_3 = a + b + e + f$

Maximum likelihood estimates of *x* using alternative models		
d.f.[a]	Model	Estimator[¥]
3	Independent	$\hat{x} = \hat{N} - N_{obs}$ $\left(\hat{N} = is\ the\ solution\ of\ (\hat{N} - N_1)(\hat{N} - N_2)(\hat{N} - N_3) = \hat{N}^2(\hat{N} - N)\right)$
2	1 – 2 interactions	$\hat{x} = (c + d + g)(f)/(a + b + e)$
2	1 – 3 interactions	$\hat{x} = (c + d + f)(g)/(a + c + e)$
2	2 – 3 interactions	$\hat{x} = (e + f + g)(d)/(a + b + c)$
1	1 – 2, 1 – 3 interactions	$\hat{x} = gf/a$
1	1 – 2, 2 – 3 interactions	$\hat{x} = gd/c$
1	1 – 3, 2 – 3 interactions	$\hat{x} = df/d$
0	1 – 2, 1 – 3, 2 – 3 interactions	$\hat{x} = (adfg)/(bce)$

Two source restricted estimators of same population			*Continued*
	$\hat{N}_{MLE}$	$\hat{N}_{NUE}$	
1 versus 2	$(N_1)(N_2)/(a+c)$	$(N_1)(N_2)/(a+c+1)$	
1 versus 3	$(N_1)(N_3)/(a+d)$	$(N_1)(N_3)/(a+b+1)$	
2 versus 3	$(N_2)(N_3)/(a+e)$	$(N_2)(N_3)/(a+e+1)$	

*d.f., degrees of freedom; MLE, maximum likelihood estimator; NUE, nearly unbiased estimator.

¥Where not given explicitly, $\hat{N} = \hat{x} + N_{obs}$

In order to select the most appropriate estimate a non-Bayesian method could be used to select the estimate that fits the model best.[10] This method is based on the value of the likelihood ratio statistic, G^2.

For any model *i* and observed data set,

$$G^2 = -2\sum Obs_j \log\left(\frac{Obj_j}{Exp_{ji}}\right), \quad (3.4)$$

Where Obs_j is the observed number in each cell, *j*, and Exp_{ji} is the expected number in cell *j* under the model *i*. The lower G^2, the better the model fit. It is usually best to select the least complex model that fits best i.e. the least saturated or the most parsimonious.[10]

Alternative methods for selecting the most appropriate estimate can be based on information cirteria of which two are widely used.[10] The first is known as the Akaike Information Criterion (AIC)[71] proposed by Akaike and second is the Bayesian Information Criterion (BIC).[72] In the cases of log-linear models the formulas for two criteria for any model *i* are:

$$AIC = G^2 - 2(d.f.) \quad (3.5)$$

$$BIC = G^2 - \left(\log\frac{N_{obs}}{2\pi}\right)(d.f.), \quad (3.6)$$

where degrees of freedom (df) is the number of degrees of freedom of the model and N_{obs} is the number of observed cases. The best estimate is the estimate of the model with the lowest value of the associated information criterion. As the complexity of the models increase, the estimates tends to become more unreliable and the confidence intervals around the estimate becomes wider.[9,10] Both the AIC and BIC penalises complex models more heavily. In general, in the log-linear

capture-recapture estimation procedure the least complex (the least saturated model), whose fit appears adequate, is preferred.[73]

Most of the limitations of capture-recapture analysis pertain to the possibility (often almost a certainty), of violation of the underlying requirements.[10] In contrast to animal population studies the requirements as outlined above are unlikely to be satisfied in epidemiological applications.[64,74] In epidemiological studies capture-recapture analysis often uses existing administrative registers that are not designed for capture-recapture analysis, instead of random surveys of the population according to a common protocol.[75] When deciding to employ routine data registers, the accuracy of the registers regarding such aspects as correct diagnosis and coding and sufficient information for appropriate record-linking, is important. In capture-recapture analysis, errors are highly likely to have a more than additive effect on estimates. Registers containing poor quality data lead to poor capture-recapture outcomes.[10,75]

Dependence of sources is often a problem in epidemiological capture-recapture applications. Such dependence can result from co-operation between the agencies that maintain the different registers, exchange of information or a more or less predictable flow of patients along various institutions due to referral arrangements. The probability of ascertainment by any particular source should be equal but in epidemiological settings often it is not, due to the intrinsic nature of human variation, e.g. socioeconomic differences or variation of severity of disease. Human populations furthermore are rarely closed.[64]

It has been argued that estimates from capture-recapture studies in epidemiology are wholly unreliable unless supported by a wide variety of sensitivity analyses, and by careful medical and social discussion of variability between individuals, and the reasons why a particular individual may fail to be recorded in a particular register.[64] Although many apparently successful capture-recapture studies have been published, only few have been reported to have failed and confidence in the validity of capture-recapture results may reflect publication bias in favour of successful capture-recapture studies rather than the inherent strength of this methodology.[76]

Numerous capture-recapture studies have been used in the field of infectious diseases, of which most focused on HIV/AIDS, malaria, meningitis and pertussis. The majority of these studies were conducted using the two source capture-recapture method. Following is a summary of some of these studies:

- A two source capture-recapture method was used to estimate the completeness of the French AIDS surveillance system for a period from 1990-1993.[77] The sources used were the mandatory AIDS surveillance system and a hospital database on HIV infection. The study estimated that the AIDS surveillance system was 83.6% (95%CI 82.9-84.3%) complete. The main concern raised by the researchers was the fact that they had no way of validating false matches (negative or positive) as confidentiality of the data sources limited accessed to personal information in order to fulfil the validation. The researchers made reasonable efforts to determine if there was any dependence between the data sources by conducting interviews with those involved in collecting and managing the data. The influence or direction of dependence between the data sources could be taken into consideration when the two source estimate was determined. The requirement of homogeneity in the study population was not considered or discussed in the study.
- A two-source capture-recapture method was used to estimate the number of people hospitalised for suspected dengue in Puerto-Rico in 1991-1995.[78] The two data sources used were a laboratory based surveillance system and hospital based surveillance system. The study estimated that in non-epidemic years on average 2 791 (95%CI 1553-3481) and in epidemic years on average 9 479 (95%CI 9076-9882) suspected dengue patients were hospitalised. By including the hospital based system case registration was improved by 12.6%.[77] This study used two data sources in which the identification of a case differed considerably. The one data source had laboratory confirmation while the other only had suspected cases. This could cause a considerable over-estimation.
- A three-source log-linear capture-recapture method was used to estimate underreporting of Legionnaires' disease and the feasibility of a laboratory based reporting system in France in 1995.[79] The three data sources used were the national notification system, a reference laboratory database and a hospital laboratory survey. The study estimated 528 patients with of Legionnaires' disease with a sensitivity of the notification of 9%.[78] This study used a stratified analysis in order to account for heterogeneity of the study population. A limitation of this study was that Legionnaires' disease is reasonably rare and the overlap between the data sources is small (6-14 cases). This could have had a negative effect on the estimate.
- A three-source log-linear capture-recapture method was used to estimate the completeness of three data sources for meningococcal disease in the Netherlands from 1993-1999.[80] The three data sources used were a notification register, hospital episode statistics and data from a reference laboratory for bacterial meningitis. The completeness of the notification, hospital and

laboratory registers was estimated at 49%, 67% and 58% respectively before correction of false-positive diagnoses. After correcting for false-positive cases the completeness was estimated at 52%, 70% and 62% respectively. Due to confidentiality constraints the researchers could not verify record linking on name and surname. Some false matches or missed matches were therefore possible. The researchers also expected some false-positive cases as those identified from the hospital were not laboratory confirmed and case could have had incorrect registration or discharge diagnosis of meningococcal disease.

- A two-source capture-recapture method was used to estimate the number of deaths due to measles in the USA from 1987-2002.[81] The study also evaluated the efficiency of the two reporting systems. The two systems evaluated were death certificates from the national centre for health statistics and the national measles surveillance system. The total number of measles deaths was estimated at 259 (95%CI 244-274). The completeness of the national centre for health statistics was 64% and for the national measles surveillance system 71%. This study was conducted on two databases with deaths related to measles, although neither of these databases has any methods for confirming that an individual has measles. There is therefore no way of identifying false-positive cases. The authors mention that the two databases are independent, but do not substantiate this claim. Overall the number of measles cases is very low, with one year when an outbreak of measles occurred, and this could have a negative impact on the two source capture-recapture estimate. Heterogeneity in the study population is also questionable as children have a higher risk of acquiring measles. The perfect record linking requirement could also not be assessed, as a name or surnames were not available in either database.

As can be seen from the above reviews the use of both two- and three-source capture-recapture methods have been used in an array of diseases and in settings all over the world. Capture-recapture analysis has been used to assess the completeness of registration of tuberculosis in various countries, also in resource-limited settings and have been well documented for public healthcare. The following are some capture-recapture studies conducted particularly pertaining to tuberculosis:

- The incidence of pulmonary TB in a shantytown in Las Pampas de San Juan de Miraflores, Peru, was estimated using a two-source capture-recapture for a period 1989-1993.[82] The data sources used were interviews with local residents and laboratory sputum smear records. The study estimated the average annual incidence of pulmonary TB per 100 000 inhabitants at 364

(95%CI 293-528) and completeness of Ministry of Health reports at 37% (95%CI 25-46%). The study population consist of migrants and therefore the closed population requirement was most likely violated. Two different methods were used to identify cases, the one being laboratory confirmed cases and the second reported TB through interviews. This would have introduced false positive cases and therefore caused an overestimate of TB cases.

- A three-source log-linear capture-recapture method was used to estimate the level of under-reporting and to improve the incidence of TB in Cayenne, French Guyana for 1996-2003.[83] The three sources used were a mycobacterial laboratory database, hospital information database and the TB Control Service database. The study estimated a total of 462 (95%CI 423-536) and an under-reporting of 49.1%, 38.7% and 41.3% for the three data sources respectively. Due to the long study period and large geographical area, the requirements of a closed population and the equal probability of being captured and recaptured were difficult to evaluate.
- The annual incidence of TB and the completeness of TB registration in England, for 1999-2002, were estimated using a three-source log-linear capture-recapture method.[84] The three data sources used were the Mandatory TB surveillance system, national mycobacteriology reference laboratory records and a hospital admission database. Annual estimated completeness of notification between 1999 and 2002 was 48.1%, 51.1%, 59.0% and 66.5% respectively.[79] The use of the selected data sources in this study highlighted a number of limitations for capture-recapture when the underlying requirements for a capture-recapture application are violated. The TB cases identified from the hospital could not be confirmed as true TB cases and therefore false positive cases could have been introduced. In England the policy was to treat TB patients as outpatients and to isolate infectious patients at home. This could explain the reason for such a low number of TB cases being identified from the hospital. The perfect record linking requirement could also not be assured as names and surnames were not available in the data sources. Violation of these two requirements could have led to an over-estimation of TB cases. Violation of the homogeneity requirement was also likely as the difference in age, site of disease and infectiousness, among others, can cause different probabilities of being observed in a TB data source.

All of the above mentioned studies violated the underlying requirements for capture-recapture in some or other form. However the process followed, e.g. record linking, in order to complete the capture-recapture analyses produced a better and more complete number of reported cases than

any one particular dataset alone. In all the above capture-recapture studies the estimates produced could be questioned, but still produced useful results as long as the estimates are considered in the context of what was previously known of the population studied. Any known limitation of the data sources will also help interpret the estimates correctly, as long as the researchers report on these limitations and not take the estimates on face value.

A three-source log-linear capture-recapture study was conducted in Egypt for the year 2007.[80] The study used record linking and three-source capture-recapture analysis of data collected through active prospective longitudinal surveillance within the public and private non-NTP sector in four Egyptian governorates. This study indicated that for all TB cases, the estimated case detection rate of NTP surveillance was 55% (95%CI 46-68) and 62% (95%CI 52-77) for completeness of case ascertainment after record linking. For sputum smear-positive TB cases, these proportions were respectively 66% (95%CI 55-75) and 72% (95%CI 60-82). Some of the key differences between the Egypt study and the current study were that the Egypt study had a prospective data collection approach and included TB cases from the NTP and non-NTP sector, which also included cases from the private health sector. The Egypt study also looked at all forms of TB and spear-positive TB cases. The current study used retrospective data and only includes TB cases from the NTB services and did not include TB cases from the private sector.

2.5 REFERENCES TO CHAPTER 2

1. US Department of Health and Human Services. Principles of epidemiology. An introduction to applied epidemiology and biostatistics. In: Self-study course 3030-G. 2nd ed. Atlanta: Centre for Disease Control; 1992.
2. Greenberg RS, Daniels SR, Flanders WD, Eley JW, Boring JR. Medical epidemiology. 2nd ed. New York: Lange Medical Books/McGraw-Hill; 2001.
3. Stone DH. A method for the validation of data in a register. Public Health 1986;100:316-24.
4. Bain MR, Chalmers JW, Brewster DH. Routinely collected data in national and regional databases--an under-used resource. J Public Health Med 1997;19:413-8.
5. Roos LL, Sharp SM, Wajda A. Assessing data quality: a computerized approach. Soc Sci Med 1989;28:175-82.
6. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep 2001;50:35 pp.
7. Hook EB, Regal RR. The value of capture-recapture methods even for apparent exhaustive surveys. The need for adjustment for source of ascertainment intersection in attempted complete prevalence studies. Am J Epidemiol 1992;135:1060-7.
8. Desenclos JC, Bijkerk H, Huisman J. Variations in national infectious diseases surveillance in Europe. Lancet 1993;341:1003-6.
9. Capture-recapture and multiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995;142:1047-58.
10. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17:243-64.
11. Goldberg MS, Carpenter M, Theriault G, Fair M. The accuracy of ascertaining vital status in a historical cohort study of synthetic textiles workers using computerized record linkage to the Canadian Mortality Data Base. Can J Public Health 1993;84:201-4.
12. Herrchen B, Gould JB, Nesbitt TS. Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. Comput Biomed Res 1997;30:290-305.
13. Howe GR. Use of computerized record linkage in cohort studies. Epidemiol Rev 1998;20:112-21.

14. Muse AG, Mikl J, Smith PF. Evaluating the quality of anonymous record linkage using deterministic procedures with the New York State AIDS registry and a hospital discharge file. Stat Med 1995;14:499-509.
15. Waien SA. Linking large administrative databases: a method for conducting emergency medical services cohort studies using existing data. Acad Emerg Med 1997;4:1087-95.
16. Wendel HF. Medical record linkage--we need it now. J Clin Comput 1984;13:72-9.
17. Dunn HL. Record linkage. Am J Public Health 1946;36:1412-6.
18. Roos LL, Wajda A. Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods Inf Med 1991;30:117-23.
19. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate Record Detection: A Survey. IEEE Transactions on Knowledge and Data Engineering 2007;19:1-16.
20. Finney JM, Walker AS, Peto TEA, Wyllie DH. An efficient record linkage scheme using graphical analysis for identifier error detection. BMC Med Inform Decis Mak 2011;11:1-7.
21. Fair ME, Lalonde P, Newcombe HB. Application of exact ODDS for partial agreements of names in record linkage. Comput Biomed Res 1991;24:58-71.
22. Howe GR, Lindsay J. A generalized iterative record linkage computer system for use in medical follow-up studies. Comput Biomed Res 1981;14:327-40.
23. Meray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidemiol 2007;60:883-91.
24. Pacheco AG, Saraceni V, Tuboi SH, 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. Am J Epidemiol 2008;168:1326-32.
25. Roos LL, Jr., Wajda A, Nicol JP. The art and science of record linkage: methods that work with few identifiers. Comput Biol Med 1986;16:45-57.
26. Wajda A, Roos LL, Layefsky M, Singleton JA. Record linkage strategies: Part II. Portable software and deterministic matching. Methods Inf Med 1991;30:210-4.
27. Friedman C, Sideli R. Tolerating spelling errors during patient validation. Comput Biomed Res 1992;25:486-509.
28. Blakely T, Salmond C. Probabilistic record linkage and a method to calculate the positive predictive value. Int J Epidemiol 2002;31:1246-52.
29. Newcombe HB. Handbook of record linkage: methods for health and statistical studies, administration and business. Oxford: Oxford University Press; 1988

30. "NYSIIS", in Dictionary of Algorithms and Data Structures. U.S. National Institute of Standards and Technology, 2009. (Accessed 29/08/2011, at http://www.nist.gov/dads/HTML/nysiis.html.)
31. Schnell R, Bachteler T, Reiher J. Privacy-preserving record linkage using Bloom filters. BMC Med Inform Decis Mak 2009;9:41.
32. Newcombe HB, Kennedy JM, Axford SJ, James AP. Automatic linkage of vital records. Science 1959;130:954-9.
33. Fellegi IP, Sunter A. A theory of record linkage. J Am Statist Assoc;64:1183-210.
34. Scheuren F. Methodologic issues in linkage of multiple data bases. Vital Health Stat 4 1988:75-95.
35. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995;14:491-8.
36. Fair M, Cyr M, Allen AC, Wen SW, Guyon G, MacDonald RC. An assessment of the validity of a computer system for probabilistic record linkage of birth and infant death records in Canada. The Fetal and Infant Health Study Group. Chronic Dis Can 2000;21:8-13.
37. Beebe GW. Record linkage systems--Canada vs the United States. Am J Public Health 1980;70:1246-8.
38. Registry Plus™ Link Plus. Centers for Disease Control and Prevention, 2007. (Accessed August, 2010, at http://www.cdc.gov/cancer/npcr/tools/registryplus/.)
39. Korzeniewski SJ, Grigorescu V, Copeland G, et al. Methodological innovations in data gathering: newborn screening linkage with live births records, Michigan, 1/2007-3/2008. Matern Child Health J 2010;14:360-4.
40. Bierrenbach AL, Stevens AP, Gomes AB, et al. [Impact on tuberculosis incidence rates of removal of repeat notification records]. Rev Saude Publica 2007;41 Suppl 1:67-76.
41. Holian J. Client and birth record linkage: a method, biases, and lessons. Eval Pract 1996;17:227-35.
42. van Hest NA, Smit F, Baars HW, et al. Completeness of notification of tuberculosis in The Netherlands: how reliable is record-linkage and capture-recapture analysis? Epidemiol Infect 2007;135:1021-9.
43. Edginton ME, Wong ML, Phofa R, Mahlaba D, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: numbers of patients diagnosed and outcomes of referrals to district clinics. Int J Tuberc Lung Dis 2005;9:398-402.

44. Botha E, den Boon S, Lawrence KA, et al. From suspect to patient: tuberculosis diagnosis and treatment initiation in health facilities in South Africa. Int J Tuberc Lung Dis 2008;12:936-41.
45. Edginton ME, Wong ML, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: an intervention to improve patient referrals to district clinics. Int J Tuberc Lung Dis 2006;10:1018-22.
46. Botha E, Den Boon S, Verver S, et al. Initial default from tuberculosis treatment: how often does it happen and what are the reasons? Int J Tuberc Lung Dis 2008;12:820-3.
47. National Department of Health. The South African National Tuberculosis Control Programme: practical guidelines. Pretoria: National Department of Health; 2004.
48. Petersen CG. The yearly immigration of young place into the Limfjord from the German sea. Rep Dan Biol Stat 1896;6:1-48.
49. Smith ME. Record linkage: present status and methodology. J Clin Comput 1984;13:52-71.
50. Machado CJ. A literature review of record linkage procedures focusing on infant health outcomes. Cad Saude Publica 2004;20:362-71.
51. Le Cren ED. A note on the history of mark-recapture population estimates. J Animal Ecol;34:453.
52. Shapiro S. Estimating Birth Registration Completeness. J Am Stat Assoc 1950;45.
53. Hogan H. The 1990 Post-Enumeration Survey: operations and results. J Am Stat Assoc 1993;88:1,047-60.
54. Wittes J, Sidel VW. A generalization of the simple capture-recapture model with applications to epidemiological research. J Chronic Dis 1968;21:287-301.
55. Wittes JT, Colton T, Sidel VW. Capture-recapture methods for assessing the completeness of case ascertainment when using multiple information sources. J Chronic Dis 1974;27:25-36.
56. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses.: U California Public Stat 1951131-60
57. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture-recapture estimate. Biometrics 1972:592-7.
58. Schnabel ZE. The estimation of the total fish population of a lake. Am Math Monthly 1938;45:348.
59. Seber GAF. The estimation of animal abundance and related parameters. 2nd ed ed. London: Charles Griffin & Co.; 1982

60. Fienberg SE. The multiple recapture census for closed populations incomplete 2K contingency tables. Biometrika 1972;59:591-603.
61. Pollock KH. Modeling capture, recapture and removal statistics for estimation of demographic parameters for fish and wildlife populations: past, present and future. J Am Stat Assoc 1991;86:225.
62. Seber GAF. A review of estimating animal abundance II. Int Stat Rev 1992;60:129.
63. Capture-recapture and multiple-record systems estimation II: Applications in human diseases. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995;142:1059-68.
64. Tilling K. Capture-recapture methods--useful or misleading? Int J Epidemiol 2001;30:12-4.
65. Hook EB, Regal RR. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. Am J Epidemiol 1993;137:1148-66.
66. Derroch JN, Fienberg SE, Glonek GF, Junker BW. A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. J Am Stat Assoc 1993;88:1,137-48.
67. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. Epidemiology 1995;6:42-8.
68. Desenclos JC, Hubert B. Limitations to the universal use of capture-recapture methods. Int J Epidemiol 1994;23:1322-3.
69. Neugebauer R, Wittes J. Voluntary and involuntary capture-recapture samples--problems in the estimation of hidden and elusive populations. Am J Public Health 1994;84:1068-9.
70. Tilling K, Sterne JA. Capture-recapture models including covariate effects. Am J Epidemiol 1999;149:392-400.
71. Sakamoto Y, Ishiguru M, Kitigawa G. Akaike information criterion statistics. Tokio: KTK Scientific; 1986
72. Agresti A. Categorical data analysis. New York: John Wiley and Sons; 1990
73. Papoz L, Balkau B, Lellouch J. Case counting in epidemiology: limitations of methods based on multiple data sources. Int J Epidemiol 1996;25:474-8.
74. Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources. Am J Epidemiol 2000;152:771-9.

75. Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. J Clin Epidemiol 1999;52:909-14.
76. Hay G. The selection from multiple data sources in epidemiological capture-recapture studies. Statistician 1997;46:515.
77. Dechant EJ, Rigau-Perez JG. Hospitalizations for suspected dengue in Puerto Rico, 1991-1995: estimation by capture-recapture methods. The Puerto Rico Association of Epidemiologists. Am J Trop Med Hyg 1999;61:574-8.
78. Infuso A, Hubert B, Etienne J. Underreporting of legionnaires disease in France : the case for more active surveillance. Euro Surveill 1998;3:48-50.
79. NA VANH, Story A, Grant AD, Antoine D, Crofts JP, Watson JM. Record-linkage and capture-recapture analysis to estimate the incidence and completeness of reporting of tuberculosis in England 1999-2002. Epidemiol Infect 2008;136:1606-16.
80. Bassili A, Grant AD, El-Mohgazy E, et al. Estimating tuberculosis case detection rate in resource-limited countries: a capture-recapture study in Egypt. Int J Tuberc Lung Dis 2010;14:727-32.

CHAPTER 3 : AIMS AND OBJECTIVES

The aim of this study is to investigate the limitations of the present system of TB registration in order to arrive at a quantitative assessment of the registration data for two high TB incidence communities in the Western Cape.

3.1 OBJECTIVES

- To determine the accuracy and completeness of the data recorded in the TB treatment registers and laboratory records and the accuracy and completeness of case registration.
- To assess the contribution of capture-recapture analysis in estimating the completeness of recording and case ascertainment of bacteriologically confirmed TB in two communities in South Africa using record linking of the TB treatment register and two laboratory data sources.

This research will help to identify limitations in the TB treatment register data and the TB registration process. These limitations will be reported to the health authorities in order to improve the health system.

The study will indicate if a capture-recapture method is appropriate for estimating the total number of TB cases in communities or if more data and a more advanced capture-recapture method is required. This information is needed not only by the South African Health authorities but by all countries where incomplete TB registers are a problem. The World Health Organisation is also interested in this trial run of capture-recapture methods.

The capture-recapture will provide us with a more complete estimate of the number of TB cases in the two communities and will improve TB management in the specific communities. If the capture-recapture estimation from the available data sources proves to be accurate enough, the same method could be used in other communities.

The following additional information is provided in order to present a more complete picture of the background to the study than allowed for in the limited space of the Methods sections allowed in the journal articles.

3.2 THE STUDY SITE

The present study was conducted in two high incidence TB communities in Cape Town. The two adjoining communities have a total surface area of 3.4km^2 and an estimated total population of 36,343.[1] The total TB case registration in these two communities were 746/100 000 population in 1997 and 1 089/100 000 population in 2007. These figures were calculated based on TB treatment register data captured at the Desmond Tutu TB Centre of the Stellenbosch University from the original registers found at the clinics. These two areas have been the focus of a number of studies over the past decade.[1-12]

Each community has a clinic where TB treatment is freely available and during the study period of 2007 cases were managed in accordance with the 2004 South African National TB programme guidelines, with a TB treatment register located in each clinic.[13] Sputum samples collected at these clinics are transported by courier to a centralised laboratory in Green Point, Cape Town. The area is also served by a tertiary care hospital, Tygerberg Hospital, which also serves as the referral hospital for these communities. All sputum samples collected at Tygerberg Hospital are sent to a laboratory located inside the hospital.

A number of studies have been conducted in this study area since 1992. Some of the earlier studies indicated that the area had a high prevalence of TB among children and that these children presented late to health services.[2,3,5,6,8] Further insight into the TB problem in this area was provided by studies which indicated that most of the TB transmission in these communities occurred outside the household[4,7,11] and that the TB epidemic is fueled by ongoing transmission of TB in the community and not reactivation of TB.[12] It was also shown that the TB cases with a second episode of TB, after previous successful treatment, were due to mostly reinfection and not reactivation of TB.[9,10] A TB prevalence survey was conducted in 2002 which indicated a TB prevalence of 10% (95% CI 6.2 – 13.8) of bacteriologically confirmed TB.[1] A limitation of this study was that the sample size was too small, as can be inferred from the wide confidence interval. Two tuberculin surveys in 1998 and 2005 indicated an ARTI of 3.7 (95% CI 3.4 – 4.0) and 3.9 (95% CI 3.6 – 4.3) respectively.

Even with all these studies conducted in these two communities, there is no indication on how accurate and complete TB case registration is in these two communities. A more accurate estimation of the number of TB cases will be useful in assessing why these two communities continue to have such a major problem with TB. Record linking was therefore used, as described

in chapter 4, to determine the accuracy and completeness of TB case registration in the TB treatment register. In chapter 5, the results of the record linking conducted in chapter 4 were then used to conduct a capture-recapture analysis in order estimate a more accurate number of TB cases.

3.3 REFERENCES TO CHAPTER 3

1. den Boon S, van Lill SW, Borgdorff MW, et al. High prevalence of tuberculosis in previously treated patients, Cape Town, South Africa. Emerg Infect Dis 2007;13:1189-94.
2. Beyers N, Gie RP, Schaaf HS, et al. Delay in the diagnosis, notification and initiation of treatment and compliance in children with tuberculosis. Tuber Lung Dis 1994;75:260-5.
3. Beyers N, Gie RP, Schaaf HS, et al. A prospective evaluation of children under the age of 5 years living in the same household as adults with recently diagnosed pulmonary tuberculosis. Int J Tuberc Lung Dis 1997;1:38-43.
4. Classen CN, Warren R, Richardson M, et al. Impact of social interactions in the community on the transmission of tuberculosis in a high incidence area. Thorax 1999;54:136-40.
5. Marais BJ, Gie RP, Hesseling AC, et al. A refined symptom-based approach to diagnose pulmonary tuberculosis in children. Pediatrics 2006;118:e1350-9.
6. Marais BJ, Hesseling AC, Gie RP, Schaaf HS, Beyers N. The burden of childhood tuberculosis and the accuracy of community-based surveillance data. Int J Tuberc Lung Dis 2006;10:259-63.
7. Schaaf HS, Michaelis IA, Richardson M, et al. Adult-to-child transmission of tuberculosis: household or community contact? Int J Tuberc Lung Dis 2003;7:426-31.
8. van Rie A, Beyers N, Gie RP, Kunneke M, Zietsman L, Donald PR. Childhood tuberculosis in an urban population in South Africa: burden and risk factor. Arch Dis Child 1999;80:433-7.
9. van Rie A, Warren R, Richardson M, et al. Exogenous reinfection as a cause of recurrent tuberculosis after curative treatment. N Engl J Med 1999;341:1174-9.
10. Verver S, Warren RM, Beyers N, et al. Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. Am J Respir Crit Care Med 2005;171:1430-5.
11. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet 2004;363:212-4.
12. Verver S, Warren RM, Munch Z, et al. Transmission of tuberculosis in a high incidence urban community in South Africa. Int J Epidemiol 2004;33:351-7.
13. National Department of Health. The South African National Tuberculosis Control Programme: Practical Guidelines. Pretoria: National Department of Health; 2004.

CHAPTER 4 : ACCURACY AND COMPLETENESS OF RECORDING OF CONFIRMED TUBERCULOSIS IN TWO SOUTH AFRICAN COMMUNITIES

R. Dunbar,[*] K. Lawrence,[*] S. Verver,[†‡] D. A. Enarson,[§] C. Lombard,[¶] J. Hargrove,[#] J. Caldwell,[**] N. Beyers,[*] J. M. Barnes[††]

[*] Desmond Tutu Tuberculosis Centre, Department of Paediatrics and Child Health, Stellenbosch University, Cape Town, South Africa; [†] KNCV Tuberculosis Foundation, The Hague, [‡] Centre for Infection and Immunity Amsterdam (CINIMA), Academic Medical Centre, University of Amsterdam, Amsterdam, The Netherlands; [§] International Union Against Tuberculosis and Lung Disease, Paris, France; [¶] Biostatistics Unit, Medical Research Council, Cape Town, [#] South African Centre for Epidemiological Modelling and Analysis, Stellenbosch University, Cape Town, [**] Department of Health, City of Cape Town, Cape Town, [††] Division of Community Health, Stellenbosch University, Cape Town, South Africa

INT J TUBERC LUNG DIS 2011 15(3):337–343

4.1 SUMMARY

BACKGROUND: Although tuberculosis (TB) treatment registers and laboratory records are essential tools for recording and reporting in TB control programmes, the accuracy and completeness of routinely collected data are seldom monitored.

OBJECTIVE: To assess the accuracy and completeness of TB treatment register data in two South African urban communities using record linking.

METHODS: All cases of bacteriologically confirmed TB, defined as two smear-positive results and/or at least one culture-positive result, were included. Record linking was performed between three data sources: 1) TB treatment registers, 2) the nearest central laboratory, and 3) the referral hospital laboratory.

RESULTS: The TB treatment registers had 435 TB cases recorded, of which 204 (47%) were bacteriologically confirmed. An additional 39 cases recorded as nonbacteriological cases in the TB treatment registers were reclassified as bacteriologically confirmed, and 63 bacteriologically confirmed cases were identified from the laboratory databases that were not recorded in the TB treatment registers. The final number of bacteriologically confirmed TB cases was 306, giving an increase of 50%.

CONCLUSIONS: The accuracy and completeness of the TB treatment register and central laboratory data were inadequate. A high percentage of bacteriologically confirmed cases from both laboratories were not recorded in the TB treatment registers. We are developing an electronic result management system to improve the management of laboratory results.

4.2 INTRODUCTION

IN SOUTH AFRICA, as in many other countries, a tuberculosis (TB) treatment register is used by the health services for the management of TB treatment as well as for the recording and reporting of TB cases. TB treatment registers are located in peripheral clinics, and data from the clinics are collated at the district level to record the number of TB cases and report these to the National TB Programme (NTP). Sputum samples for TB diagnosis, together with a request form, are sent from the peripheral clinics to the centralised laboratories of the National Health Laboratory Service (NHLS).

At least two publications have reported on the completeness of case registration in South Africa.[1,2] A study conducted in the Stellenbosch District of Western Cape Province indicated that 17% of TB cases with two smear-positive results were not recorded in the TB treatment registers.[1] These cases presented to health care facilities and were diagnosed, but never started treatment and were referred to as 'initial defaulters'. Another study conducted at the Chris Hani Baragwanath Hospital in Johannesburg indicated that 21% of the patients referred to a clinic from the hospital did not attend any clinic in Soweto and were also not recorded in the TB treatment register.[2] No studies in South Africa have thus far focused specifically on the accuracy and completeness of the data found in TB treatment registers or on the accuracy and completeness of the data found in the databases of centralised laboratories.

Routinely collected data to address specific research questions cannot be accepted without evaluation. These data sources are primarily designed as management tools and should first be thoroughly evaluated to determine whether they are adequate for the research question being posed. Such data have limitations, including lack of accuracy and completeness.[3] The aim of this study was to determine the accuracy and completeness of the data recorded in the TB treatment registers and laboratory records and the accuracy and completeness of case registration.

4.3 SETTING

The present study was conducted in two high-incidence TB communities in Cape Town. The two adjoining communities have a total surface area of 3.4 km2 and an estimated total population of 36 343. The total TB case registration in these two communities was 746 per 100 000 population in 1997 and 1089/100 000 in 2007. These figures were calculated based on TB treatment register data

captured at the Desmond Tutu TB Centre of the Stellenbosch University from the original registers found at the clinics. These two areas have been the focus of a number of studies over the past decade[4-9]

Each community has a clinic where TB treatment is freely available and is managed in accordance with the 2004 South African NTP guidelines,[10] with a TB treatment register located in each clinic. Sputum samples collected at these clinics are transported by courier to a centralised laboratory in Green Point, Cape Town. The area is also served by a tertiary care hospital, Tygerberg Hospital, which also serves as the admission hospital for these communities. All sputum samples collected at the Tygerberg Hospital are sent to a laboratory located inside the hospital.

4.3.1 Data sources

Data from the TB treatment registers and from the NHLS database for centralised and hospital laboratories were included in the study. The line data for the TB treatment register from the two community clinics for 2007 were captured and validated from copies of the original TB treatment registers located at the clinics. The centralised NHLS data from the two community clinics were received in electronic format from the NHLS data warehouse for sputum samples collected from 1 October 2006 to 31 March 2008. This included smear and culture results of all sputum samples. The hospital laboratory data were received directly from the laboratory in Tygerberg Hospital for cases with residential addresses in these two communities.

4.3.2 Case definition

A TB case was defined as an individual with bacteriologically confirmed TB having at least two positive smears or at least one culture-positive sputum result as per definition by the South African NTP Guidelines.[10]

4.4 METHODS

Record linking Link Plus, a probabilistic linking software package, was used to perform record linking.[11,12] Once records were matched within and between data sources using Link Plus, all matches and non-matches were manually reviewed for correctness. Uncertain matches were reviewed with the help of experienced nurses who worked in the clinics and knew the patients. The data fields used for linking purposes were name, surname, age and address. Name and surnames were converted to New York State Identification and Intelligence System (NYSIIS), a variation of

the more well-known Soundex phonetic coding system. This conversion reduces the possibility that slight misspellings in names or surnames result in records not being matched.

Linking was performed in two ways (Figure 4-1), using 'within-data-source' linking and 'between-data-source' linking. 'Within-data-source' linking was used to identify individuals who were recorded more than once in the TB treatment register for 2007. For these individuals, the subsequent record was considered as a separate TB case if the previous record had a treatment outcome of 'cured' or 'treatment completed'. 'Within-data-source' linking was used for both the centralised and hospital laboratory data sources to create a laboratory database for each laboratory of unique individuals from all the specimens received in the laboratory data-sources and to identify all individuals with bacteriologically confirmed TB.

'Between-data-source' linking was performed between the TB treatment register and both centralised and hospital databases produced from the 'within-data-source' linking. Any individuals from the laboratory databases who could not be linked to the TB treatment register for 2007 were looked for in the 2006 and 2008 TB treatment registers. Those who could be identified in the 2006 or 2008 TB treatment registers whose record was directly related to the TB case that was not recorded in the 2007 TB treatment register but was identified in the laboratory databases were excluded.

After 'between-data-source' linking, all results in the TB treatment register were linked with the results found in both laboratory databases. Only pre-treatment (diagnostic) smear and culture results from the TB treatment register were linked to the laboratory results after 'between-data-source' linking. This included both positive and negative smear and culture results to identify incorrectly recorded results in the TB treatment register.

4.4.1 Accuracy and completeness of recorded data

Accuracy in assessing data quality in information systems is defined as the proportion of data fields in the information system with a given characteristic that truly have this attribute.[13] The individual record will often contain several data fields in addition to the field by which the

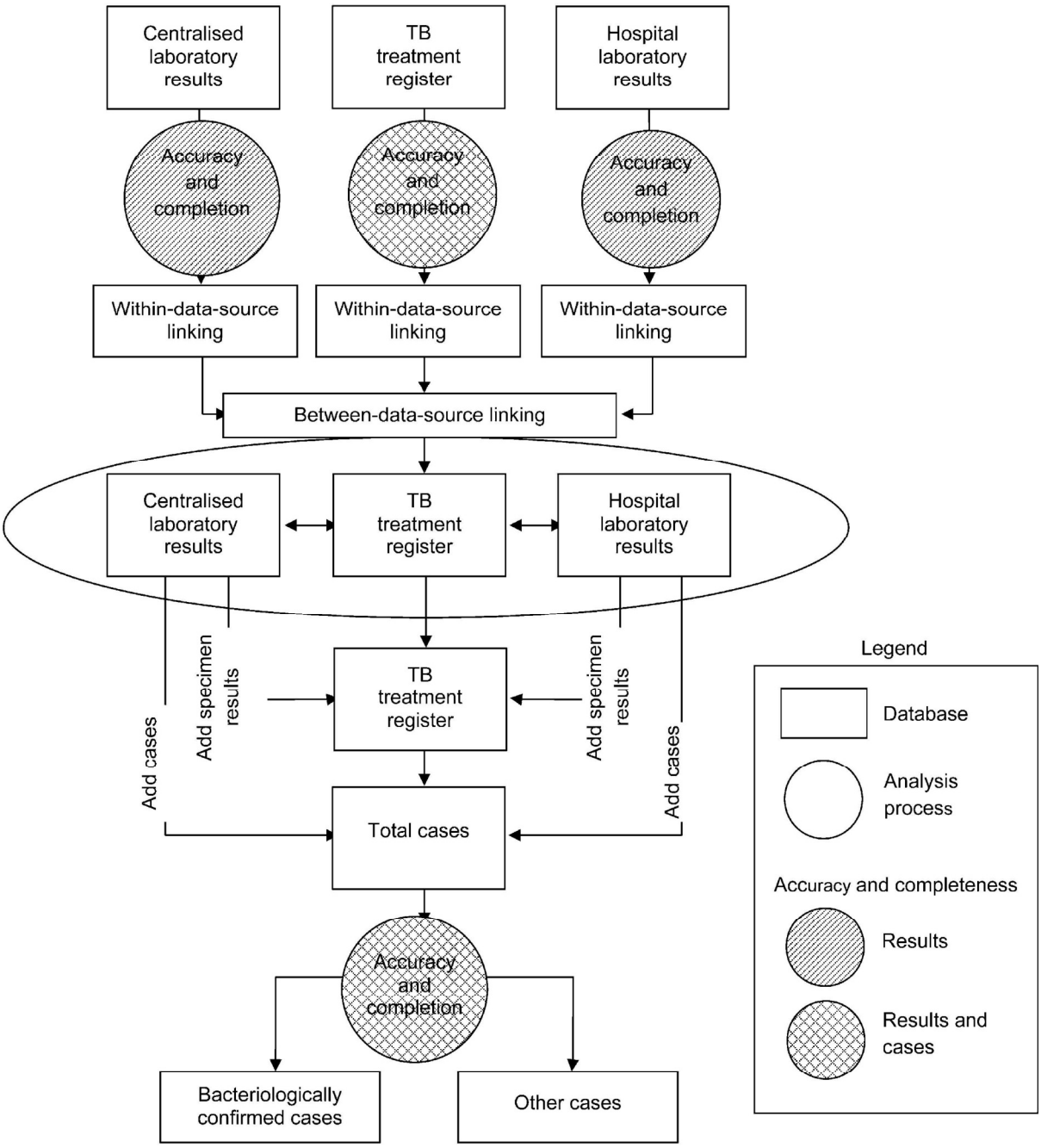

Figure 4-1: Process flow for assessing the accuracy and completeness of data sources and record linking. TB = tuberculosis.

person is identified, for example, the results of certain diagnostic tests, diagnoses, age, sex and demographic data. The completeness of data relates to how many data fields have missing data, and is thus a subcategory of data accuracy. The accuracy of each data source was calculated as the number of records with no discrepancies divided by the total number of records in the data sources.

4.4.2 Before record linking

Before record linking, we assessed whether the data in each data field were accurately recorded and if no data were missing.

4.4.3 After record linking

After linking the results from the laboratories to the TB treatment registers, the accuracy and completeness of the data were assessed. All data fields available in all the TB treatment registers and laboratory databases were compared. Names and surnames were compared between TB treatment registers and the laboratory databases to determine how many differed to such an extent that it would hamper tracing the result back to a TB case. Ages were compared to determine if any differed by more than 2 years between data sources.

For discrepancies in demographic information, timing of results (diagnostic or follow-up) and patient category (new or retreatment), the TB treatment register was regarded as the most accurate.

For discrepancies in sample results, laboratory data source was considered the most accurate source. In case of discrepancy in the sample result between the TB treatment register and the laboratory data source, the case recorded in the TB treatment register was re-evaluated to determine if any TB case already recorded in the TB treatment register as a non-bacteriological case should be reclassified as a bacteriologically confirmed case, based on the corrected results.

For assessment of completeness, all laboratory sputum results collected within 3 months before and 6 weeks after the date of recording the TB case in the TB treatment register were included in the analysis. The case classification in the TB treatment register was re-assessed to determine if any case classified as extra-pulmonary or non-bacteriological TB could be reclassified as pulmonary TB or bacteriologically confirmed TB, based on positive results in the laboratory database that had not been transcribed into the TB treatment register. Results of cases that were recorded as retreatment cases were checked to determine if all retreatment cases had culture results.

4.4.4 Accuracy and completeness of case recording

Before record linking

Before record linking, determination of the accuracy of case recording was possible only for the TB treatment registers. Accuracy of case recording could only be determined for pulmonary

disease, and was conducted based on criteria from the NTP guidelines.[10] All cases with smear-positive sputum results were assessed to determine if these had been recorded as pulmonary TB.

After record linking

After record linking, to estimate the accuracy of case recording, results in the TB treatment registers were compared to the linked results in the laboratory data source to determine if the results had been correctly recorded in the TB treatment register. Completeness of case recording was determined by counting how many TB cases in the centralised and hospital laboratory data sources were not recorded in the TB treatment register within 2 months of receiving the specimen at the laboratory.

4.4.5 Quality control in the NHLS laboratories

For external quality control in the NHLS laboratories, the laboratory receives two fixed slides and two samples to stain and evaluate every 3 months. Any unsatisfactory results are investigated and documented.

4.4.6 Ethics approval

Ethics approval was obtained from the Stellenbosch University Committee for Human Research. Permission for the study was granted by the City of Cape Town Health Directorate and the Department of Health, Western Cape Province.

All data were stored on a secure, password protected data server, with access only by the data manager. All identifying information was eliminated after record linking and a unique identifier was provided that was used for any further analysis and reporting. The researcher and study nurses completed confidentiality agreements as part of their contracts.

4.5 RESULTS

In the TB treatment register for 2007, 435 TB cases were recorded. Two cases had two treatment episodes each, with the first episode having an outcome of ‘cured’ or ‘treatment completed’. These treatment episodes were on each occasion counted as a case of TB, resulting in 435 cases treated for TB in 433 individuals. Of these 435 cases, 204 (47%) were bacteriologically confirmed. There were 263 bacteriologically confirmed cases in the centralised laboratory database and 38 in the

hospital laboratory database. The classification of the TB cases in the three data sources based on the number of smear and/or culture results is summarised in Table 4-1.

4.5.1 Accuracy and completeness of data recorded

Before linking

The data accuracy rate of the data sources before record linking was 89% in the TB treatment register, 81% in the centralised laboratory and 99% in the hospital laboratory. Of the 435 cases in the TB treatment register, 31 (7%) had inaccurate dates or results. The data fields that were most frequently incomplete in the 10 429 centralised laboratory results were date of birth (862, 8% absent), sex (734, 7% absent) and address (1140, 11% absent).

After linking

Accuracy and completeness after record linking between the TB treatment registers and the centralised laboratory and the TB treatment registers and the hospital laboratory are summarised in Table 4-2. The data fields with the largest number of discrepancies between the TB treatment registers and centralised laboratory results were age (180, 26%), sex (59, 9%) and patient category (53, 8%). After adding and correcting results in the TB treatment register from both laboratories, an additional 42 cases already recorded and treated as bacteriologically negative cases were reclassified as bacteriologically confirmed cases in the TB treatment registers, while three cases recorded as bacteriologically confirmed cases were reclassified as bacteriologically negative cases. The number of bacteriologically confirmed cases in the TB treatment register increased from 204 to 243, an increase of 16%.

4.5.2 Accuracy and completeness of case recording

The distribution of bacteriologically confirmed cases found in each data source after linking all three data sources is shown in a Venn diagram (Figure 4-2). The TB treatment registers had a total of 435 TB case records, of which 243 were bacteriologically confirmed. An additional 63 bacteriologically confirmed cases were identified from laboratory results that were not recorded in the TB treatment register: 45 (71%) from the centralised laboratory, 16 (25%) from the hospital laboratory and 2 (3%) from both laboratories. Of the 63 additional bacteriologically confirmed cases, eight (13%) had at least two positive smear results, 41 (65%) had at least one positive culture result and 14 (22%) had a positive smear and a positive culture result.

Table 4-1: Number of cases initially found in each register, reason for excluding cases and classification of cases based on the number of positive smear or culture results (based on treatment register year 2007).

Category	2007 TB treatment register n(%)	Centralised laboratory database (1 October 2006 to 31 March 2008) n(%)	Hospital laboratory database (1 October 2006 - 31 March 2008) n(%)
Initial number of cases	435	785	278
Reason for exclusion			
Cases in 2006 not recorded in TB treatment register in 2006 or 2007	0	33 (4)	0
Cases recorded in TB treatment register 2006	0	122 (16)	0
Cases in 2008 not recorded in TB treatment register in 2008 or 2007	0	43 (5)	0
Cases recorded in TB treatment register 2008	0	79 (10)	0
Total cases excluded	0	277 (35)	0
Final number of cases in each data source	435	508 (65)	278
Categorisation of cases based on number of smear and/or culture results			
1 smear-positive	82 (19)	171 (34)	8 (17)
≥ 2 smear positive*	84 (19)	65 (13)	0
≥ 1 culture-positive*	46 (11)	75 (15)	23 (50)
Smear- and culture-positive*	74 (17)	123 (24)	15 (33)
Total bacteriologically confirmed cases*	204 (47)	263 (52)	38 (14)

*Cases meeting the definition of a bacteriologically confirmed case: two smear- and/or at least one culture-positive results.
TB = tuberculosis

Table 4-2: Accuracy and completeness of data recorded after linking data sources.

	TB treatment register vs centralised laboratory n(%)	TB treatment register vs hospital laboratory n(%)
Cases linked between data sources	350	25
Samples linked between data sources	689	14
Number incorrect, assumed correct in TB treatment register		
Name/surname	0	0
Age (>2 years difference in age)	180 (26)	0
Sex	59 (9)	0
Date registered	0	0
	53 (8)	0
Diagnosis vs Follow-up	4 (1)	0
Number incorrect, assumed correct in laboratory		
Smear results		
False-positive	14 (3)	0
False-negative	8 (1)*	0
Culture-results		
False-positive	0	0
False-negative	2 (1)*	0
In addition to the linked results		
Smear results not recorded in TB treatment register	105	17
Positive†	38 (36)	9 (53)
Negative	67 (64)	8 (47)
Culture results not recorded in TB treatment register	67	25
Positive†	45 (67)	23 (92)
Negative	22 (33)	2 (8)

* 10 corrected results led to 3 extra bacteriologically confirmed cases and 3 bacteriologically confirmed cases being excluded

†115 additional results led to 1 EPTB episode becoming PTB and 39 additional bacteriologically confirmed cases.

TB = tuberculosis; EPTB = extra-pulmonary TB; PTB = pulmonary TB.

The final number of bacteriologically confirmed cases identified was 306 (Figure 4-2). These 306 bacteriologically confirmed cases included the 204 (67%) bacteriologically confirmed cases originally identified in the TB treatment register, an additional 39 (13%) cases recorded as non-bacteriological cases and reclassified as bacteriologically confirmed after incorporating all the results from the laboratory data sources, and 63 (21%) bacteriologically confirmed cases from the laboratories that had not been recorded in the TB treatment register. This resulted in an overall increase of 102 (50%) bacteriologically confirmed cases.

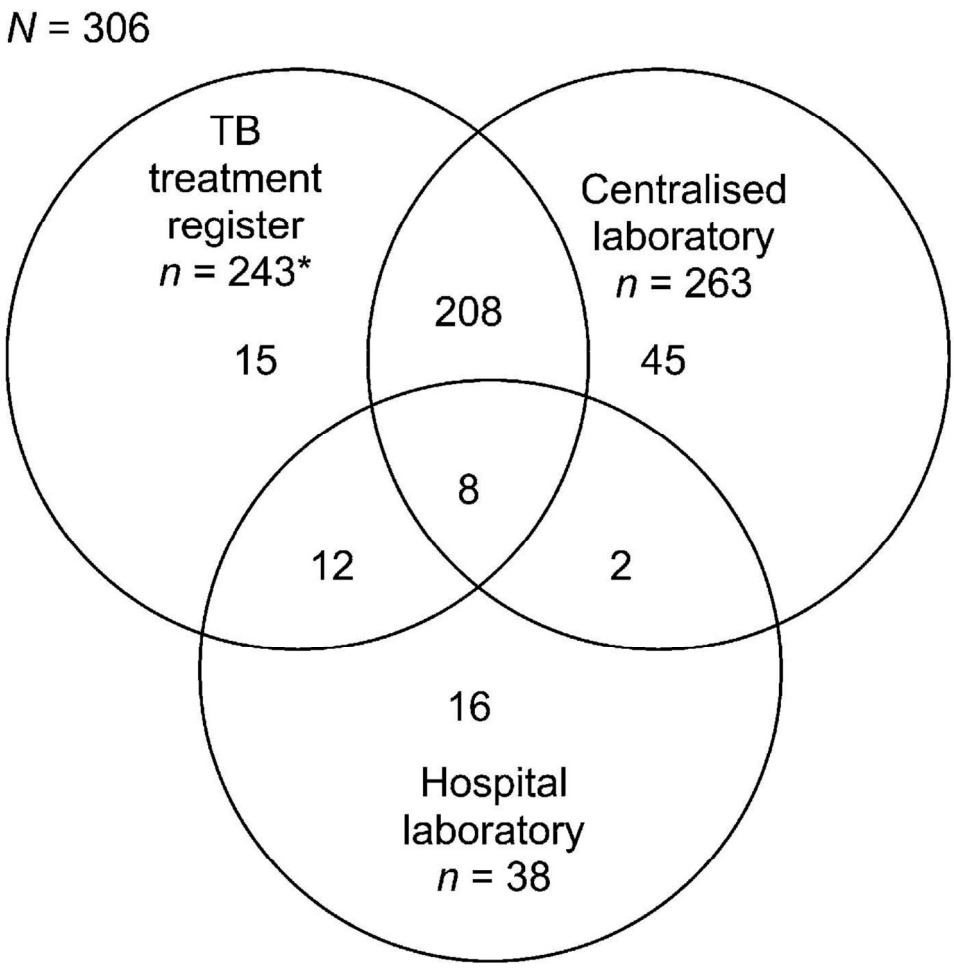

Figure 4-2: Distribution of bacteriologically confirmed TB cases after linking all three data sources. * The number of TB cases in the TB treatment register is the final number of TB cases after results were corrected and added.
TB = tuberculosis.

4.6 DISCUSSION

This study identified substantial inaccuracy and incompleteness in routine reporting of TB cases. It is possible that some of the 63 additional cases identified were treated at a health care facility outside the study area, completed treatment but were not recorded in the TB treatment registers or died before starting treatment. However, if this was not the case, the loss of these patients could have implications for the case detection rate, management and transmission of TB. If such cases are not treated, not only would it be detrimental to their health, but they would also be a continuing source of TB transmission. It is therefore important to identify and trace these individuals.

A high percentage of those not recorded in the TB treatment register were diagnosed only on culture (65%). Although these cases are less infectious than smear-positive cases, it is important to place them on treatment. A possible reason for these culture-positive cases not being recorded is the slow turnaround time of culture. It is encouraging to note, however, that 95% of retreatment episodes had at least one culture result.

Cases who were already recorded in the TB treatment register, but whose case definition changed due to correcting results from the centralised laboratory data, were already on treatment. There would therefore be no treatment implications for these patients. However, because some smear-negative cases already on treatment was reclassified as smear-positive cases, the smear-positive TB case detection rate and the number of TB cases detected would increase for each facility. This will be propagated up to district, provincial and finally the national level, and could represent a considerable cumulative increase.

Previous studies have reported high initial default rates,[1,2] of respectively 17% and 21%, although under different protocol definitions. A study conducted in Ho Chi Minh City, Viet Nam, reported an initial default rate of 8%.[14] A study in South Africa identified that 24% of initial defaulters died and 45% could not be located for interview.[15] Among those who could be interviewed, half indicated that the reason for not starting treatment was directly linked to service delivery.[15]

We are confident that the linking of the different data sources was accurate, as the data accuracy rate of the data sources was reasonable (81–99%). Hospital laboratory data were the most accurate, as they were linked to the patient management system. Any person who could not be linked could be verified by a nursing sister with many years of experience in the study communities.

The main limitation of this study is that we cannot be certain that all 63 additional cases are truly missed cases that did not access care elsewhere. Further studies are required to investigate whether these were indeed missed cases. This may be done by using additional data sources such as the electronic TB register and death registers. A further limitation that could also lead to an overestimation of non-registered TB cases is incorrect recording of names or ages, resulting in individuals who thus could not be linked.

These findings could be extrapolated to other countries, as most countries use centralised laboratories for culture diagnosis. This method can also be applied to paper-based systems. It is advisable to keep the number of variables used in assessing a surveillance system to a minimum.[16]

These methods have previously been used in cancer, measles, diabetes and malaria research.[17-20] The methods used in this study have also been recommended by the World Health Organisation as an approach that could be used in improving national case detection.[21] Alternative methods for estimating the number of TB cases are 1) prospective cohort studies, 2) surveys of the annual risk of TB infection and 3) using mortality data recorded in vital registration systems.[21]

The present study indicates a clear need to improve and monitor the accuracy and completeness of the routine forms used in TB care and to ensure that all patients identified in the laboratory are started on treatment. Better communications are needed at all levels of the health service. Inaccuracy of the laboratory data may be due to 1) inaccuracies in the specimen request form and 2) inaccuracies in data recording in the laboratory.

We conclude that the recording of laboratory results in TB treatment registers is not optimal and that this could have a negative impact on both the NTP and the utilisation of data records for research. The proportion of bacteriologically confirmed cases not recorded in the TB treatment register is high, and could be partly related to the slow turnaround time for culture results. These results indicate a need for improved recording of TB cases and TB management, which is a significant clinical, social and financial problem for South Africa. The national figures are used by the NTP for budgeting, monitoring TB outbreaks and to identify areas for targeted interventions, and inaccurate case detection rates may lead to incorrect decisions.

We recommend that a more integrated computerised health management system should be implemented to integrate all aspects of health care. A computerised system should include all types of health facilities and laboratories as well as a unique identifier which could identify an individual nationally throughout the health services. We are currently implementing a sputum result management system to ensure that all individuals with positive sputum results are placed on treatment. This system will be piloted in a number of clinics in Cape Town.

4.7 ACKNOWLEDGEMENTS

The authors express their appreciation to all the contributors to this study: National, Provincial and District TB Programme staff, and staff at Ravensmead and Uitsig clinics. They also thank the National Health Laboratory Service Corporate Data Warehouse for assisting with data extraction.

4.8 REFERENCES TO CHAPTER 4

1. Botha E, den Boon S, Lawrence KA, et al. From suspect to patient: tuberculosis diagnosis and treatment initiation in health facilities in South Africa. Int J Tuberc Lung Dis 2008;12:936-41.
2. Edginton ME, Wong ML, Phofa R, Mahlaba D, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: numbers of patients diagnosed and outcomes of referrals to district clinics. Int J Tuberc Lung Dis 2005;9:398-402.
3. Greenberg RS, Daniels SR, Flanders WD, Eley JW, Boring JR. Medical epidemiology. 2nd ed. New York: Lange Medical Books/McGraw-Hill; 2001.
4. Kritzinger FE, den Boon S, Verver S, et al. No decrease in annual risk of tuberculosis infection in endemic area in Cape Town, South Africa. Trop Med Int Health 2009;14:136-42.
5. Warren R, Hauman J, Beyers N, et al. Unexpectedly high strain diversity of Mycobacterium tuberculosis in a high-incidence community. S Afr Med J 1996;86:45-9.
6. Ellis JH, Beyers N, Bester D, Gie RP, Donald PR. Sociological and anthropological factors related to the community management of tuberculosis in the Western Cape communities of Ravensmead and Uitsig. S Afr Med J 1997;87:1047-51.
7. Munch Z, Van Lill SW, Booysen CN, Zietsman HL, Enarson DA, Beyers N. Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. Int J Tuberc Lung Dis 2003;7:271-7.
8. Verver S, Warren RM, Munch Z, et al. Proportion of tuberculosis transmission that takes place in households in a high-incidence area. Lancet 2004;363:212-4.
9. Verver S, Warren RM, Munch Z, et al. Transmission of tuberculosis in a high incidence urban community in South Africa. Int J Epidemiol 2004;33:351-7.
10. National Department of Health. The South African National Tuberculosis Control Programme: Practical Guidelines. Pretoria: National Department of Health; 2004.
11. Division of Cancer Prevention and Control. Registry Plus™ Link Plus. In. 2.0 ed. Atlanta: Centers for Disease Control and Prevention; 2007.
12. Korzeniewski SJ, Grigorescu V, Copeland G, et al. Methodological innovations in data gathering: newborn screening linkage with live births records, Michigan, 1/2007-3/2008. Matern Child Health J 2010;14:360-4.

13. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep 2001;50:35 pp.
14. Buu TN, Lonnroth K, Quy HT. Initial defaulting in the National Tuberculosis Programme in Ho Chi Minh City, Vietnam: a survey of extent, reasons and alternative actions taken following default. Int J Tuberc Lung Dis 2003;7:735-41.
15. Botha E, Den Boon S, Verver S, et al. Initial default from tuberculosis treatment: how often does it happen and what are the reasons? Int J Tuberc Lung Dis 2008;12:820-3.
16. Rieder HL, Watson JM, Raviglione MC, et al. Surveillance of tuberculosis in Europe. Working Group of the World Health Organization (WHO) and the European Region of the International Union Against Tuberculosis and Lung Disease (IUATLD) for uniform reporting on tuberculosis cases. Eur Respir J 1996;9:1097-104.
17. Couris CM, Colin C, Rabilloud M, Schott AM, Ecochard R. Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. J Clin Epidemiol 2002;55:386-91.
18. Gindler J, Tinker S, Markowitz L, Atkinson W, Dales L, Papania MJ. Acute measles mortality in the United States, 1987-2002. J Infect Dis 2004;189 Suppl 1:S69-77.
19. Giarrizzo ML, Pezzotti P, Silvestri I, Di Lallo D. [Estimating prevalence of diabetes mellitus in a Lazio province, Italy, by capture-recapture models]. Epidemiol Prev 2007;31:333-9.
20. Hansen HL, Hansen KG, Andersen PL. Incidence and relative risk for hepatitis A, hepatitis B and tuberculosis and occurrence of malaria among merchant seamen. Scand J Infect Dis 1996;28:107-10.
21. World Health Organization. Global tuberculosis control : epidemiology, planning, financing : WHO report 2009. Geneva: World Health Organization; 2009.WHO/CDS/TB/2003.313, 303 pp.

CHAPTER 5: CAPTURE-RECAPTURE TO ESTIMATE COMPLETENESS OF TUBERCULOSIS SURVEILLANCE IN TWO COMMUNITIES IN SOUTH AFRICA

R. Dunbar,[*] R. van Hest,[†] K. Lawrence,[*] S. Verver,[‡§] D. A. Enarson,[¶] C. Lombard,[#] N. Beyers,[*] J. M. Barnes[**]

[*] Desmond Tutu TB Centre, Department of Paediatrics and Child Health, Faculty of Health Sciences, Stellenbosch University, Cape Town, South Africa; [†] Department of Tuberculosis Control, Municipal Public Health Service Rotterdam-Rijnmond, Rotterdam, [‡] KNCV Tuberculosis Foundation, The Hague, [§] Centre for Infection and Immunity Amsterdam (CINIMA), Academic Medical Centre, University of Amsterdam, The Netherlands; [¶] International Union Against Tuberculosis and Lung Disease, Paris, France; [#] Biostatistics Unit, Medical Research Council, Cape Town, [**] Division of Community Health, Stellenbosch University, Cape Town, South Africa

INT J TUBERC LUNG DIS 2011 15(8):1038–1043

5.1 SUMMARY

BACKGROUND: Reliable surveillance is essential for any tuberculosis (TB) control programme; however, under-registration of TB cases due to under-notification of patients on treatment or failure to initiate treatment has been well-documented internationally.

OBJECTIVE: To determine the contribution of capture-recapture methods in estimating the completeness of bacteriologically confirmed pulmonary TB registration in two high-incident communities in South Africa.

METHODS: Record linking between the TB treatment register and two laboratory sputum TB result registers and three-source log-linear capture-recapture analysis.

RESULTS: The number of bacteriologically confirmed pulmonary TB cases in the TB treatment register was 243, with an additional 63 cases identified in the two laboratory databases, resulting in 306 TB cases. The observed completeness of the TB treatment register was 79%. The log-linear model estimated 326 (95%CI 314–355) TB cases, resulting in an estimated completeness of registration of 75% (95%CI 68–77).

CONCLUSION: Capture-recapture can be useful in evaluating the completeness of TB control surveillance and registration, including in resource-limited settings; however, methodology and results should be carefully assessed. Interventions are needed to increase the completeness of registration and to reduce the number of initial defaulters.

5.2 INTRODUCTION

TUBERCULOSIS (TB) is a global public health problem, especially in low- and middle-income countries experiencing a human immunodeficiency virus (HIV) epidemic, increasing poverty and population increase, where TB is prevalent.[1-3] In absolute numbers, in 2007 South Africa ranked fourth among the 22 highest TB burden countries in the world, with an estimated 461 000 incident TB cases, translating into an incidence rate of 948 per 100 000 population, by far the highest in these countries.[3] The World Health Organisation (WHO) implementation targets for TB control include a sputum smear-positive detection rate of 70%.[4] Meaningful quantification of bacteriologically confirmed pulmonary TB in a community is therefore an important task for any TB control programme, and surveillance is essential.[5] Most countries with a high burden of TB are resource-limited, do not yet have robust surveillance systems and the estimates of case detection rates according to the National TB Programmes (NTPs) are often unreliable.[6,7]

In resource-limited countries such as South Africa, under-notification, i.e., failure to report a patient on treatment to the NTP, is often attributed to the private sector.[8] Under-registration can also be caused by failure to initiate treatment and registration in the primary health care (PHC) clinics, due to gaps in the referral systems between the laboratories and hospitals, where suspected TB is bacteriologically diagnosed, and the PHC facilities.[9-11] Patients failing to report back for results and start treatment are known as initial defaulters, and will not appear in the case detection statistics. In the Stellenbosch District of South Africa, initial default proportions of 17% have been described for sputum smear-positive TB suspects.[10]

Correct interpretation of data on TB incidence and trends necessitates assessment of completeness of recording.[12] An important first step in this assessment is record linking, i.e., comparing patient data in the notification or official recording register with multiple other related data sources.[13] Completeness of recording can then be assessed by comparison with the case ascertainment, i.e., the total number of patients observed in at least one register. We recently performed a record linking study in two high-incidence communities in the Western Cape. We found an observed completeness of the TB treatment register in the PHC clinics of 79% after record linking with two laboratory registers.[14] Completeness of case-ascertainment can subsequently be estimated through capture-recapture analysis. Based on certain assumptions, capture-recapture methods use the information on the overlap between various registers to estimate the number of cases that are not recorded in all registers and thus the estimated total number of cases.[15,16] The preferred capture-

recapture method is log-linear modelling of at least three linked registers.[15,17-19] Capture-recapture analysis has been used to assess the completeness of registration of TB in various countries, including in resource-limited settings.[20]

The objective of the present study was to assess the contribution of capture-recapture analysis in estimating the completeness of recording and case ascertainment of bacteriologically confirmed TB in two communities in South Africa using record linking of the TB treatment register and two laboratory data sources.

5.3 METHODS

The study was conducted between 1 January and 31 December 2007 in two high-incidence TB communities in Cape Town. Details of the study setting, data sources, case definition and record linking have been described elsewhere.[14] Briefly, the study area is 3.4 km^2 and has 36 343 inhabitants; each community has a PHC clinic where free TB treatment is available, with a TB treatment register located in each clinic (data source 1). Sputum samples collected at these clinics are transferred to a centralised laboratory in Green Point, Cape Town (data source 2). The area has a tertiary care hospital (Tygerberg Hospital), which also serves as the admission hospital for these communities. All sputum samples collected at Tygerberg Hospital are sent to the hospital laboratory (data source 3). Data were collected from 1 October 2006 to 31 March 2008 to correct for misclassification due to late laboratory results or registration. A bacteriologically confirmed pulmonary TB case was defined as an individual with at least two smears positive and/or at least one culture-positive sputum result.[21]

For record linking, Link Plus, a free probabilistic software package (Centers for Disease Control and Prevention, Atlanta, GA, USA),[22] was used, with name, surname, age and address as identifiers. Duplicate records in each register were removed. Matches and non-matches were manually reviewed for correctness. Near-matches were reviewed with the help of experienced research nurses who worked in the PHC clinics and knew the patients. Names and surnames were converted to New York State Identification and Intelligence System (NYSIIS), a variation of the Soundex phonetic coding system, reducing mismatching due to minor misspellings. Observed completeness of the TB treatment register was defined as the number of bacteriologically confirmed pulmonary TB cases, divided by the case-ascertainment after record linking, expressed as a percentage.

5.3.1 Capture-recapture analysis

The estimation of the total number of unobserved TB cases was calculated based on the distribution of observed TB cases in the three data sources after record linking. The independence of data sources and other assumptions underlying capture-recapture analysis have been described previously.[15,18] Interdependencies between the three TB-related registers are probable, positive or negative, causing possible bias in two-source capture-recapture estimates. Three-source log-linear capture-recapture analysis was employed to take possible interdependencies into account. The log-linear model with the lowest Akaike information criterion (AIC) was selected as the most valid model. Internal validity analysis was performed[23] using Chapmans's Nearly Unbiased Estimator.[24,25] Estimated completeness of the TB treatment register and case ascertainment after record linking was defined as the number of bacteriologically confirmed pulmonary TB cases in the TB treatment register or recorded in at least one register divided by the estimated number of bacteriologically confirmed pulmonary TB cases by capture-recapture analysis, expressed as a percentage. The 95% confidence interval (CI) for estimated completeness was also calculated based on the 95%CI of estimated number of cases.

We also estimated the number of TB cases with alternative estimators, the so-called truncated models, to cross-validate the selected log-linear model, as described previously.[26] We used a truncated binomial model, a truncated Poisson mixture model (Zelterman)[27] and a truncated Poisson model (Chao).[28] We chose this combination of truncated models as they have been described as an alternative to capture-recapture methods,[18,29] can be used on the same data needed for the three-source log-linear model and are easy to apply.[30]

Ethics approval was obtained from the Stellenbosch University Committee for Human Research. Permission for the study was granted by the City of Cape Town Health Directorate and the Department of Health, Western Cape Province.

5.4 RESULTS

5.4.1 Record linking

The overall number of bacteriologically confirmed pulmonary TB cases found in the three data sources was 306, with 243 cases recorded in the TB treatment register, 263 cases found in the centralised laboratory data source and 38 cases registered in the hospital laboratory data source

(Figure 5-1). Overlap was substantial between the TB treatment register and centralised laboratory (74%), but low between the TB treatment register and hospital laboratory (8%) and the centralised laboratory and hospital laboratory (3%). The total number of bacteriologically confirmed pulmonary TB cases not recorded in the TB treatment register was 63 (20%); 45 of the cases were observed in the centralised laboratory, 16 cases in the hospital laboratory and two cases in both laboratories. The observed completeness of the TB treatment register for bacteriologically confirmed pulmonary TB case was 79%.

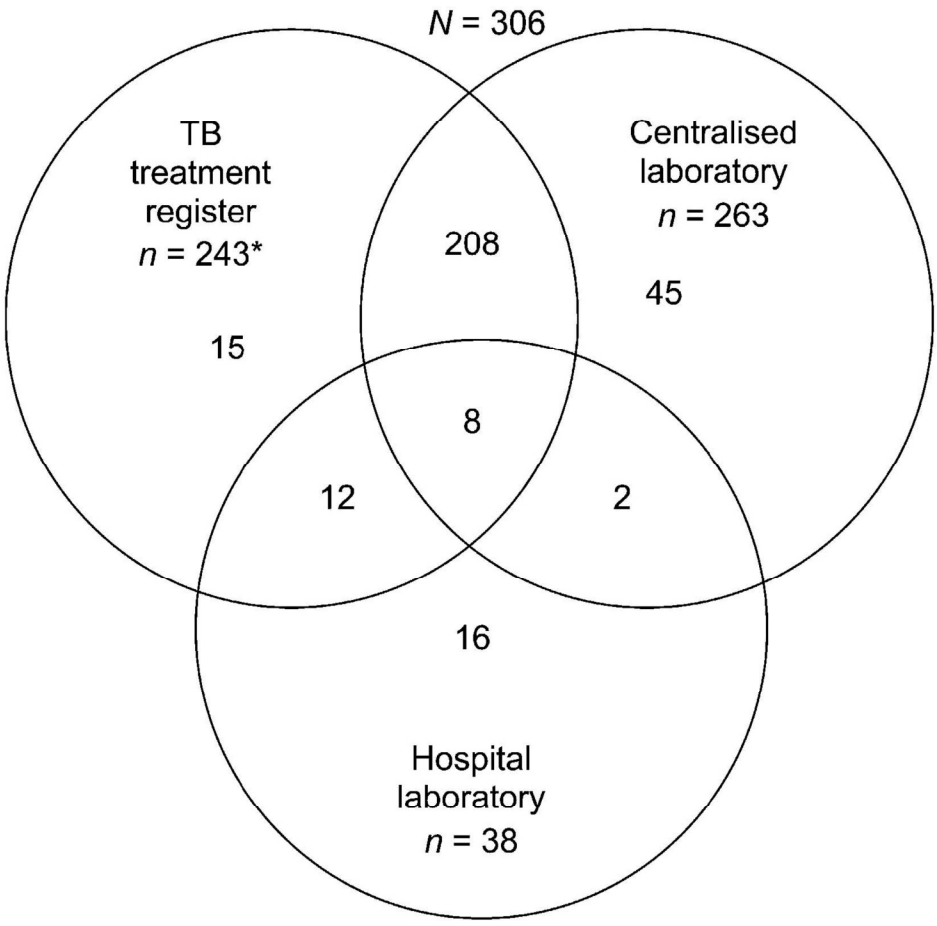

Figure 5-1: Distribution of bacteriologically confirmed TB cases after linking all three data sources.[14] * The number of TB cases in the TB treatment register is the final number of TB cases after results were corrected and added.
TB = tuberculosis.

5.4.2 Three-source capture-recapture analysis

The eight possible three-source log-linear capture-recapture models with likelihood ratio, AIC and estimated number of bacteriologically confirmed pulmonary TB cases with 95%CIs are presented in Table 5-1. The log-linear model selected based on the lowest AIC, apart from the main effects, was a parsimonious model incorporating only two interaction effects: the centralised laboratory and TB treatment register interaction and the centralised laboratory and hospital laboratory interaction. This model estimated the number of bacteriologically confirmed pulmonary TB cases not registered in any of the data sources at 20, resulting in a total number of bacteriologically confirmed pulmonary TB cases of 326 (95%CI 314–355). The completeness of registration of bacteriologically confirmed pulmonary TB cases was 75% (95%CI 68–77). The observed completeness is outside the 95%CI of the estimated completeness. The estimated completeness of case ascertainment after record linking was 94%.

Table 5-1: The eight possible three-source log-linear capture recapture models (N = 306 cases after record linking).

Model	G^2	*P* value	df	AIC	N_{est}	95% CI
Independent model	96.6	<0.001	3	90.6	317	313 - 323
Centralised laboratory × TB treatment register interaction	33,7	<0.001	2	29,7	501	407 - 681
Centralised laboratory × hospital laboratory interaction	18,9	<0.001	2	14,9	310	308 - 313
TB treatment × hospital laboratory interaction	84,1	<0.001	2	80,1	315	312 - 320
Centralised laboratory × TB treatment register, Centralised laboratory × hospital laboratory interactions	0,03	0,860	1	-1,97	326	314 - 355
TB treatment register × hospital laboratory, Centralised laboratory × hospital laboratory interactions	4,3	0,038	1	2,3	312	308 - 312
TB treatment register × hospital laboratory, Centralised laboratory × TB treatment register interactions	32,6	<0.001	1	30,6	666	386 - 1918
Saturated model	0,00	<0.001	0	0	321	309 - 389

G^2 = likelihood ratio; df = degrees of freedom; AIC = Akaike Information Criterion; Nest = estimated number of TB patients; CI = confi dence interval; TB = tuberculosis.

Table 5-2: Interval validity analysis through three two-source capture-recapture analyses.

Source 1	Source 2	Source 1 obs(*n*)	Source 2 obs(*n*)	Source 1, 2 obs(*n*)	Total obs(*n*)	est(*n*)	95% CI
Central Laboratory	TB treatment	263	243	216	290	296	290-301
Central Laboratory	Hospital Laboratory	263	38	10	291	935	560-1439
TB treatment	Hospital Laboratory	243	38	20	261	452	339-585

obs(*n*) = number of TB patients observed; est(*n*) = number of TB patients estimated; CI = confidence interval; TB = tuberculosis

Internal validity analysis through two-source capture-recapture analysis (Table 5-2) indicated some possible positive interaction between the TB treatment register and the centralised laboratory, because this estimate of 296 (95%CI 290–301) is slightly lower than the parsimonious capture-recapture model selected. The internal validity analysis further indicates a strong and very strong negative interaction between the almost mutually exclusive TB treatment and hospital laboratory registers and the centralised laboratory and hospital laboratory registers. This is shown by the considerably higher estimated number of bacteriologically confirmed pulmonary TB cases, at respectively 452 (95%CI 339–585) and 935 (95%CI 560–1439), with wide 95%CIs.

The truncated binomial, truncated Poisson heterogeneity and truncated Poisson mixture models estimated respectively 315 (95%CI 279–346), 319 (95%CI 283–355) and 307 (95%CI 273–341) bacteriologically confirmed pulmonary TB cases. Two-source capture-recapture analysis using the combined laboratory databases estimated 314 (95%CI 307–320) bacteriologically confirmed pulmonary TB cases, a completeness of registration of 78%.

5.5 DISCUSSION

Record linking between the TB treatment register and two laboratory data sources in the two communities in Cape Town identified the vast majority of the observed unregistered bacteriologically confirmed TB cases (n = 63). Capture-recapture estimated only 20 additional cases.

5.5.1 Capture-recapture assumptions and limitations

The assumptions that should be respected for valid capture-recapture analysis have been described elsewhere, in general[15] and in detail.[31] Violation of the perfect record linking assumption is

probably limited, as record linking between data sources was conducted through three different processes: use of probabilistic linking software, manual review of all matched and non-matched records and verification of near-matches by research nurses working in the PHC clinics. Phonetic coding reduced mismatching due to misspelling of names.

Violation of the closed population assumption is possible, as the study area is small and patients may have attended PHC clinics outside the study area or private practitioners (and private diagnostic services) in the study area, leading to underestimation of treatment recording, although private practitioners are required to refer patients to their local PHC clinics.

Violation of the homogeneous population assumption may exist but is assumed to be limited due to the small study area and free TB treatment. However, this assumption was not further examined, for example with stratified or covariate capture-recapture analysis.[32]

Violation of the independence assumption is expected to be limited, as three-source capture-recapture analysis was performed and a parsimonious model was selected, identifying and incorporating two relevant interactions. Internal validity analysis showed that these interactions were expected, namely, some positive interaction between the TB treatment and central laboratory registers and stronger negative interaction between the central and hospital registers. Elimination of duplicate cases prevented overestimation and, as only bacteriologically confirmed pulmonary TB cases were analysed, the number of false-positive cases is assumed to be limited.

In three-source capture-recapture analysis, sufficient data should preferably be collected in three fairly independent registers, approximately equal in number, each capturing more than 15% of cases and having sufficient overlap. In our study, the hospital laboratory register was much smaller than the other registers and the overlap between the hospital laboratory and the two other registers was poor. This had only a limited effect on the capture-recapture estimate as a two-source analysis of the TB treatment register linked with the two laboratory registers combined had a result just within the 95%CI of the three-source estimate, which also had a relatively small 95%CI. Examination of data 3 months before and after the study period reduced the number of cases that did not belong to the time period of the investigation.[14]

The truncated models for cross-validation produced results similar to the selected log-linear model. It has been postulated that independent and parsimonious three-source log-linear capture-recapture models are preferable but that truncated models can be used to identify possible failure in log-linear models.[26]

5.5.2 Other limitations

The two areas under study are small and have been the focus of a number of studies over the past two decades. This could mean that the results of this study cannot necessarily be extrapolated to other high incidence areas in South Africa, where, for example, due to lack of health education, to stigma and to distances to PHC clinics, initial default proportions could be much higher. A number of previous studies from Stellenbosch and Johannesburg have indicated that a considerable number of TB patients, many co-infected with HIV, die during hospital admission or soon afterwards, and are never recorded in the TB treatment register, although it is the duty of the hospital staff to notify the PHC clinic.[9] This issue might also contribute to an underestimation of TB-related mortality. When we excluded those cases only recorded in the hospital and those cases recorded in both laboratories who died during admission, the observed completeness of the TB treatment register increased from 79% to 84%. It cannot be ruled out that some of the cases only known to the central laboratory also died at home before seeking treatment at the PHC clinics, which would further increase the observed completeness of the TB treatment register.

The case definition of bacteriologically confirmed pulmonary TB was based on the 2004 South African NTP guidelines[21] as an individual with at least two sputum smears positive and/or at least one culture-positive sputum result. Chest X-ray abnormalities and clinical illness were not included in this case definition. This may lead to an underestimation of bacteriologically confirmed pulmonary TB, as cases with one sputum smear-positive result and radiological or clinical indications were not included. TB cases with negative sputum smear results but radiological abnormalities consistent with active TB and fulfilling the 2004 South African NTP guideline criteria for a bacteriologically negative TB case were also not included. This study did not observe or estimate extra-pulmonary TB cases. Patients with a zero probability of access to health care are not included in the estimate of TB patients; however, in this study area this is assumed to be limited.

5.5.3 Initial treatment default

The establishment of a TB care centre in the Chris Hani Baragwanath Hospital in Johannesburg considerably improved the proportion of patients being recorded and successfully referred to PHC clinics after TB suspects were given health education.[33] Simple health system interventions could reduce the number of initial treatment defaulters, for example by improved completion of sputum request forms by the PHC clinics, especially the address details of the TB suspects to facilitate

follow-up, and by adequate return of the results by the laboratories, e.g., implementing an integrated computerised TB treatment and results management system. Possible reasons for initial treatment default could be health system related,[11] or due to a lack of knowledge of TB and fear of stigmatisation by the community, including being considered HIV-positive.[34,35] A combination of interventions is currently being tested in the study area.

5.6 CONCLUSION

Capture-recapture can be useful for evaluation of TB control programmes, including in resource-limited settings, but methodology and results should be carefully assessed. Interventions are needed to increase the low completeness of registration and to reduce the high number of initial defaulters, as this problem is of great concern and could be worse elsewhere in South Africa.

5.7 ACKNOWLEDGEMENTS

The authors express their appreciation to all the contributors to this study: staff involved with the national, provincial and district TB programmes, and staff at Ravensmead and Uitsig clinics. They also thank the National Health Laboratory Service Corporate Data Warehouse for assistance with data extraction.

5.8 REFERENCES TO CHAPTER 5

1. Corbett EL, Watt CJ, Walker N, et al. The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. Arch Intern Med 2003;163:1009-21.
2. Davies PD. The world-wide increase in tuberculosis: how demographic changes, HIV infection and increasing numbers in poverty are increasing tuberculosis. Ann Med 2003;35:235-43.
3. World Health Organization. Global tuberculosis control : epidemiology, planning, financing : WHO report 2009. Geneva: World Health Organization; 2009.WHO/CDS/TB/2003.313, 303 pp.
4. World Health Organization. Resolution WHA 44.8 of the forty-fourth World Health Assembly. WHA44/1991/REC/1. Geneva, Switzerland: World Health Organization; 1991.WHO/GPA/INF/89.10,
5. Dye C, Scheele S, Dolin P, Pathania V, Raviglione MC. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA 1999;282:677-86.
6. Borgdorff MW. New measurable indicator for tuberculosis case detection. Emerg Infect Dis 2004;10:1523-8.
7. van der Werf MJ, Borgdorff MW. Targets for tuberculosis control: how confident can we be about the data? Bull World Health Organ 2007;85:370-6.
8. Masjedi MR, Fadaizadeh L, Taghizadeh Asl R. Notification of patients with tuberculosis detected in the private sector, Tehran, Iran. Int J Tuberc Lung Dis 2007;11:882-6.
9. Edginton ME, Wong ML, Phofa R, Mahlaba D, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: numbers of patients diagnosed and outcomes of referrals to district clinics. Int J Tuberc Lung Dis 2005;9:398-402.
10. Botha E, den Boon S, Lawrence KA, et al. From suspect to patient: tuberculosis diagnosis and treatment initiation in health facilities in South Africa. Int J Tuberc Lung Dis 2008;12:936-41.
11. Botha E, Den Boon S, Verver S, et al. Initial default from tuberculosis treatment: how often does it happen and what are the reasons? Int J Tuberc Lung Dis 2008;12:820-3.
12. Migliori GB, Spanevello A, Ballardini L, et al. 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.

13. Mukerjee AK. Ascertainment of non-respiratory tuberculosis in five boroughs by comparison of multiple data sources. Commun Dis Public Health 1999;2:143-4.
14. Dunbar R, Lawrence K, Verver S, et al. Accuracy and completeness of recording of confirmed tuberculosis in two South African communities. Int J Tuberc Lung Dis 2011;15:337-43.
15. Capture-recapture and multiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995;142:1047-58.
16. Laska EM. The use of capture-recapture methods in public health. Bull World Health Organ 2002;80:845.
17. Fienberg SE. The multiple recapture census for closed populations incomplete 2K contingency tables. Biometrika 1972;59:591-603.
18. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17:243-64.
19. Capture-recapture and multiple-record systems estimation II: Applications in human diseases. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995;142:1059-68.
20. Bassili A, Grant AD, El-Mohgazy E, et al. Estimating tuberculosis case detection rate in resource-limited countries: a capture-recapture study in Egypt. Int J Tuberc Lung Dis 2010;14:727-32.
21. National Department of Health. The South African National Tuberculosis Control Programme: Practical Guidelines. Pretoria: National Department of Health; 2004.
22. Division of Cancer Prevention and Control. Registry Plus™ Link Plus. In. 2.0 ed. Atlanta: Centers for Disease Control and Prevention; 2007.
23. Hook EB, Regal RR. Internal validity analysis: a method for adjusting capture-recapture estimates of prevalence. Am J Epidemiol 1995;142:S48-52.
24. Chapman CJ. Some properties of the hypergeometric distribution with applications to zoological censuses.: U California Public Stat 1951131-60
25. Wittes JT. On the bias and estimated variance of Chapman's two-sample capture-recapture estimate. Biometrics 1972:592-7.
26. 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;136:14-22.

27. Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. J Stat Plan Inf 1988;18:225-37.
28. Chao A. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 1987;43:783-91.
29. Hook EB, Regal RR. Validity of Bernoulli census, log-linear, and truncated binomial models for correcting for underestimates in prevalence studies. Am J Epidemiol 1982;116:168-76.
30. Van der Heijden PG, Cruyff MJ, Van Houwelingen H. Estimating the size of a criminal population from police registrations using the truncated Poisson regression model. Stat Neerl 2003;57:289-304.
31. Brenner H. Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. Epidemiology 1995;6:42-8.
32. Hook EB, Regal RR. Effect of variation in probability of ascertainment by sources ("variable catchability") upon "capture-recapture" estimates of prevalence. Am J Epidemiol 1993;137:1148-66.
33. Edginton ME, Wong ML, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: an intervention to improve patient referrals to district clinics. Int J Tuberc Lung Dis 2006;10:1018-22.
34. Westaway MS, Wolmarans L. Cognitive and affective reactions of black urban South Africans towards tuberculosis. Tuber Lung Dis 1994;75:447-53.
35. Cramm JM, Finkenflugel HJ, Moller V, Nieboer AP. TB treatment initiation and adherence in a South African community influenced more by perceptions than by knowledge of tuberculosis. BMC Public Health 2010;10:72.

CHAPTER 6: OVERALL CONCLUSIONS AND RECOMMENDATIONS

6.1 OVERALL CONCLUSIONS

Epidemiology is focused mainly on the study of disease in large groups of people or populations.[1] It thus differs from the more conventional clinical approaches where disease processes in affected individuals are the main concern. While the objective of the latter is to treat diseases in individuals already affected, epidemiology is basically concerned with the interplay between causative factors and vulnerabilities in individuals who became diseased.

Epidemiology is considered a basic science of public health and is involved in the promotion and protection of the health of the public. Thus the study of health events, health characteristics, or patterns of health determinants in a community is of crucial concern to the health services in any country.

Globally, in TB control programmes, health events, health characteristics, or patterns of health determinants are measured through the use of routinely collected data by means of a recording and reporting system. The system of routinely collected TB data is an essential tool for (1) establishing the burden of TB in a country, (2) managing the TB programme and (3) identifying those communities with gaps in their TB programme and the management system.[2-4] Once gaps and challenges have been identified, health services can be strengthened in either the whole country or in specific vulnerable communities through targeted interventions. The appropriate allocation of resources to address these challenges can then be based on verifiable information. However, this system can only work optimally if the routinely collected data is accurate and complete. It is well known that there are often difficulties with the accuracy and completeness of routinely collected data[1] and therefore more emphasis should be placed on the accuracy and completeness of such data, especially if the data are used for important decisions and recommendations.

According to the WHO[3] the basic information sources and systems used to record and report TB cases in each country should include:

1. A laboratory register which should be maintained at primary health care centres and all samples examined for TB using onsite microscopy should be recorded in this register.
2. A TB treatment register in which all cases placed on treatment are recorded.

3. A TB treatment card on which all information about treatment is recorded.

Table 6-1: WHO recommended TB recording and reporting system compared to South African NTP recording and reporting system.

Recommended by World Health Organisation	South African National TB Programme approach
At health facility in central, regional or peripheral level:	
TB laboratory register	Not implemented
TB treatment register	TB treatment register
TB treatment card	TB treatment card
TB treatment referral and/transfer	TB treatment referral and/transfer
At Basic Management Unit level:	
TB laboratory register	Not implemented
TB treatment register	TB treatment register (implemented as electronic TB register)

The work for this thesis focused on the integration of the laboratory register and the TB treatment register as it is essential that all individuals diagnosed with TB should start treatment and these persons' names should appear in the TB treatment register. This is usually not a difficult task in most countries because the laboratory register and the TB treatment register are both located in the clinic where the diagnosis of TB is made and where the individual will receive treatment.

However, in South Africa there is a centralised laboratory service (NHLS) with a complex system of sending samples from the clinic to the laboratory and sending the results from the laboratory back to the clinic. In this complex system there are many potential gaps where samples or results can get lost. In addition, in South Africa there is a complicated system of collating the data from the TB treatment register. Paper-based TB treatment registers are located at PHC clinics, TB hospitals and correctional facilities where TB patients receive treatment and have their names recorded. These paper-based registers are used to manage the treatment of cases and to track and record the treatment outcome of each case. These TB treatment registers are updated on a daily basis by TB staff at the PHC facilities by entering new TB cases and noting the ongoing progress of TB cases already on treatment. As pages from the TB treatment registers are completed these pages are sent to the sub-district office where the data is captured into the electronic TB register

(ETR). The data from the ETR is sent electronically from the sub-district or district office to the Provincial Department of Health where further analyses are conducted and the data is aggregated before being sent to the National Department of Health. Aggregate data from the ETR is also sent to the District Health Information System and the Notifiable Medical Disease System and then incorporated into national disease notification data. Data sent to the National Department of Health is used as a source for completing the WHO standard collection forms.

The work for this thesis built on research conducted between 2004 and 2005 at 13 primary health care facilities in the Stellenbosch district of the Western Cape.[5,6] That study indicated that 17% of TB cases with two smear positive results were not recorded in the TB treatment register.[5] These patients accessed health care facilities and were diagnosed but never started treatment and were referred to as 'initial defaulters'. A follow-up study was conducted in 11 of the 13 primary health care facilities in order to determine what happened to these 'initial defaulters'.[6] Of the 58 initial defaulters 14 (24%) had died. Of the 44 remaining initial defaulters, 26 could not be found due to address-related reasons. The remaining 18 initial defaulters (41%) were found and interviewed. Reported reasons for not starting treatment were directly linked to services in 56% of cases and not directly linked to service in 44% of cases.

The study area selected for the present study has been the focus of various TB related research studies over the last decade and therefore it seemed reasonable to assume that the recording and reporting of TB cases in this area would be considerably better than in other communities in Cape Town. The research attention this area received and the close interaction between research staff and PHC workers should therefore have instilled better attainment due to increased attention and quality control. This however, did not materialise! The results presented in this thesis highlighted the concern regarding the accuracy and completeness of routinely collected TB recording and reporting data. It was clearly demonstrated through the work done for this thesis, that there is still a huge challenge with the accuracy and completeness of the data which leads to underreporting of cases in this area. The challenges identified in this study and the questions raised by this work can give broad indications of the extent of the problem of underreporting of TB in South Africa.

The implications of poor accuracy and completeness of data leading to underreporting of TB cases are discussed in relation to the implications for the National TB Programme, the individual and the community. In addition, the cost implications for patients and the NTP are briefly discussed.

6.1.1 Implications for the National TB Programme:

Accurate management information is an essential component of any effective control programme.[7] In this way, in the NTP, good quality routine data is required to control TB effectively and to implement appropriate intervention strategies. The data is also needed to determine the progress towards the 70% case detection rate and 85% cure rate set by the Stop TB strategy.[8] Routinely collected data from the TB treatment registers which give information about TB case finding and treatment outcome provide the information for these assessments. If however the routine data are not accurate or complete, TB cases are underreported and do not represent the true situation of TB in a community, district, province or country. The decisions made regarding health care, resource allocation and service delivery on a community, district, provincial and national level would be incorrect.

This could compound the already unequal distribution of health care and resources in the country by not directing sufficient health care resources to where it is needed most. Health care services in South Africa are already strained due to the lack of qualified staff and inadequate health care facilities. The PHC clinics with poor quality routinely collected data would be placed under even more pressure due to inadequate allocation of staff, equipment, drugs and resources. Service delivery to the individual patients and community would therefore deteriorate due to the number of additional cases that the system is unaware of and which will place the health care workers and services under increased pressure. Health care workers will have less time to deliver appropriate health care and complete patient files and the required NTP documents (sputum request forms, TB treatment register, etc.). This in turn will decrease the accuracy and completeness of recording and reporting and increase the number of TB cases not recorded and reported and the vicious circle will continue.

6.1.2 Implications for the individual:

Underreporting of TB cases and not starting those diagnosed with TB (especially smear positive TB cases) on treatment does not only have an impact on the TB statistics. There are also serious implications to the individual who is diagnosed with TB and not subsequently placed on treatment as quickly and efficiently as possible. If this individual stays untreated he/she could progressively develop more serious TB disease and possibly die from this treatable disease.

6.1.3 Implications for the community:

Apart from the implications to the individual, there are also implications for the community, because if individuals diagnosed with TB are not placed on treatment as soon as possible, they will continue transmitting TB. The longer it takes to place an individual with TB on treatment the longer he/she will stay infectious and the more chance the untreated individual will have to contribute to the continued exposure and transmission of TB in the community. Many of the TB cases in high burden countries, including South Africa, are young people and often young parents of small children. Children are often exposed to TB in their homes and therefore finding and treating smear positive adult TB cases remains the best approach to controlling and preventing TB, especially childhood TB.[2,3]

6.1.4 Cost implications for the patient:

TB places an extraordinary burden on sufferers, their families, and communities and on government budgets. The greatest burden of TB falls on productive adults who are often unable to work. The burden of taking care of sick individuals usually falls on other family members. In addition to putting them at greater risk of developing TB themselves, care giving can lower their productivity. Besides loss of employment, the cost of having TB can also be significant. Even though TB care is "free of charge" to the individual, it has clearly been shown that there is enormous cost incurred by those on TB treatment through transport costs, days of work lost etc.[9,10]

6.1.5 Cost implications for the National TB Programme:

The Department of Health financial resources are wasted when cases are diagnosed but left untreated as the tests have been done and paid for without acting on the results and without treating the patients. The NHLS is a parastatal company and not integrated in the NTP and the Department of Health. The Department of Health is billed for all TB tests, even those tests not acted on by the PHC. This is unproductive expenditure to the Department of Health and because each Province is allocated a limited budget for specific use, each Province needs to use its available funds sparingly.

Currently the TB recording and reporting system and the NTP have too many gaps where individuals diagnosed with TB could be and are lost to the system and therefore lost to care. The study described in the preceding chapters has identified some alarming facts regarding the recording and reporting of TB cases and the accuracy and completeness of the related data. Some

already established issues identified by previous studies[5,6,11] regarding the underreporting of TB cases have also been re-emphasised. A number of studies have already indicated that underreporting of TB cases in TB recording and reporting systems is a concern.[12-14] The WHO has also identified the need to evaluate the accuracy and completeness of the recording and reporting of TB and has therefore developed a policy and recommendation for how to assess the epidemiological burden of TB and the impact of TB control.[4] In South Africa this policy and recommendations have however not yet been implemented by the NTP.

The present study (Chapter 5)[15] used a WHO recommended approach, the capture-recapture methodology, to estimate the number of TB cases not recorded and reported by the NTP system. Due to the fact that no previous capture-recapture study has ever been conducted in South Africa on the available NTP data sources, it was decided to assess the feasibility of conducting such a study. Before any capture-recapture study can however be considered it was necessary to first assess the accuracy and completeness of the proposed data sources to be used. In order to assess the accuracy and completeness of the data sources a record linking study was carried out (Chapter 4).[16]

6.1.6 Record linking

The accuracy and completeness of the data sources used for record linking were assessed before performing the record linking. This is essential for the linking process as a high level of accurate and complete data, especially on the data fields used for linking, will increase the number of individuals whose data from the one data source can be accurately matched with the data from a different data source (true matches). In the absence of a common unique identifier (e.g. a unique identification number) between all data sources, the four most essential data fields for a successful record linking study is surname, name, gender and date of birth. If date of birth is however not available the age could be used. Using age does not however produce the best record linking results as age changes over time. For any study where record linking will be used and the data sources are deemed to be inadequate on the primary linking data fields, the researcher should carefully consider if it is worth continuing with the study as the effort needed to conduct the record linking could outweigh the benefit of conducting the study. This is especially true when data sources with a large number of records are considered.

Record linking was used to determine the accuracy and completeness of (1) the results received from NHLS and subsequently recorded in the TB treatment registers and (2) the accuracy and

completeness of case registration (Chapter 4).[16] This was done through record linking of the TB treatment registers and the NHLS database for centralised and hospital laboratories. Record linking is a valuable tool for the use in healthcare research and epidemiology, especially when utilising data residing in different databases. Record linking provides an opportunity to effectively link these databases to provide a more accurate and complete combined data source. Data fields from the separate data sources could be added to the final data source. Data fields available across the separate data sources can be verified against each other for accuracy and completeness. Additional cases can also be identified in the separate data sources resulting in more complete case recording.

For the present study the data sources were deemed to be adequate with an accuracy and completeness of 89% in the TB treatment register, 81% in the centralised laboratory and 99% in the hospital laboratory. The centralised laboratory data source had the highest level of incomplete data on date of birth and gender (Chapter 4).[16] The TB treatment register did not have a date of birth field and therefore age had to be calculated in the laboratory data source in order to have a common age field in all 3 data sources. The above limitations of the data sources increased the number of records that needed to be manually reviewed in order to verify true matches.

After record linking, date of birth was added to the TB treatment register as an additional data field from the laboratory data. It was also possible to verify and add additional sputum results from both the laboratory data sources. By doing this an additional 42 bacteriologically confirmed TB cases were identified in the TB treatment register which were recorded as bacteriologically unconfirmed before record linking (Chapter 4).[16] There are a number of reasons why these results were possibly not recorded in the TB treatment register with the following being the most probable reasons:[5,6,11]

- The laboratory did not send the result to the PHC clinics and the health care worker did not follow up the unresolved result with the laboratory.
- The laboratory did send the result to the PHC clinic and the result was received by the health care worker, but the health care worker did not recall the client and therefore treatment was not started. The results were therefore also not recorded into the TB treatment register or the results were recorded incorrectly.
- The sputum request form sent to the laboratory with the individual's sputum was not completed adequately. The result returned to the PHC clinics would therefore not have adequate information to be associated with the person to whom the result pertained.

Correcting 10 results in the TB treatment register from the laboratory data resulted in 3 bacteriologically confirmed TB cases being reclassified as bacteriologically unconfirmed and therefore excluded from the study. On the other hand, 3 bacteriologically unconfirmed TB cases were reclassified as bacteriologically confirmed and therefore added to the study. With the additional 115 results added to the TB treatment register from the laboratory data, one EPTB case was reclassified as a PTB case and 39 bacteriologically unconfirmed TB cases were reclassified as bacteriologically confirmed. These 42 reclassified cases were already in the TB treatment register and therefore on TB treatment and would not contribute to any further transmission of TB in the community. The treatment for these cases would also not have to be changed or reassessed due to the reclassification.

These reclassified cases would however have an effect on the TB statistics for these communities. The key indicator for measuring the success of the NTP is to monitor the treatment outcome of smear positive TB cases, especially new smear positive cases. When the TB treatment registers located at the PHC clinic are sent to the district office in order to be captured into the ETR, the errors in the TB treatment register will be carried forward into the ETR. Laboratory results not recorded, or recorded incorrectly, in the TB treatment register will also be missing or incorrect in the ETR. The total number of smear positive cases would therefore be underrepresented in the ETR data.

An additional 63 TB cases were identified in the two laboratory data sources which were not recorded in the TB treatment register - 45 (71%) from the centralised laboratory, 16 (25%) from the hospital laboratory and 2 (3%) that were in both laboratory data sources. For the centralised laboratory the 45 cases comprised 17% of the total 263 bacteriologically confirmed TB cases identified from the centralised laboratory and the 16 cases identified from the hospital laboratory comprised 42% of the total number of cases indentified from the hospital laboratory. These TB cases were not on treatment in the clinic situated closest to their residential address. It is possible that some of the 45 additional TB cases identified from the centralised laboratory were treated at a PHC clinic outside the study area and were therefore not recorded in the TB treatment registers corresponding to their address. The 16 cases identified in the hospital laboratory could have completed treatment in the hospital or could have died in hospital. These cases could also have been referred to a PHC clinic outside the study area and therefore commenced treatment outside the study area. The policy in Cape Town is to treat patients at any clinic, the clinic of the patient's choice. This however, seldom happens and most patients are treated in the clinic closest to where

they reside. Some of these cases could also have died in the hospital or died before treatment started at the PHC clinics to which they were referred. It would be possible to determine if any of the additional cases were placed on treatment at a PHC clinic outside the study area by looking at ETR data of the sub-district or of the whole of Cape Town. Mortality data could be used to determine if any of the additional cases died. This data was however not available for this study.

6.1.7 Capture-recapture

After record linking, the next logical step is to conduct a capture-recapture analysis. This component of the present study aimed to determine if a capture-recapture analysis would be a feasible technique for the proposed study area and if the data sources available would produce a valid capture-recapture estimation of TB in these two communities. For any capture-recapture analysis careful attention should be giving to the underlying requirements for a valid capture-recapture study.[17,18] In the current study a three-source log-linear capture-recapture model was used as well as a two-source capture-recapture model. The two-source capture-recapture was used to assess the internal validity of the three-source log-linear estimate and to assess the nature of dependence between the three data sources.

The underlying requirements for capture-recapture were discussed in chapter 5[15], but the ways they impacted on the present study are briefly discussed.

- Violation of the perfect record linking requirement: This was limited in this study as record linking between data sources was conducted through three different processes: use of probabilistic linking software, manual review of all matched and non-matched records and verification of near-matches by research nurses working in the PHC clinics. Phonetic coding reduced mismatching due to misspelling of names.
- Violation of the closed population requirement: This requirement was not examined in detail. However after discussions with research nurses and other researchers working in the study area it was concluded that individuals frequently change residence within the study area to another residence within the same area. Some individuals could have moved out of the study area but the frequency of this is not known, but is considered to be limited. A further unexplored limitation was that patients may have attended PHC clinics outside the study area or private practitioners in the study area, leading to underestimation of treatment recording, although private practitioners are required to refer patients to their local PHC clinics. This would however lead to an underestimation in the capture-recapture analysis.

- Violation of the homogeneous population requirement: This may exist but is assumed to be limited due to the small study area and free TB treatment. However, this requirement was not further examined, for example with stratified or covariate capture-recapture analysis.
- Violation of the independence requirement: This is expected to be limited, as three-source capture-recapture analysis was performed and a parsimonious model was selected, identifying and incorporating two relevant interactions. Internal validity analysis showed that these interactions were expected, namely, some positive interaction between the TB treatment register and central laboratory registers and stronger negative interaction between the central and hospital registers.
- Violation of including false-positive cases and duplicate cases: Elimination of duplicate cases prevented overestimation and, as only bacteriologically confirmed pulmonary TB cases were analysed, the number of false-positive cases is assumed to be limited. Examination of data 3 months before and after the study period reduced the number of cases that did not belong to the time period of the investigation.
- Violation of adequate overlap between data sources: In three-source capture-recapture analysis, sufficient data should preferably be collected in three fairly independent registers, approximately equal in number, each capturing more than 15% of cases and having sufficient overlap. In our study, the hospital laboratory register was much smaller than the other registers and the overlap between the hospital laboratory and the two other registers was poor. This had only a limited effect on the capture-recapture estimate as a two-source analysis of the TB treatment register linked with the two laboratory registers combined had a result just within the 95%CI of the three-source estimate, which also had a relatively small 95%CI.

This study indicated that a capture-recapture analysis for these two communities, and on the available data sources, is feasible. The 20 additional cases estimated by means of capture-recapture are presumed to be in line with the number of TB cases not recorded in the NTP data sources. The estimate was discussed with research nurses who have worked in these communities for many years and confirmed by fellow researchers at the Desmond Tutu TB Centre. We cannot however determine or validate the capture-recapture estimate as being 100% correct. We know of no other capture-recapture studies conducted in a similar setting in order to compare capture-recapture estimates. The most obvious difference between the setting in this study and other studies is the centralised nature of the laboratory services in South Africa.

In the current study (Chapter 5)[15] the total number of TB cases identified as being not reported and recorded by means of capture-recapture was estimated as 20 TB cases and by record linking as 63 cases (Chapter 5).[15] The current study showed what a huge benefit record linking, even without capture-recapture, could be in identifying issues with recording and reporting systems. It is therefore not always necessary to conduct a capture-recapture analysis, but is still useful to identify any additional cases that would normally not be identified by record linking alone.

The current study identified some major concerns with the recording and reporting system in these two communities. The completeness of recording of bacteriologically confirmed pulmonary TB was 79% in the TB treatment register. After the capture-recapture estimate the completeness of registration of bacteriologically confirmed pulmonary TB cases was only 75% (Chapter 5).[15] If the finding of this study is seen as a best case scenario, the NTP TB recording and reporting system only represents 75% of the TB cases in our community. If we extrapolate these findings to a national level then the South African NTP recording and reporting system misses 115 000 TB cases for the country per year.

6.1.8 A word of warning!

The use of record linking and capture recapture techniques is not a quick fix for the shortcomings in the NTP recording and reporting system. Such techniques can only identify if and where any problems exist. It is ultimately the responsibility of the NTP to address the gaps within their system.

6.2 WHERE DOES THE SYSTEM BREAK DOWN?

In South Africa the diagnostic services for TB in the NTP are centralised, not delivered as a point of care diagnostic service and not integrated within the NTP. In other countries smear microscopy is a point-of-care service located at PHC clinics. The WHO recommends that laboratory services be integrated within the NTP.[19] The most obvious reason for the underreporting of TB cases in the NTP seems to be the centralised laboratory services. A second reason is the diagnosis of TB cases in the hospital and the referral system from hospital to PHC clinics and the subsequent recording of cases at the PHC clinic in the TB treatment register. Studies[5,6,11,12,20] have already documented

the breakdown in the referral system and clearly interventions are needed to address these concerns.

The additional TB cases identified through the linking of data sources were not recorded and reported on, but have accessed health services, and were in fact diagnosed with TB, although no action was taken on the results. These cases should have been recorded and reported, and should have been on TB treatment. The basic data management skills and the importance of accurate and complete recording and reporting seem to be lacking at all levels of the NTP. There also seems to be no accountability when it comes to the lack of accuracy and completeness of the data in the recording and reporting system. This lack of accountability appears to be widespread throughout the NTP structure, but no direct substantiation of this is available. This obviously needs to be studied further.

An important point however is the responsibility borne by the TB cases themselves. Health care workers can make all the efforts to identify TB suspects and diagnosed TB cases, but it is still the responsibility of the individual to present back to the PHC clinic for treatment. Some individuals may even refuse to commence treatment and health care workers have little or no authority to force TB cases to start treatment. It is also the responsibility of each individual accessing the PHC clinic to provide the health care worker with correct personal and demographic information. If however the individual provides the health care worker with an incorrect residential address for instance, there will be no way to follow the individual up in the community in order to start them on TB treatment. Some individuals may provide inaccurate information intentionally as they do not wish the health care workers to know where they live, possibly due to stigma, mistrust or fear. Some individuals may also provide inaccurate information due to illiteracy. Some individuals may not know some of their personal details, for instance in cases where parents did not register their birth and they therefore may not be sure of their accurate birth date.

6.2.1 Model for change

Even though there are concerns regarding the NTP recording and reporting system, it is still a functional concept. There are some advocates who are in favour of the abolishment of the NTP recording and reporting system since it is not accurate. This would however be a retrograde step and not a solution to the poor functioning of the system. The WHO structure and recommendations have been successfully implemented in many countries and have been trialled and tested over many years. The way in which the NTP recording and reporting system in South Africa has been

implemented, monitored and evaluated is however not adequate and needs to be improved. The only way this can be corrected though is by approaching all stakeholders and working towards a functional solution together. The usual approach by many is however to identify shortcomings within departments or systems and to place blame. This approach only alienates those responsible, especially governmental departments, and finding working solutions becomes much harder.

Building a trusting and working relationship with all the stakeholders does take a long time and could be very frustrating. It is however worth the effort as working against or without the stakeholders could delay a solution to the problem even further. The finding from this study, and others[5,6], have identified that a problem exists in the recording and reporting of TB cases. This has led to discussions with City of Cape Town Health Directorate, the Western Cape Health Department and the National Department of Health, as well as NHLS, in order to find a solution to the problem. These discussions are continuing and there are monthly meetings attended by colleagues from City of Cape Town Health Directorate, the Western Cape Health Department and Desmond Tutu TB Centre.

6.3 WHAT THIS WORK HAS LED TOO

Collaboration

- The Desmond Tutu TB Centre attends a monthly government partners meeting where various health, operational research and academic research issues are discussed.
- The Desmond Tutu TB Centre and NHLS meet together on the Line Probe Assay (LPA) Steering Committee. It is hoped that through this Steering Committee for the LPA study, which aims to investigate the impact of the molecular diagnostic method for TB, many of the challenges identified in the work done for this thesis, can be addressed.
- The National Department of Health has requested the Desmond Tutu TB Centre to assist with monitoring and evaluation of their health information systems.

Other studies

- LPA Study: This study aims to assess the impact of improved TB and MDR-TB diagnosis at primary health care facilities in Cape Town. The current routine NTP systems and data will be strengthened through monitoring and evaluation. The study proposes to establish monitoring and evaluation systems which could be used to evaluate the new changed diagnostic

algorithms and to set systems in place that will ensure that all cases diagnosed, will start treatment.

- Initial defaulter study: This study aims to identify the initial default rate (individuals diagnosed but not started on treatment) at primary health care facilities in 5 provinces in South Africa. This study also aims to identify facility and individual risk factors for high initial default rates at primary health care facilities.
- All the above studies are done in close collaboration with government partners and NHLS and the final aim is to change the way in which the NHLS and NTP integrate results, thereby improving care to patients.

6.4 OVERALL RECOMMENDATIONS

Table 6-2: Summary of recommendations.

1	NTP should engage with experts in order to improve health information systems.
2	Develop systems to make sure those diagnosed with TB are placed on treatment.
3	Strengthen referral systems between the hospital and PHC clinics.
4	Train health care workers.
5	Monitor and evaluate, and responsibility.
6	Incorporate record linking as part of the monitoring and evaluation process.
7	Develop and validate effective record linking strategies.
8	Make record linking user-friendly.
9	Integrate health information systems.
10	Verify that non-reported cases are truly not reported.
11	Find out what happened to those not recorded.

Recommendation 1: The NTP should engage with experts in order to improve health information systems.

Background:

Various concerns have been raised regarding the accuracy and completeness of health information systems in all levels of the NTP. The work done for this thesis clearly identified that there are major concerns with the accuracy and completeness of data and that many individuals are diagnosed with TB but never start treatment.

Recommendation:

The NTP at national, provincial and district level should engage with experts in order to identify and correct limitations in current health information systems. These experts should include researchers familiar with the field of TB, epidemiologists and application/data systems engineers. An initiative to improve information systems and thereby improve the routinely collected data might have implications for NTP budget, skills training etc.

Recommendation 2: Develop systems to make sure those diagnosed with TB are placed on treatment.

Background:

This study, and others, has identified that many people access health care services and are subsequently diagnosed with TB but are not always placed on TB treatment.

Recommendation:

Systems should be developed in order to make sure that those diagnosed with TB receive their result and are placed on TB treatment in a timely manner. A PEPFAR funded study managed by the Desmond Tutu TB Centre, has already been implemented and is currently being tested in one sub-district in Cape Town. The initial findings from this study is promising and once finalised could be extended to the rest of South Africa. The PEPFAR study receives all TB results electronically from NHLS on a daily basis. All positive TB diagnostic results are reported to the district NTP office. A TB coordinator at the district office follows these positive results up with the PHC clinics in order to determine if the individuals to whom the results belong started TB treatment. A weekly report is then sent back to the research office indicating if all positive results have been acted on. If a result has not been acted on the coordinator indicates the reason for this. This system will be improved and evaluated and once finalised, will be presented to the Department of Health and NTP for possible implementation in other sub-districts and perhaps in other programmes as well.

Recommendation 3: Strengthen referral systems between the hospital and PHC clinics.

Background:

TB cases diagnosed in the hospital are often not recorded in the TB treatment register. These TB cases could have completed treatment in full at the hospital, could have been referred to a PHC clinic or have died in hospital.

Recommendation:

It is recommended to implement a similar system of TB care, referral and education as was successfully implemented at Chris Hani Baragwanath Hospital.[20] A dedicated TB care centre was established in the hospital where those diagnosed with TB are referred to before being discharged from hospital. An electronic TB register was implemented in the hospital to record TB cases. Patients also received education regarding TB, HIV and the location of available PHC clinics in their district. Patients referred to PHC clinics were followed up to make sure they continued their treatment and were recorded at the PHC clinic. The data from the work done for this thesis will be presented to the Cape Town City Health Directorate and it will be recommended that a similar approach be followed. The possibility of establishing a TB Treatment register in the hospital will be discussed.

Recommendation 4: Training health care workers.

Background:

It is imperative for those responsible at any level of data processing to know what the data they are responsible for means and how to interpret the data. Currently very little or no basic data management or analysis skills seem to be present in most levels of the NTP. There also seems to be a shortage of basic computer skills among health care workers.

Recommendation:

The Desmond Tutu TB Centre can facilitate discussions about the training of health care workers to develop a training programme of the skills necessary for accurate and complete recording of TB data. Basic data management and analysis training should be available to key health care personnel and NTP staff. This also includes basic computer skills training. It would be advisable to identify staff within health services, for instance the head TB nurse, to be trained. It would not be

financially viable for the Department of Health to train all staff. Training in data handling would place the responsibility of the data into the health care worker's hands and would also increase their knowledge and understanding of the data.

The personnel selected to be trained should also be encouraged to regularly analyse the data for their own PHC clinics and discuss their findings with the rest of the staff at the facility on a regular basis. The analysis and discussions should however not only focus on the number of cases recorded and successfully treated, but should also include:

- Incomplete patient and TB treatment register records.
- Missing laboratory results.
- TB cases that were not followed up and therefore not placed on treatment.

Health care workers should also be trained on how to educate patients on general health care and TB/HIV. The patient education should also include the effects and consequences of their illnesses, especially if they are not treated appropriately. Some training material is available as part of the DOTS strategy but it is not always used or adhered to by the health care workers.

Recommendation 5: Monitoring and evaluation, and responsibility.

Background:

Any health information system should have a monitoring and evaluation system. However any monitoring and evaluation system is only as good as its implementation. There is currently a monitoring and evaluation system in the South African NTP[21] but its performance is unknown.

Recommendation:

The current monitoring and evaluation system should be re-assessed and adapted in order to address the gaps currently in the NTP recording and reporting system. A study should be conducted to evaluate the current system at all levels of the NTP and to determine where the current system works well and where it breaks down. The results from this study could then be used to strengthen the current system. Key personnel with appropriate training and skills should then be placed at all levels of the NTP to be responsible and accountable for the system. Retraining programmes should keep staff turnover in mind.

Recommendation 6: Incorporate record linking as part of the monitoring and evaluation process.

Background:

The current study, and many other studies, has shown that record linking is a useful tool to be used as part of the monitoring and evaluation system.

Recommendation:

It is recommended that the methodology used in this study should be used by the NTP to evaluate the recording and reporting system on an ongoing basis. The record linking could be conducted at district level on the ETR data and NHLS laboratory data. All positive TB results could be sent electronically to each district office from the centralised NHLS data warehouse in order to be linked to the ETR data. A computerised system could be developed in order to assist with the data management, record linking and reporting of the ETR and TB results.

This would be useful for the NTP to identify problem PHC facilities where accuracy and completeness should be addressed. Conducting a capture-recapture study is however not always necessary, but could be done at less regular intervals and for only selected areas.

Recommendation 7: Developing and validating effective record linking strategies.

Background:

Record linking is not a quick and easy process and the record linking may produce unusable results if not properly implemented. Most record linking strategies and algorithms have been developed in high income countries and have not been tested thoroughly in a South African setting, especially taking the spelling of African names into consideration.

Recommendation:

It is recommended that a study be conducted in order to develop a valid and easily adaptable record linking strategy for South African health information data. This could be accomplished by testing and adapting the different record linking strategies already available for the South African context. It is especially important to thoroughly test phonetic coding (Soundex etc.) for use with African names.

Recommendation 8: Making record linking user-friendly.

Background:

In order to do accurate record linking, keeping in mind the amount of effort and time needed, a unique identifier should be available to be linked in all the data sources before record linking is seriously considered.

Recommendation:

A national unique patient health care number available on all health care related documents would be the ultimate solution to many identification problems in health care, not only in record linking issues. This unique number could be maintained centrally with all of an individual's heath related information. This however would not be an easy and quick process as all systems and procedures throughout the country will need to be adapted. This would also be a very costly exercise. This is not an infallible solution though, as these unique identifiers could still be recorded incorrectly, some people may lose their identifier record and some people may even use other people's identifiers.

Recommendation 9: Integration of health information systems.

Background:

No health information systems and the data from these systems can produce accurate and complete data if the data sources are not appropriately integrated.

Recommendation:

The South African NTP should follow the WHO recommendation of integrating laboratory services within the NTP. The data sources used within the NTP should especially be better integrated. This could be done by feeding the laboratory results directly into the NTP recording and reporting system. This would only be possible if all patients have a national unique health care number in order to link the individual's laboratory results to his/her other laboratory results and TB treatment information. The ETR should also be moved to the PHC clinic. The functionality of the ETR could be extended to be more than just a data capture and data aggregation system by using the ETR as a patient treatment management and laboratory sample management system. This would create a better integrated health information system

Recommendation 10: Verify that non-reported cases are truly not reported.

Background:

A limitation of this study is that non-recorded TB cases were only verified within the study area. It is therefore possible that some of these cases accessed healthcare elsewhere or died before starting treatment.

Recommendation:

Non-reported cases should be searched for in the ETR data of the sub-district or of the whole of Cape Town in order to determine if any of these cases were placed on treatment at a PHC clinic outside the study. Mortality data (death certificates) should be used to determine if any of the additional cases died. Patient confidentiality issues should however always be kept in mind. Death certificates may not state the prime cause of death correctly should the person not have been seen by health services in the recent past and in some instances death certificates are issued by the police service with cause of death stated merely as "natural causes" when no foul play was suspected and no medically trained person can be found to certify the death. This happens from time to time in impoverished and rural areas.

Recommendation 11: Find-out what happened to those not recorded.

Background:

Currently little is known regarding the reasons why TB cases are not recorded and what eventually happens to these cases.

Recommendation:

A study is recommended in order to track and follow up those TB cases not recorded and reported on in order to determine the reasons for non-recording, as well as to determine what happened to these cases. Such studies should concentrate on both the clinic setting and the community setting to capture the reasons from both the system functioning and the human behaviour side. A study of this kind to look at the systems failures at clinic level in relation to the type of patient who most likely will go unreported is an urgent priority for the TB services in South Africa.

6.5 REFERENCES TO CHAPTER 6

1. Greenberg RS, Daniels SR, Flanders WD, Eley JW, Boring JR. Medical epidemiology. 2nd ed. New York: Lange Medical Books/McGraw-Hill; 2001.
2. World Health Organization. Global tuberculosis control : epidemiology, planning, financing : WHO report 2009. Geneva: World Health Organization; 2009.WHO/CDS/TB/2003.313, 303 pp.
3. World Health Organization. Stop TB Dept. Implementing the WHO Stop TB Strategy : a handbook for national tuberculosis control programmes. Geneva: World Health Organization; 2008.WHO/HTM/TB/2008.401, 184 pp.
4. World Health Organization. Stop TB Dept. Stop TB policy paper : 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.WHO/HTM/TB/2009.416, 58 pp.
5. Botha E, den Boon S, Lawrence KA, et al. From suspect to patient: tuberculosis diagnosis and treatment initiation in health facilities in South Africa. Int J Tuberc Lung Dis 2008;12:936-41.
6. Botha E, Den Boon S, Verver S, et al. Initial default from tuberculosis treatment: how often does it happen and what are the reasons? Int J Tuberc Lung Dis 2008;12:820-3.
7. Castro KG. Tuberculosis surveillance: data for decision-making. Clin Infect Dis 2007;44:1268-70.
8. Attaran A. An immeasurable crisis? A criticism of the millennium development goals and why they cannot be measured. PLoS Med 2005;2:e318.
9. Laxminarayan R, Klein EY, Darley S, Adeyi O. Global investments in TB control: economic benefits. Health Aff (Millwood) 2009;28:w730-42.
10. Laxminarayan R, Chow J, Shahid-Salles SA. Intervention Cost-Effectiveness: Overview of Main Messages. In: Disease Control Priorities in Developing Countries. 2nd edition ed. Washington, DC: Oxford University Press and the World Bank; 2006.
11. Edginton ME, Wong ML, Phofa R, Mahlaba D, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: numbers of patients diagnosed and outcomes of referrals to district clinics. Int J Tuberc Lung Dis 2005;9:398-402.
12. Dye C, Watt CJ, Bleed DM, Williams BG. What is the limit to case detection under the DOTS strategy for tuberculosis control? Tuberculosis (Edinb) 2003;83:35-43.

13. Rieder HL, Watson JM, Raviglione MC, et al. Surveillance of tuberculosis in Europe. Working Group of the World Health Organization (WHO) and the European Region of the International Union Against Tuberculosis and Lung Disease (IUATLD) for uniform reporting on tuberculosis cases. Eur Respir J 1996;9:1097-104.
14. van der Werf MJ, Borgdorff MW. Targets for tuberculosis control: how confident can we be about the data? Bull World Health Organ 2007;85:370-6.
15. Dunbar R, van Hest R, Lawrence K, et al. Capture-recapture to estimate completeness of tuberculosis surveillance in two communities in South Africa. Int J Tuberc Lung Dis 2011;15:1038-43.
16. Dunbar R, Lawrence K, Verver S, et al. Accuracy and completeness of recording of confirmed tuberculosis in two South African communities. Int J Tuberc Lung Dis 2011;15:337-43.
17. Capture-recapture and multiple-record systems estimation I: History and theoretical development. International Working Group for Disease Monitoring and Forecasting. Am J Epidemiol 1995;142:1047-58.
18. Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev 1995;17:243-64.
19. Narvaiz de Kantor I, Kim SJ, Frieden TR, et al. Laboratory services in tuberculosis control. Geneva: World Health Organization; 1998.WHO/TB/98.258, 47 pp.
20. Edginton ME, Wong ML, Hodkinson HJ. Tuberculosis at Chris Hani Baragwanath Hospital: an intervention to improve patient referrals to district clinics. Int J Tuberc Lung Dis 2006;10:1018-22.
21. National Department of Health. The South African National Tuberculosis Control Programme: Practical Guidelines. Pretoria: National Department of Health; 2009.

ANNEXURE 1 : REGISTRY PLUS™ LINK PLUS

Link Plus is a probabilistic record linking program developed at CDC's Division of Cancer Prevention and Control in support of CDC's National Programme of Cancer Registries (NPCR). Link Plus is a record linking tool for cancer registries. The software has two modes:[1]

- To detect duplicates in a cancer registry database.
- To link a cancer registry file with external files.

Although originally designed to be used by cancer registries, the program can be used with any type of data.

Link Plus performs probabilistic record linking based on the theoretical frame work developed by Fellegi and Sunter (1969). It uses the EM algorithm Expectation-maximisation (EM) algorithm, a method for maximum likelihood estimation in problems involving incomplete data (Dempster, A. P., Laird, N. M., and Rubin, D.B. 1977), to estimate parameters in the model proposed by Fellegi and Sunter.

7.1 USING LINK PLUS FOR RECORD LINKING

Link Plus can be configured in a number of ways in order to accommodate the data available and to produce the required results. The first step in configuring Link Plus is to import the required data files containing the variable to be used in the linking process. Once the data is available the following steps can be followed to complete the configuration.

- Selecting Blocking Variables
- Selecting Matching Methods
- Selecting Cut-Off Value
- Run Matching Process
- Manual Review

7.2 BLOCKING VARIABLES AND PHONETIC SYSTEMS

Blocking variables are variables common to the two files that are used to ‘block’ (or partition) the two files. Only within these blocks are matching variables compared between the records.

Blocking is a way to reduce the computing cost by portioning files into mutually exclusive and exhaustive blocks and performing comparisons only on records within each block.

Link Plus provides a simple blocking ("OR blocking") mechanism by indexing the variables for blocking and comparing the pairs with the identical values on at least one of those variables. For instance, if users select Social Security Number, Soundex of last name, and Date of Birth for blocking, Link Plus compares the pairs with the same Social Security Number OR the same Soundex OR the same Date of Birth instead of the same Social Security Number AND the same Soundex AND the same Date of Birth.

Phonetic coding involves coding a string based on how it is pronounced. Link Plus offers a choice of 2 Phonetic Coding Systems:

7.2.1 Soundex

The Soundex code for a name consists of a letter followed by three numbers: the letter is the first letter of the name, and the numbers encode the remaining consonants. Zeroes are added at the end if necessary to produce a four-character code. Additional letters are disregarded.

Example: Washington is coded W-252 (W, 2 for the S, 5 for the N, 2 for the G (remaining letters disregarded)

Using the Soundex code phonetic system reduces matching problems due to different spellings, and is simple and fast.

7.2.2 New York State Identification and Intelligence System (NYSIIS)

NYSIIS was developed in New York State in 1970. NYSIIS maps similar phonemes to the same letter and maintains relative vowel positioning. In addition, the codes can be pronounced by the reader without decoding.

Example: Deborah Walker = DABARAWALCAR

NYSIIS offers an improvement to the Soundex algorithm, with a reported accuracy increase of 2.7% over Soundex. NYSIIS is more distinctive than Soundex; people are more likely to have the same Soundex than the same NYSIIS. Some studies suggest NYSIIS performs better than Soundex when Spanish names are used.

7.3 AVAILABLE MATCHING METHODS

The following nine Matching Methods are available in Link Plus. In addition to the exact matching method, there are several approximate Matching Methods, or Comparators, which find partial, approximate, or fuzzy matches, and generate values of match on a particular field that can be other than “yes” or “no”, 1 or 0.

These matching methods incorporate partial matching, value-specific matching, or both, and are customised for the content of specific data items or types.

7.3.1 Exact

This is a character-for-character string comparison method, but is case insensitive. Results are either yes or no.

7.3.2 Last Name and First Name

These Matching Methods incorporate both partial and value-specific matching and NYSIIS phonetic code to account for minor typographical errors, misspellings, and hyphenated names. For a hyphenated name, the name matching methods compare each sub-string separated by the hyphen with the other name of the comparison pair.

For a comparison pair having the same name, the frequency of this name will be incorporated into computing the weight of this pair. A common name results in low weight and a rare name results in high weight. By default, the frequencies of names are derived from File 1.

For a comparison pair having different first names, the First Name Matching Method will use a file of nick names (provided by the program) to match the names if one of the names is regarded as a nick name. If the names are regarded as no match after the nick name checking, this method will try the partial matching.

The partial matching is based on the *Jaro-Winkler metric* that is widely used to measure the similarity between two names. A name match status (yes or no) is determined according to whether the similarity score (between 0 and 1) is greater than a threshold value, or less than and equal to the threshold value. For a hyphenated name, the name matching methods compare each sub-string separated by the hyphen with the other name within the comparison pair.

Jaro-Winkler Metric

The Jaro-Winkler Metric is a string comparator which measures the partial agreement between two strings. In many matching situations, it is not possible to compare two strings exactly (character-by-character) because of typographical errors.

The basic Jaro algorithm consists of three procedural components:

- compute the string length,
- find the number of common characters in the two strings, and
- find the number of transpositions between the two strings.

The definition of common characters used is that any agreeing characters must be within half the length of the shorter string. The definition of transposition is that the character from one string is out of order with the corresponding common character from the other string. Winkler enhanced the Jaro string comparator by assigning increased value to agreement on beginning characters of a string. This enhancement was based on ideas from a very large empirical study by Pollock and Zamora for the Chemical Abstract Service. The study showed that the fewest errors typically occur at the beginning of a string and that error rates by character position increase monotonically as the position moves to the right.

The formula for the basic Jaro string comparator is as follows:[1]

$$\Phi = (c / l_1)W_1 + (c / l_2)W_2 + ((c - n_t) / c)W_t$$

Where

W_1 = weight associated with string in the first file

W_2 = weight associated with string in the second file

W_t = weight associated with transpositions

l_1 = length of string in first file

l_2 = length of string in second file

n_t = number of transpositions

c = number of common characters

The number of transpositions is calculated as follows: The first common character on one string is compared to the first common character on the other string. If the characters are not the same, half of a transposition has occurred. Then the second common character on one string is compared to

the second common character on the other string, etc. The number of mismatched characters is divided by two to yield the number of transpositions.

7.3.3 Middle Name

This method accounts for occurrence of the middle initial only versus the full middle name.

7.3.4 SSN (National identification number)

This method is specifically for Social Security Number. It incorporates partial matching to account for typographical errors and transposition of digits. The SSN Matching Method also enables the match between a 9 digit social security number in one file and a 4 digit social security number in the other file. If the last 4 digits of the 9 digit number are the same as the 4 digit number, the comparison pair will receive a higher score.

7.3.5 Date

This method incorporates partial matching to account for missing month values and/or day values. The Date matching method checks to see if two dates are the same on day, month, and year components. If they are the same on all three components, the comparison pair will get a high weight (w). If they agree on year and month but are missing on day, the weight (w1) will be positive but less than w. If they agree on year but are missing on month and day the weight (w2) will still be positive but less than w1. If they are not missing values, the Date matching method will check if the day and month are swapped. The method also checks for transposition.

7.3.6 Value-Specific (Frequency-Based)

Intended for advanced users, this method sets weights for matching values, based on the frequencies of values in the files being compared. A match on a frequent value is associated with a low weight, while a match on a rare value is associated with a high weight. For example, if this Matching Method is applied to the Matching Variable of Race in a file with a high proportion of records with the value of 01 for White, a match on value 01 would be weighted lower than a match on the value 03, American Indian.

7.3.7 Generic String

This method incorporates partial matching to account for typographical errors. The Generic String Matching Method uses an edit distance function (Levenshtein distance) to compute the similarity

of two long strings. The edit distance is defined as the minimum number of operations (insertion, deletion, or substitution of a single character) needed to transform one string into the other.

7.3.8 Zip Code

The Zip Code Matching Method enables the match between a 9 digit zip code and a 5 digit zip code. If the first 5 digits of the 9 digit zip code is the same as the 5 digit zip code, the comparison pair will get a high weight.

7.4 M-PROBABILITY

The M-probabilities determine the maximum agreement weight and minimum disagreement weight for each field, and so define the agreement and disagreement weight ranges for each field and for the entire record. These probabilities allow you to specify which fields provide the most reliable matching information and which provide the least. For example, in person matching, the gender field is not as reliable as the SSN field for determining a match since a person's SSN is more specific. Therefore, the SSN field should have a higher m-probability than the gender field. The more reliable the field, the greater the m-probability for that field should be.

By default, the M-probabilities are derived from the frequencies of the matching variables in the first file in hand; however, Link Plus provides an option that allows to use the frequencies from 2000 US Census data or 2000 US Nation Death Index data, for instance for names. Link Plus computes the M-probabilities based on the data at hand using the EM algorithm as the second Decision Method option. To compute the default M-probabilities, Link Plus uses the data in File 1 to generate the frequencies of last names and first names and then computes the weights for last name and first name based on the frequencies of their values.

7.5 DIRECT METHOD

"Direct Method" refers to the method used to derive the M-probabilities used in linking. The Direct Method is more robust and it consumes roughly half of the CPU time needed to have Link Plus compute the M-probabilities. However, using the EM Algorithm may improve results, because computed M-probabilities are likely to be more reflective of the true probabilities, since they were computed by capturing and utilising the information dynamically from the actual data

being linked. This is especially true when the files are large and the selected Matching Variables provide sufficient information to identify potential linked pairs.

7.6 EM ALGORITHM IN LINK PLUS

The EM algorithm is a method for maximum likelihood estimation in problems involving incomplete data such as estimating parameters in latent models. Latent models are widely used on data from unknown subpopulations and latent (hidden) factors. In the situation of linking, there are two latent populations:

- linked pairs and
- unlinked pairs

because we don't know the true status of a pair. The EM algorithm empowers us to automatically obtain matching parameters without relying on prior empirical values of these matching parameters or using training data.

EM alternates between performing an expectation (E) step, which computes an expectation of the likelihood by including the latent variables as if they were observed, and maximisation (M) step, which computes the maximum likelihood estimates of the parameters by maximising the expected likelihood found on the E step. The parameters found on the M step are then used to begin another E step, and the process is repeated.

Since the data varies on quality, the estimates obtained from the EM algorithm also reflect the characteristics of the data dynamically. In spite of its slow numerical convergence, the EM algorithm has become a very popular computational method in statistics. The beauty of the EM algorithm is its generality, simplicity, flexibility, and impressive numerical stability. The basic principle of the EM algorithm is to derive a solution in a complicated case from a corresponding solution in a simple case.

The M-probability measures the reliability of each data item. A Value of 0 means the data item is totally unreliable (0%) and a value of 1 means that the data item is completely reliable (100%). Reasonable values range from 0.9 (90% reliable) to 0.9999 (99.99% reliable). If first name is less reliable than last name, then the M-probability of first name should be less than the M-probability of last name.

To compute the default M-probabilities, Link Plus uses the data in the first data source to generate the frequencies of last names and first names and then computes the weights for last name and first name based on the frequencies of their values.

7.7 CUT OFF VALUE

The Cut Off Value is the linking score value above which comparison pairs are accepted as potential links. For a comparison pair, the overall weight over all matching variables; a higher score means a higher likelihood of being a match.

7.8 PROBABILISTIC MATCHING

The total score for a linking between any two records is the sum of the scores generated from matching individual fields. The score assigned to a matching of individual fields is:

- Based on the probability that a matching variable agrees given that a comparison pair is a match.
- M Probability - similar to "sensitivity".
- Reduced by the probability that a matching variable agrees given that a comparison pair is not a match.
- U Probability - similar to "specificity".

7.9 MATHEMATICAL MODEL

In an application with two files, A and B, denote the rows (records) by α(a) in file A and β(b) in file B. Assign *K* characteristics to each record. The set of records that represent identical entities is defined by

$$M = \{(a, b); a = b; a \in A; b \in B\}$$

and the complement of set *M*, namely set *U* representing different entities is defined as

$$U = \{(a, b); a \neq b; a \in A; b \in B\}$$

A vector, γ is defined, that contains the coded agreements and disagreements on each character:

$$\gamma[\alpha(a), \beta(b)] = \{\gamma^1[\alpha(a), \beta(b)], \ldots \gamma^K[\alpha(a,), \beta(b)]\}$$

where K is a subscript for the characteristics (sex, age, last name, first name, etc.) in the file. The conditional probability of observing a specific vector γ given $(a, b) \in M, (a, b) \in U$ are defined as

$$m(\gamma) = P\{\gamma[\alpha(a), \beta(b)]|(a, b) \in M\} \sum_{(a,b)\in M} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b)|M]$$

and

$$u(\gamma) = P\{\gamma[\alpha(a), \beta(b)]|(a, b) \in U\} \sum_{(a,b)\in U} P\{\gamma[\alpha(a), \beta(b)]\} \cdot P[(a, b)|U]$$

respectively.

7.10 REFERENCES TO ANNEXURE 1

1. Division of Cancer Prevention and Control. Registry Plus™ Link Plus. In. 2.0 ed. Atlanta: Centers for Disease Control and Prevention; 2007.

ANNEXURE 2 : RECORD LINKING PROCEDURE FOLLOWED FOR THE CURRENT STUDY

The following section details the process followed in linking records within-data-sources and between-data-sources. The two methods available in Link Plus namely to detect duplicates in a register, within-data-source, and linking register files with external files, between-data-source, are in principle the same. The linking process described in this section is therefore based on the between-data-source linking method. Due to ethics and patient confidentiality the record linking conducted in annexure 2 was based on fictitious data and real patient information was not used.

The first step was to export the data from the source database into a format compatible with Link Plus, in this case a tab delimited text file. Two text files exported from the TB treatment register and the laboratory data source consisted of the data fields shown in Table 8-1:

Table 8-1 : The data fields exported from the two data sources to be used in Link Plus.

TB Treatment Register	**Laboratory data source**
Surname	Surname
First name	First name
Gender	Gender
Age	Age
Address	Address
Date registered	Date specimen registered in laboratory
Date treatment started	
Date treatment ended	

Step two was to configure Link Plus for record linking. In the data type section, the format of the files which contains the data to be link should be specified and the location of the files on the hard drive. In this case delimited was selected as the data was in tab delimited text files. The block variable section provides a means of limiting the number of unnecessary matching. Only gender was selected as a blocking variable and therefore matching was done between the two files where gender matched between the two files. In essence records of males were matched with records of males and records of females were matched with records of females. A limitation of selecting a blocking variable is that if gender was recorded incorrectly in one of the data sources, a match will

not be found. This limitation was dealt with during the manual review of non-matched records after record linking with Link Plus was performed.

The matching variables and methods section indicate the data fields from the two data sources which will be used for matching purposes. In this case surname, first name, gender and age were selected. The matching methods indicate the method that should be used and indicates to Link Plus the type of data expected in the data field. By selecting the matching method “Last Name” for Surname and “First Name” for First Name, Link Plus knows to treat these two data fields appropriately as discussed in Annexure 1. Additionally Link Plus also conducts a cross-over match between Surname and First name, in case these to field were recorded incorrectly. For this study “Direct Matching” was not selected, which means Link Plus would use Probabilistic matching, and the ‘Cutoff Value” was set to 0 in order to maximise the possibility of including very low scoring matched records in the manual review part of Link Plus.

In step three the manual review are configured. The manual review section automatically contains the data fields select during step two as matching variable. The additional data fields selected was Address and Date registered from the TB treatment register, and Date specimen registered in laboratory from the laboratory data source. These additional data fields were used to identify low scoring matched as true matches or to change matches to non-matches.

During the Link Plus manual review process there are three different matching types that can be selected, namely:

- True match
- False match
- Uncertain match

Table 8-2 contains a few records ranging from high matching scores to low matching scores and the selection of a true- or false match. The selection of a matching type is not done by Link Plus, but by the user. The process followed in the current study was to identify a matching score where records were considered not to be matches. This decision is up to the researcher as to where the cut-off should be. The matching scores calculated by Link Plus was therefore only used as a guideline as to where the cut-off should be and not as a definitive decision as to if two records were true matches or false matches.

All records above the selected cut-off matching score were then marked as true matches and all record below as false matches. All the records were then manually reviewed. If a record marked as a true match could be identified as being a false match, then match status was changed to false match. On the other hand, if a false match could be identified as a true match the matching status was corrected. If however there were not enough information available to make a selection as a true- or false match, the record was assigned a matching status of uncertain match. Uncertain matches were then given to research nurses in order to make a final decision on matching status and their decision was then applied to the record.

Base on Table 8-2 the following decision was made on matching status. In Table 8-2, File 1 corresponds to the TB treatment register and File 2 corresponds to the laboratory data source.

Line 1: True match – all the matching fields were identical.

Line 2: True match – all the matching fields were identical, except first name in file 2 was misspelled.

Line 3: False match – this was originally assigned as a true match, but changed to false match due to the age being different with more the 2 years. The address and date of the specimen also did not correspond to that of File 1.

Line 4: True match - the surname and first name were swapped around in File 2 and the age differed with only one year.

Line 5: False match – only the surname was the same between the two file.

Table 8-2: Example of the manual review in Link Plus assigning matching status.

Line #	Match	Score	Class	Link ID	File	Record #	Surname	First name	Gender	Age	Address	Date Reg
1	True match	15.2	1	1	1	39	Bentley	Ali	Male	50	7369 Cursus Str	03/01/2007
			1	1	2	1070	Bentley	Ali	Male	50	7369 Cursus Str	04/01/2007
2	True match	14.5	0	59	1	215	Berg	Fletcher	Male	65	5866 Natoque Ave	05/03/2007
			0	59	2	538	Berg	Fletser	Male	65		09/03/2007
3	False match	9.0	0	100	1	55	Vinson	Nina	Female	32	Tincidunt Street	13/09/2007
			0	100	2	77	Vinson	Nina	Female	16	56 2nd Ave	01/01/2007
4	True match	8.66	0	145	1	22	Jensen	Vielka	Female	21	77 Odio Rd	15/06/2007
			0	145	2	56	Vielka	Jensen	Female	20		17/06/2007
5	False match	3.3	0	200	1	79	Gardner	Carol	Female	19	66 Convallis Str	18/09/2007
			0	200	2	97	Gardner	Dean	Male	20	66 Convallis Str	01/01/2008

In step four all the records which were defined as true matches or false matches were exported to an analysis database. Each record in the TB treatment register (Table 8-3) was updated with the record number of the laboratory data source if a true match was found. Similarly each record in the laboratory data source was updated with the record number of the TB treatment register if a true match was found (Table 8-4).

Table 8-3: Example of TB treatment register data after record linking.

TB Reg record #	**Lab record #**	**Surname**	**First name**	**Gender**	**Age**	**Address**	**Date Reg**
39	1070	Bentley	Ali	Male	50	7369 Cursus Str	03/01/2007
215	538	Berg	Fletcher	Male	65	5866 Natoque Ave	05/03/2007
55		Vinson	Nina	Female	32	Tincidunt Street	13/09/2007
22	56	Jensen	Vielka	Female	21	77 Odio Rd	15/06/2007
79		Gardner	Carol	Female	19	66 Convallis Str	18/09/2007

Table 8-4: Example of Laboratory data source after record linking.

Lab record #	**TB Reg record #**	**Surname**	**First name**	**Gender**	**Age**	**Address**	**Date Reg**
1070	39	Bentley	Ali	Male	50	7369 Cursus Str	04/01/2007
538	215	Berg	Fletser	Male	65		09/03/2007
77		Vinson	Nina	Female	16	56 2nd Ave	01/01/2007
56	22	Vielka	Jensen	Female	20		17/06/2007
97		Gardner	Dean	Male	20	66 Convallis Str	01/01/2008

During step 5 a final manual review was conducted in order to make sure that no matches were incorrect and that no records we missed during the matching process.