Changes in Mortality Patterns and Associated Socioeconomic Differentials in a Rural South African Setting: Findings from Population Surveillance in Agincourt, 1993-2013

Chodziwadziwa Whiteson Kabudula

A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy (by publications)

20th December 2017

Supervisors

Professor Samuel J. Clark, Department of Sociology, The Ohio State University, United States of America.

Associate Professor Mark Collinson, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa.

Professor Stephen Tollman, School of Public Health, University of the Witwatersrand, Johannesburg, South Africa.

Dedication

To the memory of my late uncle, Wilson Drill Namaona (may his soul continue resting in peace), whose guardianship, support and encouragement during the early years of my education continue to be a prime source of inspiration and motivation in my career.

Declaration

I declare that *Changes in Mortality Patterns and Associated Socioeconomic Differentials in a Rural South African Setting: Findings from Population Surveillance in Agincourt, 1993-2013* is my own unaided work. It is being submitted for the degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg, South Africa. It has not been submitted before for any degree or examination at any other University, and I declare that all the sources I have used or quoted have been indicated and acknowledged as complete references.

Chodziwadziwa Whiteson Kabudula 20th December 2017

Signed:..........................

Acknowledgements

I would like to extend my heartfelt gratitude to my supervisors: Professor Samuel Clark, Associate Professor Mark Collinson and Professor Stephen Tollman for their encouragement, guidance, advice and support throughout my doctoral studies.

Dr Brian Houle deserves special credit and acknowledgement for his immense contributions towards my studies. His guidance and thoughtful suggestions and comments in the analysis and write-up of three of the four papers of this thesis are highly appreciated.

A special thank you also to Dr. Georges Reniers for contributing immensely towards my professional and career growth. I would not have developed the expertise in record linkage without his guidance and encouragement.

I am also grateful to the co-authors of the papers in this thesis for their respective constructive ideas, comments and suggestions during the write-up of the papers.

The respective contributions to the production of the data used in the studies of this thesis from the members of the communities in the Agincourt Health and Demographic Surveillance System study area, the staff of the Agincourt Health Centre and present and past field staff and management team of the MRC/Wits Rural Public Health and Health Transitions Research Unit are also appreciated and acknowledged.

To all my past and present colleagues within the MRC/Wits Rural Public Health and Health Transitions Research Unit both in Johannesburg and Bushbuckridge, among whom: Francesc Xavier Gómez-Olivé, Ryan Wagner, Paul Mee, Wendy Pearsall, Dawn Dalby, Nicolette Pingo and Rhian Twine, I say thank you for supporting me in many different ways throughout my studies.

For the financial support towards my studies I am deeply grateful to the leadership of the MRC/Wits Rural Public Health and Health Transitions Research Unit - particularly Professors Stephen Tollman and Kathleen Kahn and the University of the Witwatersrand, Faculty of Health Sciences.

To my dear wife Memory Nyasha Mhembere, son Precious, daughter Theodora, mum, sister, aunt and cousins I am forever grateful for the endless love, support and encouragement that kept me going during the tough moments.

Above all, I am mostly grateful to the Almighty God for making it possible for me to complete this PhD and for the infinite blessings.

Abstract

Understanding a population's mortality and disease patterns and their determinants is important for setting locally-relevant health and development priorities, identifying critical elements for strengthening of health systems, and determining the focus of health services and programmes. This thesis investigates changes in socioeconomic status (SES), cause composition of overall mortality and the socioeconomic patterning of mortality that occurred in a rural population in Agincourt, northeast South Africa over the period 1993-2013 using Health and Demographic Surveillance Systems (HDSS) data. It also assesses the feasibility of applying record linkage techniques to integrate data from HDSS and health facilities in order to enhance the utility of HDSS data for studying mortality and disease patterns and their determinants and implications in populations in resource-poor settings where vital registration systems are often weak. Results show a steady increase in the proportion of households that own assets associated with greater modern wealth and convergence towards the middle of the SES distribution over the period 2001-2013. However, improvements in SES were slower for poorer households and persistently varied by ethnicity with former Mozambican refugees being at a disadvantage. The population experienced steady and substantial increase in overall and communicable diseases related mortality from the mid-1990s to the mid-2000s, peaking around 2005-07 due to the HIV/AIDS epidemic. Overall mortality steadily declined afterwards following reduction in HIV/AIDS-related mortality due to the widespread introduction of free antiretroviral therapy (ART) available from public health facilities. By 2013, however, the cause of death distribution was yet to reach the levels it occupied in the early 1990s. Overall, the poorest individuals in the population experienced the highest mortality burden and HIV/AIDS and tuberculosis mortality persistently showed an inverse relation with SES throughout the period 2001-13. Although mortality from non-communicable diseases (NCDs) increased over time in both sexes and injuries were a prominent cause of death in males, neither of these causes of death showed consistent significant associations with household SES. A hybrid approach of deterministic followed by probabilistic record linkage, and the use of an extended set of conventional identifiers that included another household member's first name yielded the best results for linking data from the Agincourt HDSS and health facilities with a sensitivity of 83.6% and a positive predictive value (PPV) of 95.1% for the best fully automated approach. In general, the findings highlight the need to identify the chronically poorest individuals and target them with interventions that can improve their SES and take them out of the vicious circle of poverty. The results also highlight the need for integrated health-care planning and programme delivery strategies to increase access to and uptake of HIV testing, linkage to care and ART, and prevention and treatment of NCDs especially among the poorest individuals to reduce the inequalities in cause-specific and overall mortality. The findings also contribute to the evidence base to inform further refinement and advancement of the health and epidemiological transition theory. Furthermore, the findings demonstrate the feasibility of linking HDSS data with data from health facilities which would facilitate population-based investigations on the effect of socioeconomic disparities in the utilisation of healthcare services on mortality risk.

Keywords

Agincourt

Cause of death composition

Epidemiological Transition

Health and Demographic Surveillance System (HDSS)

Household assets

HIV/AIDS

Index of Inequality

InterVA

Mortality

Non-communicable Diseases

Population Surveillance

Record linkage

Rural

Socioeconomic Status

South Africa

Verbal Autopsy

Wealth Index

Original papers

This thesis is based on the following original papers:

I **Kabudula CW**, Houle B, Collinson MA, Kahn K, Tollman S and Clark S (2016). Assessing changes in household socioeconomic status in rural South Africa, 2001-2013: A distributional analysis using household asset indicators. *Social Indicators Research.* doi:10.1007/s11205-016-1397-z

II **Kabudula CW**, Houle B, Collinson MA, Kahn K, Gómez-Olivé FX, Clark S and Tollman S (2017). Progression of the epidemiological transition in a rural South African setting: Findings from population surveillance in Agincourt, 1993-2013. *BMC Public Health*,17:424. doi: 10.1186/s12889-017-4312-x

III **Kabudula CW**, Houle B, Collinson MA, Kahn K, Gómez-Olivé FX, Tollman S and Clark S (2017). Socioeconomic differences in mortality in the antiretroviral therapy era in Agincourt, rural South Africa, 2001-2013: a population surveillance analysis. *Lancet Global Health*, 5(9): e924-e935. doi: 10.1016/S2214-109X(17)30297-8

IV **Kabudula CW**, Clark BD, Gómez-Olivé FX, Tollman S, Menken J and Reniers G (2014). The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa. *BMC Medical Research Methodology*, 14:71. doi: 10.1186/1471-2288-14-71

The papers are referred to in the text by their roman numerals (I through IV) and they are all published under a creative commons attribution license, which permits their reproduction in this thesis.

Contents

Declaration i

Acknowledgements iii

Abstract iv

Keywords vii

Original papers viii

List of Tables xiii

List of Figures xv

Acronyms and abbreviations xvii

1 Introduction 1

1.1 Theoretical perspectives on determinants of mortality and disease patterns 2

1.2 Epidemiological perspectives on changes in mortality and disease patterns 4

1.3 The use of record linkage in studying mortality and disease patterns 7

1.4 Mortality trends and their determinants in South Africa 10

1.5 Statement of the problem 14

1.6 Research aims, questions and hypotheses 15

1.6.1 Overall aim 15

1.6.2 Specific objectives 15

1.6.3 Research questions 16

1.6.4 Hypotheses 16

1.6.5 Thesis themes 17

1.7 Conceptual framework 17

2 Data and Methods 19

2.1 Study Setting 19

2.2 Study design 20

2.3 Data 20

2.3.1 Health and socio-demographic surveillance system data 20

2.3.2 Health facility data 23

2.4 Data Analysis 24

2.4.1 Measuring household SES 24

2.4.2 Assessing changes in household SES 24

2.4.3 Data file for mortality analysis 25

2.4.4 Estimating mortality risk 25

2.4.5 Assessing changes in the cause composition of overall mortality 26

2.4.6 Assessing socioeconomic differentials in mortality patterns 26

2.4.7 Assessing the feasibility of record linkage of health facility data and demographic surveillance data . . . 28

2.5 Ethics . . . 29

3 Findings on changes in household SES in Agincourt, rural South Africa, 2001-2013 30

3.1 Measuring SES . . . 30

3.2 Poverty transition: Changes in household SES . . . 31

4 Findings on mortality transition in Agincourt, rural South Africa, 1993-2013. 38

4.1 Mortality transition . . . 38

4.1.1 Overall mortality . . . 38

4.1.2 Mortality risk by age . . . 39

4.1.3 Life expectancy trends . . . 42

4.1.4 Distribution of deaths by cause categories . . . 44

4.1.5 HIV-cause-deleted life expectancy . . . 53

4.1.6 Progression of the epidemiological transition . . . 55

5 Socioeconomic disparities in mortality indicators 59

6 Feasibility of using record linkage to enhance the utility of HDSS for studying utilisation of health services and its effects on mortality 63

7 Discussion and conclusions 65

7.1 General discussion . . . 65

7.1.1 Implications of findings on changes in SES . . . 65

7.1.2 Changes in mortality patterns in Agincourt in the context of the epidemiological transition framework 66

7.1.3 Using record linkage to enhance the utility of HDSS for studying utilisation of health services and its effects on mortality 68

7.1.4 Implications of the changing mortality patterns for health policy and practice . . . 69

7.1.5 Limitations of the data and methods 69

7.1.6 Future research areas 70

7.2 Conclusions 71

Bibliography **73**

Appendix A: Published Papers **86**

List of Tables

1.1 Key themes addressed in the thesis papers 17

2.1 Data for thesis papers 21

2.2 Calculation of modified ridit scores 28

3.1 Movement of households between quintiles of absolute, PCA and MCA indices 31

3.2 Household SES quintile transition matrix between 2001 and 2013: Agincourt 37

4.1 Person years and number of deaths by time period and cause of death categories, Agincourt, South Africa, 1993-2013 39

4.2 Survivors of the synthetic birth cohort of 1000 to ages 5, 50 and 65 40

4.3 Years of life expected at selected ages 43

4.4 Female-male difference in life expectancy at selected ages 43

4.5 Top five causes of death among all individuals by time period, Agincourt 1993-2013 46

4.6 Top five causes of death among children 0-4 years old by time period, Agincourt 1993-2013 47

4.7 Top five causes of death among individuals aged 5-14 years by time period, Agincourt 1993-2013 48

4.8 Top five causes of death among individuals aged 15-49 years by time period, Agincourt 1993-2013 49

4.9 Top five causes of death among individuals aged 50-64 years by time period, Agincourt 1993-2013 50

4.10 Top five causes of death among individuals aged 65+ years by time period, Agincourt 1993-2013 51

4.11 Potential years gained in life expectancy at selected ages if HIV/AIDS and TB causes were deleted 55

List of Figures

1.1 Trends in HIV Prevalence among 15-49 year old antenatal women in South Africa 11

1.2 Thesis papers in relation to determinants of health conceptual framework 18

2.1 Location of Agincourt HDSS study area, Mpumalanga province, South Africa 19

2.2 Agincourt HDSS special census modules 22

3.1 Dwelling structure 32

3.2 Water and Sanitation 33

3.3 Power supply 33

3.4 Appliances 34

3.5 Transport 34

3.6 Livestock 35

3.7 Prevalence of selected household asset indicators in 2013 in the Agincourt population and tribal areas in Mpumalanga province and South Africa 36

4.1 Survivors of the synthetic birth cohort of 1000 by age, sex, and time period: Agincourt 1993-2013 40

4.2 Percentage distribution of deaths by age and time period in Agincourt compared with South Africa 41

4.3 Years of life expected at selected ages, Agincourt 1993-2013 42

4.4 Trends in life expectancy at birth in Agincourt and selected demographic surveillance sites in Southern Africa 43

4.5 Mortality rates with and without HIV/TB by age, sex and time period: Agincourt 1993-2013 52

4.6 Life expectancy at selected ages with and without HIV/TB by sex and time period: Agincourt 1993-2013 54

4.7 HIV-cause-deleted life expectancy at birth 55

4.8 Years of life expectancy gained at birth if HIV/AIDS and TB causes were deleted by sex, and time period: Agincourt 1993-2013 56

4.9 Cause of death distribution in South Africa, Agincourt and Africa Centre 58

5.1 Relative inequalities in summary mortality indicators: Agincourt 2001-2013 61

5.2 Relative inequalities in mortality by cause of death and sex: Agincourt 2001-2013 62

Acronyms and abbreviations

AIDS Acquired Immunodeficiency Syndrome

ART Antiretroviral therapy/treatment

CDF Cumulative distribution function

EM Expectation maximization

HDSS Health and Demographic Surveillance System

HIV Human Immunodeficiency Virus

INDEPTH International Network for the Demographic Evaluation of Populations and Their Health in Developing Countries

LMICs Low- and middle-income countries

MCA Multiple Correspondance analysis

NCD Non-communicable diseases

PCA Principal Components Analysis

PDF Probability density function

PMTCT Prevention of mother-to-child transmission of HIV

PPV positive predictive value

RII Relative index of inequality

SES Socioeconomic status

SII Slope index of inquality

TB Tuberculosis

VA Verbal Autopsy

Chapter 1

Introduction

Human populations continuously undergo changes in mortality and disease patterns (Omran, 1971; Frenk et al., 1991). The changes vary by geographical location in the way they transpire (Patton et al., 2009; Moser et al., 2005; McMichael et al., 2004; Murray and Lopez, 1997). Additionally, within a given geographical location, at any point in time mortality and disease patterns may also vary by demographic (Drevenstedt et al., 2008), racial (Harper et al., 2007) and socioeconomic characteristics (Meara et al., 2008). The heterogeneity and dynamic nature of the mortality and disease patterns across space and time creates a need to quantify and characterise these patterns and their determinants and implications in different sub-populations. This is important for setting locally-relevant health and development priorities, identifying critical elements for strengthening of health systems, and determining the focus of health services and programmes.

The data required to address the need for understanding mortality and disease patterns and their determinants and implications can be generated from many potential sources. These include general sources for demographic data, such as censuses, vital statistics registration systems, sample registration systems, and general sample surveys, and specialised sources, such as sample surveys on health, administrative records on health and mortality (for example, health facility records), health and demographic surveillance systems (HDSS) as utilised in this thesis, epidemiological studies, and clinical trials (Nolen et al., 2005; Siegel, 2011; Ye et al., 2012). However, the data generated from each source independently only addresses specific questions. Addressing a broader range of questions and providing a more comprehensive view of a society's mortality and disease patterns and their determinants and implications requires integration and combined analysis of data from disparate sources. The most cost-effective means for integrating information from different sources is record linkage. By definition, this is the process of bringing together information relating to the same individual from different sources achieved by using a limited set of basic sociodemographic factors known as "linkage variables" to uniquely and reliably identify an individual across multiple datasets (Holman et al., 1999; Jutte et al., 2011). As expressed by Clark (2004), the quality of this process can be assessed by two key measures. One is sensitivity, which gives the proportion of true matches produced by the record linkage process, and the other is positive predictive value (PPV), which gives the proportion of matches produced by the record linkage process that are true matches.

This thesis investigates changes in the cause composition of overall mortality and the socioeconomic patterning of the mortality changes that have occurred in a rural population in northeast South Africa over the period 1993-2013 using data from one of the longest running HDSSs in Southern Africa. In addition, it also assesses the feasibility of applying record linkage techniques to integrate data from HDSS and health facilities in order to enhance the utility of HDSS data for studying mortality and disease patterns and their determinants and implications.

This introductory chapter presents first an overview of the theoretical perspectives on determinants of mortality and disease patterns. This is followed by an overview of epidemiological perspectives for understanding changes in mortality and disease patterns. Thereafter, it provides examples of some of the recently published studies that have investigated mortality and disease patterns using data generated by record linkage. This is followed by an overview of trends in mortality and disease patterns and their determinants in South Africa from 1993-2013. The chapter finishes with the statement of the problem, research aims, questions and hypotheses of the thesis. Chapter two describes selected features of the study setting and provides a detailed description of the methods and approaches that have been applied in the studies that this thesis is based on. Chapters three to six present the main findings of the studies as well as findings from complementary analyses. Finally, chapter seven presents a discussion of the results of the studies and conclusions drawn from the findings.

1.1 Theoretical perspectives on determinants of mortality and disease patterns

A society's mortality and disease patterns which are key indicators of health are determined by the complex interaction of an articulated set of proximate and distal factors (Rogers et al., 2005; Frenk et al., 1991; Embrett and Randall, 2014; Moure-Eraso et al., 2007; Graham, 2004; Braveman and Gruskin, 2003; Marmot et al., 2008). The proximate factors are sometimes labeled as intermediary or downstream factors and they are those that directly impact the risk of mortality. These include genetic factors, behaviour and lifestyle, living conditions, working conditions, environmental exposure and the health care system (Lee and Paxman, 1997; Adler and Newman, 2002; Braveman et al., 2011; Frenk et al., 1991; Raphael, 2006). The distal factors are sometimes labeled as structural or upstream factors and these are those that indirectly influence mortality (Braveman et al., 2011; Frenk et al., 1991; Raphael, 2006). These include the political, economic, and social factors that determine in large part the quality of the proximate factors and the total level of wealth of a society (Braveman et al., 2011; Frenk et al., 1991; Raphael, 2006). These factors manifest themselves mainly in the form of culture, ideology, political decisions, laws, regulations, taxes, subsidies and social welfare policies (Frenk et al., 1991; Raphael, 2006).

Among the proximate factors, living conditions are accorded a central position because they exert greater influence on the susceptibility of individuals to various types of disease agents and subsequent mortality (Frenk et al., 1991; Marmot, 2005; Raphael, 2006; Commission on Social Determinants of Health, 2007). The influence is both direct through toxic exposures and indirect through the effect of living conditions together with structural factors on behaviours and lifestyles (Frenk et al., 1991). Together, the living conditions, behaviours and lifestyles either reduce or increase the susceptibility of individuals to various diseases (Frenk et al., 1991; Commission on Social Determinants of Health, 2007). For example, (i) poor sanitation and water supply increases susceptibility of individuals to diarrhoea, cholera and other water and sanitation-related diseases; (ii) susceptibility of individuals to air-borne diseases such as tuberculosis increases as a result of living in crowded housing conditions; (iii) cigarette smoking increases susceptibility of individuals to lung cancer; and (iv) physical activity protects individuals from developing chronic diseases such as heart disease, hypertension, type 2 diabetes, abnormal blood lipid (cholesterol and triglyceride) profile, stroke, and colon and breast cancers. In addition to the living conditions, behaviours and lifestyles, the susceptibility of individuals to various diseases and subsequent mortality is also affected by the preventive, diagnostic and therapeutic interventions offered by the health care system which is also a proximate factor in itself. Preventive interventions, such as immunization, reduces an individual's probability of becoming ill while therapeutic interventions such as the treatment of diseases with specific drugs (e.g. acute respiratory infections with antibiotics or acute diarrhoea with oral rehydration salts or HIV with antiretroviral drugs) diminish the probability of death among those who are already ill and reduce the risk of spreading the disease to others in the population (Frenk et al., 1991).

In every society, the factors that influence exposure to disease agents and individual susceptibility are not equally distributed. Some groups of people are more likely than others to be exposed to health damaging factors. This is because individuals occupy different socioeconomic positions in a society's hierarchical social structure (Graham, 2004), which in turn shape their access and exposure to a set of factors that influence exposure to disease agents and individual susceptibility. The differences in exposure to disease agents and individual susceptibility consequently produce differences in health outcomes including mortality.

1.2 Epidemiological perspectives on changes in mortality and disease patterns

A general framework for understanding changes in mortality and disease patterns and their determinants and implications is provided by the theory of the epidemiologic transition. As originally published by Abdel Omran in 1971, the theory consists of five propositions (Omran, 1971). The first is that mortality plays a key role in population dynamics. The second indicates that over time a long-term shifts occur in mortality and disease patterns whereby degenerative and man-made diseases gradually displace pandemics of infection as the leading cause of death. The third is that the largest changes in these health and disease patterns are experienced by children and young women. The fourth indicates that the shifts in health and disease patterns closely relate to demographic and socioeconomic transformations that come along with modernisation. The last one is that the transition in each setting can follow one of three models of epidemiological transition depending on the time of onset, pace, determinants and its consequences.

In this original formulation, the theory postulates that the shift from high and fluctuating to low and relatively stable levels of mortality occurs over three successive stages. It begins with "the age of pestilence and famine" in which mortality is high and fluctuating owing to epidemics of infectious diseases, famines and wars. Then it proceeds with "the age of receding pandemics" in which mortality declines progressively and degenerative diseases start to replace infectious diseases as the major causes of morbidity and death. Finally it ends with "the age of degenerative and man-made diseases" in which chronic diseases such as cardiovascular diseases, diabetes and cancer and accidents predominate as causes of death and mortality further declines and eventually becomes stable at a relatively low level.

The three models defined in the original formulation of the theory are the classical or western model, the accelerated model and the contemporary or delayed model. The classical or western model is for populations in which the transition had started early and progressed slowly but was already completed at the time the theory was formulated. It was postulated that these populations, commenced the transition in the eighteenth century. The transition was initially triggered by socioeconomic factors. Thereafter it was augmented by improvements in sanitary conditions in the late nineteenth century and progress in medical technology and public health in the twentieth century. Populations in developed countries in Western Europe were deemed to fit this model. The accelerated or semi-western model is for populations in which the transition started later but progressed faster than in the classical model. To a large extent, the transition in these populations was thought to have been driven by a combination of sanitary and medical advances and general social improvements. The Japanese population exemplifies populations that fit this model (Omran, 1971). Lastly, the contemporary or delayed model is for populations in which the transition started at a much later stage and was yet to be completed at the time of the formulation of the theory. All populations in developing countries were deemed to fit this model since substantial declines in mortality started to be registered in these populations only in the period following the end of World War II. In these populations the transition was thought to be driven by improvements in public sanitation, use of immunisation and decisive therapies (Omran, 1971).

After Omran's initial formulation of the theory, a number of analyses were carried out to assess the validity and applicability of its propositions in different settings around world. From such analyses it became apparent that changes in mortality and diseases patterns and their determinants in many countries diverged from some of the original propositions (Carolina and Gustavo, 2003; Caselli et al., 2002). Consequently, important modifications were proposed to be made to the epidemiological transition theory. Distinct fourth stages were proposed to be added to the originally proposed three stages of the shift in mortality and disease patterns over time. These proposals were made by Olshansky and Ault in 1986 and Rogers and Hackenberg in 1987 (Spijker and Llorens, 2009). Another modification was the adding of a new model to the original theory. This proposal was made by Frenk and his colleagues in 1989 (Frenk et al., 1989).

The fourth stage proposed by Olshansky and Ault is designated "the age of delayed degenerative diseases". In this stage: (1) rapid declines in mortality are concentrated mostly in advanced ages and occur at nearly the same pace for males and females; (2) the age pattern of mortality by cause resembles that in the third stage, but the risk of dying from degenerative causes of death is progressively shifted towards older ages; and (3) relatively rapid improvements in survival are concentrated among the population at advanced ages (Olshansky and Ault, 1986). Arrival at this stage is attributed to a combination of factors including shift in the age structure towards advanced ages, advances in medical technology, health care programs for the elderly and population-wide reductions in risk factors (Olshansky and Ault, 1986).

As summarised by Spijker and Llorens (2009), the other fourth stage proposed by Rogers and Hackenberg is named the "hybristic" stage. In this stage, mortality patterns are mainly influenced by individual health-related behaviour. The influence is either positive such as in settings where more healthy lifestyles are widely adopted or negative such as in settings where potentially health-destructive lifestyles like excessive alcohol drinking and smoking are widespread. Adoption of risk behaviours stems from individuals' overconfidence in their capabilities and a feeling that they are invulnerable. As a result of adopting risk behaviors, social pathologies like accidents, suicides, homicides, and alcohol and smoking related diseases influence mortality patterns.

Frenk et al. (1989) argued that for developing countries, there exists another model which does not conform to the linear and unidirectional progression of the transition portrayed in the original formulation of the epidemiological transition theory. This model is called the “protracted epidemiologic transition”. It is characterised by overlap of the stages outlined in the original formulation of the epidemiological transition theory. According to Frenk et al. (1989), changes in the patterns of morbidity and mortality in populations in settings following this model do not fully take place as specified. Infectious and chronic diseases coexist as the leading causes of morbidity and death in the same population. In addition, the shift in leading causes of death can be reversed giving rise to a “counter transition”. The coexistence of diseases belonging to different stages of the transition as prescribed in the original formulation of the epidemiological transition theory also leads to an “epidemiological polarization”. This is a situation whereby different sectors of the same population exhibit different stages of the transition.

Taking into account the proposed modifications to the original formulation of the epidemiological transition theory, Omran published a revised version of the theory in 1998 (Omran, 1998). In the revised form, the theory consists of four propositions. The first is that mortality and fertility are key forces in population dynamics. The second indicates that over time a long-term shift occurs in mortality, disease and survival patterns whereby degenerative, stress and man-made diseases and aging gradually, but not entirely, displace pandemics of infection and gross malnutrition as the leading cause of death. The third proposition is that during the transition profound inequalities in health and disease changes occur according to age, gender, race or ethnic origin, social class, indigenous status and geopolitical locales within or among countries. The last one is that there are five models of the transition that are distinguishable by the pattern, pace, determinants and consequences of health, survival and population changes.

As in the original formulation of the theory, populations in which the transition started early but the shift from high mortality to low mortality progressed slowly are designated the western or classical transition model. However, such populations are deemed to typically go through four instead of three stages of the epidemiological transition. From the original stage three, these populations move on to “the age of declining cardiovascular mortality, aging, lifestyle modification, emerging and resurgent diseases” in which mortality from cardiovascular diseases declines and life expectancy at birth increases to 80-90 years or more. Similarly, populations in which the transition started later but progressed faster than those in the classical mode are considered to follow the semi-western or accelerated transition model. Three non-western models, which are distinguishable by the pace, timing and magnitude of change in patterns in mortality and life expectancy and fertility levels after World War II were defined for populations in which the transition started at a much later stage. The first model, named “the rapid transition model”, is designated to populations in rapidly industrialising countries and societies like China. The second model, titled “the intermediate transition model”, is designated to populations in middle or lower-middle income countries. The third model, named “the slow transition model”, is designated to populations in least-developed and some less-developed countries. Timely and successful commencement of the transition in all populations that follow the non-western model was hindered by poverty, limited education, low status of women and slow pace of economic development.

In the revised version of the theory, Omran postulated that populations that follow non-western models go through three stages. They move from “the age of pestilence and famine” to “the age of receding pandemics” to “the age of triple health burden” (Omran, 1998). The characteristics of the first two stages are similar to those in the original formulation of the theory. In the age of triple health burden, populations experience simultaneously three kinds of health burdens. Firstly, health problems of the previous stages still persist. These include communicable diseases, perinatal and maternal morbidity and mortality, malnutrition, poor sanitation, persistent problems of poverty, low literacy, overpopulation and limited access to health care and safe water, particularly in rural areas. Secondly, a new set of health problems emerge. Degenerative diseases such as heart diseases, stroke, cancer and metabolic disorders, stress and man-made diseases gradually increase. These diseases continue to play a role in the mortality pattern of the population as long as lifestyles, risky health habits and lack of special technologies persist. Lastly, the health systems and physicians and other health professionals are ill-prepared to effectively deal with the mentioned health problems.

1.3 The use of record linkage in studying mortality and disease patterns

Substantive questions about a society’s mortality and disease patterns and their determinants and implications cannot be adequately answered from a single data source. A number of studies have demonstrated that a wide range of important and often unique investigations concerning mortality and disease patterns are conducted using data generated by record linkage of data from multiple disparate sources. This section provides examples of some of the recently published studies that have investigated mortality and disease patterns using data generated by record linkage.

In Australia, data generated by record linkage of data from the Disability Services Minimum Data Set, Admitted Patients Data Collection, Emergency Department Data Collection, Australian Bureau of Statistics Death Registry and Registry of Births, Deaths and Marriages were used to explore the health and mental health profiles, mortality, pattern of health service use and associated costs between 2005 and 2013 for people with intellectual disability in New South Wales (Reppermund et al., 2017).

In the United Kingdom, a study by Simmons et al. (2013) examined national trends in death rates, the proportion of deaths attributable to AIDS and risk factors associated with an AIDS-related death in the era of effective antiretroviral therapy (ART) between 1999 and 2008 using data created by record linkage of data from the national HIV and AIDS surveillance system and death reports from the death register of the Office of National Statistics (ONS). As described by Simmons et al. (2013) , the data from the national HIV and AIDS surveillance system included demographic and risk factor information on all adults aged 15 years and above newly diagnosed with HIV infection. It also included prospective clinical information such as CD4 T-lymphocyte cell counts, viral load and ART use at date last seen that was collated on an annual basis from all HIV clinics and supplementary CD4 cell counts from laboratories. In addition, first AIDS event and deaths from any cause reported by clinicians were also included in the data. The data provided by the ONS contained all deaths with a known HIV- and/or AIDS-related cause and all deaths of persons aged 65 years and under at the time of death with the International Classification of Diseases Tenth Revision (ICD10) code causes of death (Simmons et al., 2013).

Another study in the United Kingdom used data created by record linkage of national inpatient hospital admissions and mortality data across England and Wales to investigate whether socially deprived patients had an increased risk of dying following hip fracture compared with affluent patients between 2004 and 2011 (Thorne et al., 2016). The study retrieved all emergency admissions to the English and Welsh hospitals where hip fracture was recorded as the principal diagnosis on the discharge record from the Hospital Episode Statistics and the Patient Episode Database for Wales, which holds records of inpatient admissions in England and Wales respectively. Similar to the study by Simmons et al. (2013), the retrieved inpatient data were linked to death certificate data from the ONS. In addition, the data were also linked to death certificate data from the Welsh Demographic Service which also registers deaths for confirmatory purposes (Thorne et al., 2016).

In the Netherlands, record linkage was used to produce a dataset that was used to investigate first-time utilization of long-term care services among the general population (Slobbe et al., 2017). The data linked consisted of data from (i) a large primary care database with information on chronic diseases as registered by general practitioners, maintained by the Netherlands Institute for Health Services Research, (ii) the national long-term care register, with data on long-term care use for the entire population aged 20 years and above, and (iii) several administrative databases with data on predisposing and enabling factors available at Statistics Netherlands.

In Italy, Alicandro et al. (2017) used data created by record linkage to measure differences in cause-specific premature mortality by educational level. They linked data from the 2011 Italian census with 2012 and 2013 death registries. From the census, they retrieved demographic and socioeconomic information including sex, age, region of residence, marital status and the highest educational attainment. From the death registries, they retrieved the date of death and the cause of death, coded according to the ICD10.

In British Columbia (BC), Canada, using de-identified health-related data created by record linkage of data from the BC Centre for Excellence in HIV/AIDS (BCCfE) Drug Treatment Program and Population

Data BC, Eyawo et al. (2017) characterised and compared changes in mortality rates and causes of death among a population-based cohort of persons living with and without HIV from 1996 to 2013. The data from BCCfE included demographic, immunologic, virologic, ART-use and other clinical data on all known HIV-infected individuals who had ever accessed ART. The data from Population Data BC consisted of individual-level longitudinal data for all residents of BC collected by public bodies and stored in administrative databases (Eyawo et al., 2017).

In Brazil, Mandacaru et al. (2017) used linked data from three different data sources to quantify the number of deaths and serious injuries in five representative state capitals in 2012 and 2013. They also used the matched pairs from the linked records to obtain estimates of the percentage of corrections of the underlying cause of death, the circumstances that caused the injury, and the injury severity of the victims. The data sources used were: (i) the Brazilian Ministry of Health Hospital Information System, which provided information on hospitalisations; (ii) the Brazilian Mortality Information System, which provided information on causes of death; and (iii) the Police Road Traffic database reported by the police and road traffic agents.

In New Zealand, Teng et al. (2017) used census data from 1981, 1986, 1991, 1996, 2001 and 2006 linked to mortality and cancer registry data to investigate the contribution of various cancers to socioeconomic gaps in mortality and changes over time from 1981 to 2011. The cancer registry provided prospective information on the incidence of various cancers and income collected from the census was used as a measure of socioeconomic status.

In South Africa, Ingle et al. (2010) used linked data from the Comprehensive HIV and AIDS Management program in the Free State Province, the National Health Laboratory Services Database and the National Death Register to examine pre-ART mortality and associated determinants from 2004 to 2008. The Comprehensive HIV and AIDS Management program and National Health Laboratory Services Databases provided information on CD4 cell counts and treatment eligibility while the National Death Register provided information on deaths.

Another study in South Africa by Sengayi et al. (2016) used data created by record linkage to estimate cancer incidence in HIV patients attending the Sinikithemba clinic in KwaZulu-Natal from 2004 to 2011. They linked patient data of all HIV-positive adults aged 16 years at start of ART at the clinic with cancer records in the National Cancer Registry recorded in public laboratories in KwaZulu-Natal province.

A study by Johnson et al. (2015), also conducted in South Africa using record linkage of data from different sources, estimated the completeness of recording of deaths in cohorts of HIV infected patients receiving ART between 2004 and 2014. The study compared the recording of deaths in the civil registration system and in patient files from six ART programmes. The ART programmes included in the study were Khayelitsha, Gugulethu and Tygerberg programmes in the Western Cape province; McCord and Hlabisa programmes in KwaZulu-Natal province; and Themba Lethu programme in Gauteng province.

The list of examples provided in this section indicates that studies that have investigated mortality and disease patterns using data generated by record linkage have been conducted in both low and high income settings. However, studies from low income settings are disproportionately very few in number because of lack of comprehensive record linkage systems. Examples of comprehensive record linkage systems from high income settings include: the Oxford Record Linkage Study, the Secure Anonymised Information Linkage databank in Wales, the Scottish Record Linkage System, the Rochester Epidemiology Project, the Western Australia Data Linkage System, the Centre for Health Record Linkage in New South Wales, the Manitoba Center for Health Policy, the Center for Health Services and Policy Research in British Columbia, and the Institute for Clinical and Evaluative Sciences in Ontario, Canada (Holman et al., 1999; Lyons et al., 2009; Jutte et al., 2011; Harron et al., 2015, 2017).

1.4 Mortality trends and their determinants in South Africa

From the 1960s until the early 1990s overall mortality levels were declining and life expectancy was steadily improving among South Africans (United Nations, Department of Economic and Social Affairs, Population Division, 2011). Thereafter, South Africans experienced dramatic steady increases in overall mortality and reductions in life expectancy from the mid-1990s to the mid-2000s with the highest overall mortality and lowest life expectancy at birth levels around 2005-2007 (Pillay-van Wyk et al., 2016; Karim et al., 2009; United Nations, Department of Economic and Social Affairs, Population Division, 2011; Bradshaw et al., 2004; Tollman et al., 1999; Kahn et al., 2007a; Zwang et al., 2007). Nationally age-standardised death rates increased from 1 215 per 100 000 people in 1997 to a peak of 1 670 per 100 000 people in 2006 (Pillay-van Wyk et al., 2016) and life expectancy at birth declined from 63 years in 1995 to 54 years in 2005 (Mayosi and Benatar, 2014). Since then, overall mortality levels have been declining and life expectancy at birth steadily rising. Nevertheless, as of 2012, the levels of overall mortality (age-standardised death rate of 1 232 per 100 000 population) and life expectancy at birth (60 years) were still worse than those in the early 1990s (Pillay-van Wyk et al., 2016; Mayosi and Benatar, 2014).

The dramatic increases in overall mortality and reductions in life expectancy from the mid-1990s to the mid-2000s were driven mostly by increases in mortality caused by the large HIV/AIDS epidemic and lack of treatment programmes (Pillay-van Wyk et al., 2016; Kabudula et al., 2014b; Herbst et al., 2009, 2011; Karim et al., 2009). A heterosexually transmitted generalised HIV epidemic emerged in South

Africa between 1988 and 1993 and grew steadily until 2005 (Karim et al., 2009). The national HIV seroprevalence in pregnant women rose from 0.8% in 1990 to a peak of 30.2% in 2005 before stabilising at around 29.5% thereafter (Karim et al., 2009; National Department of Health, 2015). Using the "apartheid defined (racial) population groups (black, Indian or Asian descent, white [European descent], and coloured [of mixed ancestry according to the preceding categories])" (Pillay-van Wyk et al., 2016), the burden of the HIV epidemic is disproportionately higher among the black population and varies widely between provinces and local communities with the north-eastern parts of the country experiencing highest HIV prevalences and western parts of the country experiencing the lowest prevalences. As shown in Figure 1.1, overall prevalences among pregnant women in 2013 exceeded 30% in KwaZulu-Natal, Mpumalanga and Eastern Cape provinces, were between 20% and 30% in Free State, Gauteng, North West and Limpopo provinces and were below 20% in Western Cape and Northern Cape provinces (National Department of Health, 2015). Results from the 2012 national survey also showed a similar geographical pattern. Estimated HIV prevalence in the adult population exceeded 20% in KwaZulu-Natal, Mpumalanga, Free State and North West provinces and was between 12% and 20% in the rest of the provinces except for the Western Cape province where it was under 10% (Shisana et al., 2014).

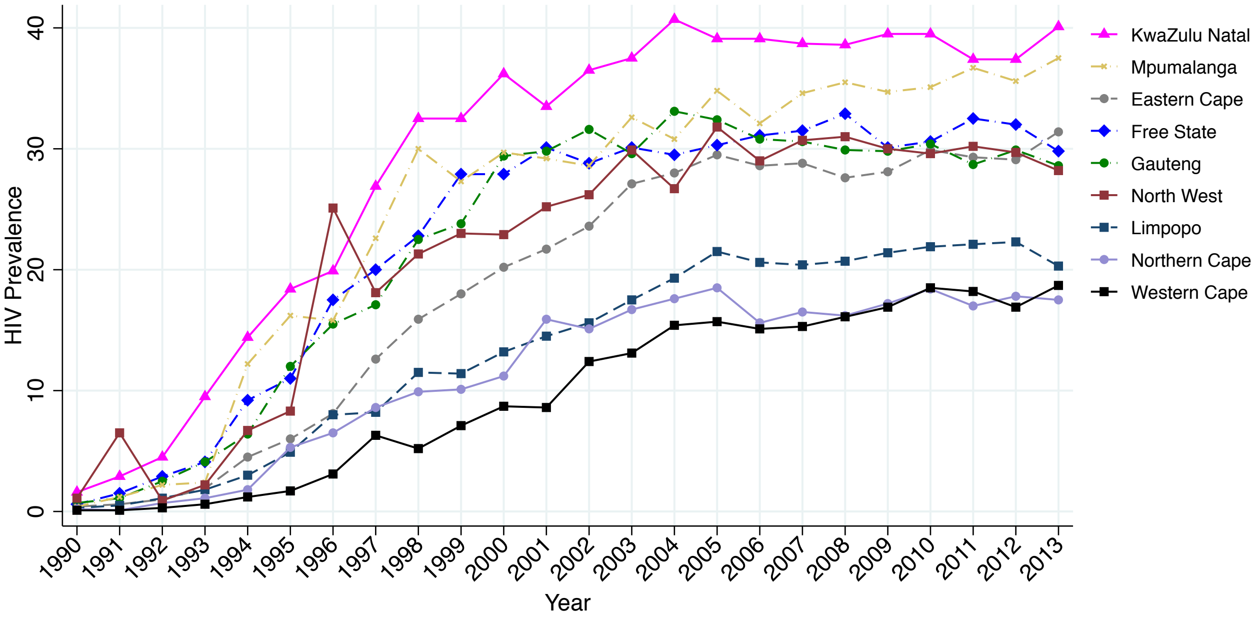


Figure 1.1: Trends in HIV Prevalence among 15-49 year old antenatal women in South Africa

Source: National Department of Health (2015)

The growth of the HIV/AIDS epidemic was mainly facilitated by the labour migration system created by the apartheid government from 1948 to 1994 (Coovadia et al., 2009). During the apartheid years, restrictions on access to land and means of production combined with coercive legislation and taxes enforced migration of black labourers (mostly males) to urban areas (Coovadia et al., 2009). The black migrant workers lived temporarily in the urban areas in overcrowded, unsanitary, single-sex hostels with regular visits to their rural homes because they were not provided proper housing and land was mostly reserved for white people in the urban areas (Coovadia et al., 2009; Karim et al., 2009). The oscillatory migration lifestyle affected the sexual practices in the black population. Usually the male labour migrants had sexual partners in both the urban areas and their rural homes and the majority of them formed second families in the urban areas (Coovadia et al., 2009). Women in the rural areas also often had another sexual partner while the men were away. These sexual practices provided a favourable environment for the efficient transmission of HIV infection among the black population (Coovadia et al., 2009; Karim et al., 2009). Although the apartheid system of government and racial segregation ended in 1994, the oscillatory migration lifestyle has persisted and continues to influence the spread of HIV infection in the black population.

Despite the dramatic increases in the prevalence of HIV infection in the general population, little was done by the South African governments to mitigate its impact on mortality until 2003 when president Thabo Mbeki's government started to provide antiretroviral therapy (ART) for free in public health services following a Constitutional Court order in favour of the Treatment Action Campaign and other civil society groups. However, the rollout of ART in the public health services progressed slowly until 2008. The availability of ART in the public health services became widespread from 2009 following the change in government leadership and by 2011 South Africa had the largest ART programme in the world with about 1.8 million people estimated to be taking antiretrovirals (Mayosi et al., 2012). Following the widespread availability of free ART in the public health services, HIV/AIDS-related and overall mortality has steadily declined and life expectancy at birth has steadily increased (Pillay-van Wyk et al., 2016; Kabudula et al., 2014b; Herbst et al., 2009, 2011).

Notwithstanding the importance of HIV/AIDS and that it remained the single leading cause of death during the period 1997-2012, the overall mortality burden in South Africa since the mid-1990s has been characterised by a unique quadruple disease burden consisting of HIV/AIDS and tuberculosis (TB); other communicable diseases (excluding HIV/AIDS and TB), maternal causes, perinatal conditions and nutritional deficiencies; NCDs; and injuries (Pillay-van Wyk et al., 2016; Herbst et al., 2011; Groenewald et al., 2010; Tollman et al., 2008; Bradshaw et al., 2005; Hosegood et al., 2004; Bradshaw et al., 2003). Causes from all these four broad cause groupings persistently constituted the top ten single causes of death over the period 1997-2012 (Pillay-van Wyk et al., 2016). In terms of magnitude, causes from the category of NCDs have been the second most important in the cause of death profile in South Africa following HIV/AIDS. Throughout the period 1997-2012, a considerable number of deaths were due to non-communicable diseases especially cardiovascular conditions including stroke, ischaemic heart disease and hypertensive heart disease (Mayosi et al., 2009; Pillay-van Wyk et al., 2016; Nojilana et al., 2016).

The contribution of NCDs to the overall mortality burden among South Africans has become more pronounced in recent years. An analysis of national cause-specific mortality data showed that in 2012, 43.4% of the deaths were attributed to NCDs, 33.6% to HIV/AIDS and tuberculosis, 13.5% to other communicable diseases, perinatal conditions, maternal causes, and nutritional deficiencies, and 9.6% to injuries (Pillay-van Wyk et al., 2016). Nevertheless, compared to other LMICs in Southeast Asia, the level of mortality attributable to NCDs in South Africa is lower. For example, as reported in a systematic review by Schröders et al. (2017), in Indonesia in 2012 about 71% of deaths were attributable to NCDs.

The increases in the contribution of NCDs to the overall mortality burden in recent years in South Africa have occurred most prominently among black people owing to population ageing and the adoption of lifestyle practices that expose individuals to a variety of risk factors for non-communicable diseases (Pillay-van Wyk et al., 2016; Mayosi and Benatar, 2014). Black people have experienced an increase in the prevalence of common risk factors for NCDs such as tobacco and alcohol use, physical inactivity, and a diet with high amounts of fat. The shifts in lifestyle practices in the black population have been driven mostly by modernisation coupled with a wide-range of social and economic reforms introduced by the post-apartheid government. The reforms include the provision of free basic services such as electricity (50 kWh per household per month), water, sanitation and housing and non-contributory social grants to vulnerable sectors of the population (Bhorat and van der Westhuizen, 2013; Collinson, 2010; Lund, 2002). These reforms which their implementation has vastly expanded since 1997, have resulted in the reduction of absolute poverty and enabled the growth of a middle class among the black population (Mayosi et al., 2012; Mayosi and Benatar, 2014). Nevertheless, racial disparities still persist for several indicators of socioeconomic development with white people being the most privileged and black people the worst off (Mayosi et al., 2012).

Overall, mortality patterns and levels in South Africa since the early 1990s have exhibited considerable racial, socioeconomic and geographical variations (Pillay-van Wyk et al., 2016; Nojilana et al., 2016; Msemburi et al., 2016). The mortality burden has consistently been highest among black people followed by coloured people and lowest among white people. The socioeconomically disadvantaged people also suffer the highest burden of mortality because many of them remain undiagnosed, untreated and at risk of preventable complications and have higher rates of ill-health than other groups (Nojilana et al., 2016). In addition, although a quadruple burden of disease is evident in each province, there are considerable variations in the levels of overall mortality and its cause composition between the nine provinces in South Africa (Pillay-van Wyk et al., 2016; Msemburi et al., 2016). For example, in 2012 the overall age-standardised death rate was highest in KwaZulu-Natal province (1 576 deaths per 100 000 people with 537 deaths per 100 000 people due to HIV/AIDS and TB), and lowest in Western Cape province

(938 deaths per 100 000 people with 149 HIV/AIDS and TB deaths per 100 000 people) (Msemburi et al., 2016).

1.5 Statement of the problem

As presented in the preceding section, changes in mortality patterns and their cause composition in South Africa over the past two decades have revealed elements of counter-transition, protractedness and polarisation of the epidemiological transition. The recently published second National Burden of Disease Study in South Africa showed that all-cause age-standardised mortality increased rapidly from 1997 as a result of a severe HIV epidemic, peaked in 2006 and then declined following the rollout of ART (Pillay-van Wyk et al., 2016; Nojilana et al., 2016; Msemburi et al., 2016). The study further showed that throughout the past two decades, despite the changes in overall mortality levels, the top ten single causes of death persistently included causes from four broad cause groupings consisting of (i) HIV/AIDS and tuberculosis, (ii) other communicable diseases, maternal causes, perinatal conditions and nutritional deficiencies, (iii) non-communicable diseases and (iv) injuries. However, **the proportions and rankings of the top causes of death did not remain constant**. The study also found considerable racial, socioeconomic and geographical variations in the mortality patterns and levels.

It is evident from the findings reported in the second National Burden of Disease Study that mortality patterns and levels in South Africa have evolved in diverse ways across sub-populations over the past two decades. However, more up to date formal and empirical assessments of the extent to which mortality patterns and levels have changed in rural settings of South Africa covering periods of the HIV epidemic without and with ART are limited. The studies in this thesis based on rigorous, comprehensive health and socio-demographic surveillance, update and extend published trends in mortality and cause of death profiles in the rural population in Agincourt (Kahn et al., 2007a; Kabudula et al., 2014b; Houle et al., 2014a) by including data from more recent years that cover the widespread availability and uptake of ART. The studies also complement the second National Burden of Disease Study in South Africa, which investigated differences between ethnic groups and provinces in mortality patterns and levels, by examining socioeconomic differences at the local level in a well-characterised resource-poor rural setting.

While the rigorous surveillance data collected through a health and socio-demographic surveillance system make it possible to investigate changes in mortality patterns, levels and some of their determinants, it is not possible, using the surveillance data alone, to assess the contribution of utilisation of health services to the changes observed in mortality patterns. Such an assessment requires integration of information on service usage patterns with the population surveillance data. Hence, this thesis also assesses the feasibility of using linkage to clinic records so as to integrate data from health facility registers with the population surveillance data.

1.6 Research aims, questions and hypotheses

1.6.1 Overall aim

The overall aim of the research reported in this thesis is to contribute to the understanding of changes in the cause composition of overall mortality and associated socioeconomic differentials that have occurred in a rural population of the Agincourt sub-district of northeast South Africa over the period 1993 to 2013. The knowledge about changes in the cause composition of overall mortality and associated socioeconomic differentials is important when setting health and development priorities, identifying critical elements for strengthening of health systems, and determining the focus of health services and programmes. A further aim is to assess the feasibility of linking population data from the HDSS and utilisation data from health facility registers in order to enhance the utility of HDSS data for providing comprehensive insights regarding mortality and disease patterns and their determinants and implications in the Agincourt population.

1.6.2 Specific objectives

The specific objectives of this thesis are:

i. To assess changes in household socioeconomic status (SES) in the population of the Agincourt sub-district in rural northeast South Africa over the period 2001 - 2013. (**Paper I**)

ii. To assess changes in mortality cause composition in the population of the Agincourt sub-district in rural northeast South Africa over the period 1993 - 2013. (**Paper II**)

iii. To assess temporal trends in socioeconomic differentials in the major cause categories of mortality in the population of the Agincourt sub-district in rural northeast South Africa over the period 2001 - 2013. (**Paper III**)

iv. To evaluate the feasibility of linking population data from HDSS and utilisation data from health facility registers in order to study patterns in utilisation and access to health services in the Agincourt sub-district. (**Paper IV**)

1.6.3 Research questions

The following are the questions investigated in this thesis and they relate closely to the objectives listed above:

i. How has household wealth evolved in the population of the Agincourt sub-district in rural northeast South Africa over the period 2001 - 2013? (**Paper I**)

ii. What have been the changes in the cause composition of overall mortality in the population of the Agincourt sub-district in rural northeast South Africa over the period 1993 - 2013? (**Paper II**)

iii. How has the socioeconomic gradient in the major cause categories of mortality evolved in the population of the Agincourt sub-district in rural northeast South Africa over the period 2001 - 2013? (**Paper III**)

iv. To what extent have changes in mortality patterns in the Agincourt sub-district in rural northeast South Africa over the period 1993 - 2013 followed the propositions of the classical epidemiological transition theory and subsequent refinements ? (**Papers II & III**)

v. How feasible is linking of population data from HDSS and clinical data from health facility registers in order to study socioeconomic differentials in access to health services in the Agincourt sub-district? (**Paper IV**)

1.6.4 Hypotheses

The hypotheses of the thesis include:

i. The Agincourt population has experienced substantial improvements in household wealth over the period 2001 - 2013. (**Paper I**)

ii. The contribution of non-communicable disease mortality to overall mortality in the population of the Agincourt sub-district has increased following a reduction in HIV/AIDS-related mortality as a result of introduction and expansion in coverage of ART programs. (**Paper II**)

iii. The socioeconomic gradient in HIV/AIDS mortality in the population of the Agincourt sub-district increased following the roll-out of ART programmes. However, the direction of the gradient has not changed significantly over the period 2001 - 2013. (**Paper III**)

iv. The changes in mortality patterns in the Agincourt sub-district in rural northeast South Africa over the period 1993 - 2013 diverge from some of the propositions of the classical epidemiological transition theory. (**Papers II & III**)

v. Population data and health facility data in the rural Agincourt sub-district can be linked effectively by a combination of deterministic and probabilistic record linkage approaches using a set of conventional identifiers (such as name, sex, date of birth and place of residence). (**Paper IV**)

1.6.5 Thesis themes

Table 1.1: Key themes addressed in the thesis papers

Themes	**Papers**			
	Paper I Changes in household SES	Paper II Changes in mortality cause composition	Paper III Socioeconomic differentials in mortality	Paper IV Record linkage of HDSS and health facility data
Measuring SES	✓			
Socioeconomic transition	✓			
Mortality transition		✓	✓	
Socioeconomic differentials in mortality			✓	
Feasibility of using record linkage to enhance the utility of HDSS data				✓

1.7 Conceptual framework

The investigation of changes in the cause composition of overall mortality and associated socioeconomic differentials that occurred in the Agincourt population in rural northeast South Africa over the period 1993 - 2013 in this thesis is guided by the conceptual framework shown in Figure 1.2. This conceptual framework is based on Mosley and Chen's analytical framework of proximate determinants of the health dynamics of a population (Mosley and Chen, 1984). Obviously, this framework is a simplification of reality, and includes variables for which data have been continuously collected in the Agincourt surveillance population over an extended period of time and are relevant to the investigation in this thesis. Age, sex, SES, and healthcare services are conceptualised as the key independent factors that affected the levels of overall and cause-specific mortality in the Agincourt population over the period 1993-2013 through their influence on health behaviour, lifestyles and the use of preventive, diagnostic and therapeutic healthcare services which directly influence the risk of morbidity and mortality. On the basis of our conceptual framework, we assessed changes in: (i) SES using indices computed from a list of household asset indicators (**Paper I**); (ii) overall and cause-specific mortality levels by sex, age and time period (**Paper II**); and (iii) overall and cause-specific mortality levels by sex, age, time period and SES (**Paper III**). In addition, we also assess the feasibility of linking population surveillance data with data from health facility registers (**Paper IV**). The linkage of population surveillance data with data from health facility registers, which never existed prior to our study, is central to studying the contribution of differentials in the utilisation of health services to differentials in mortality levels.

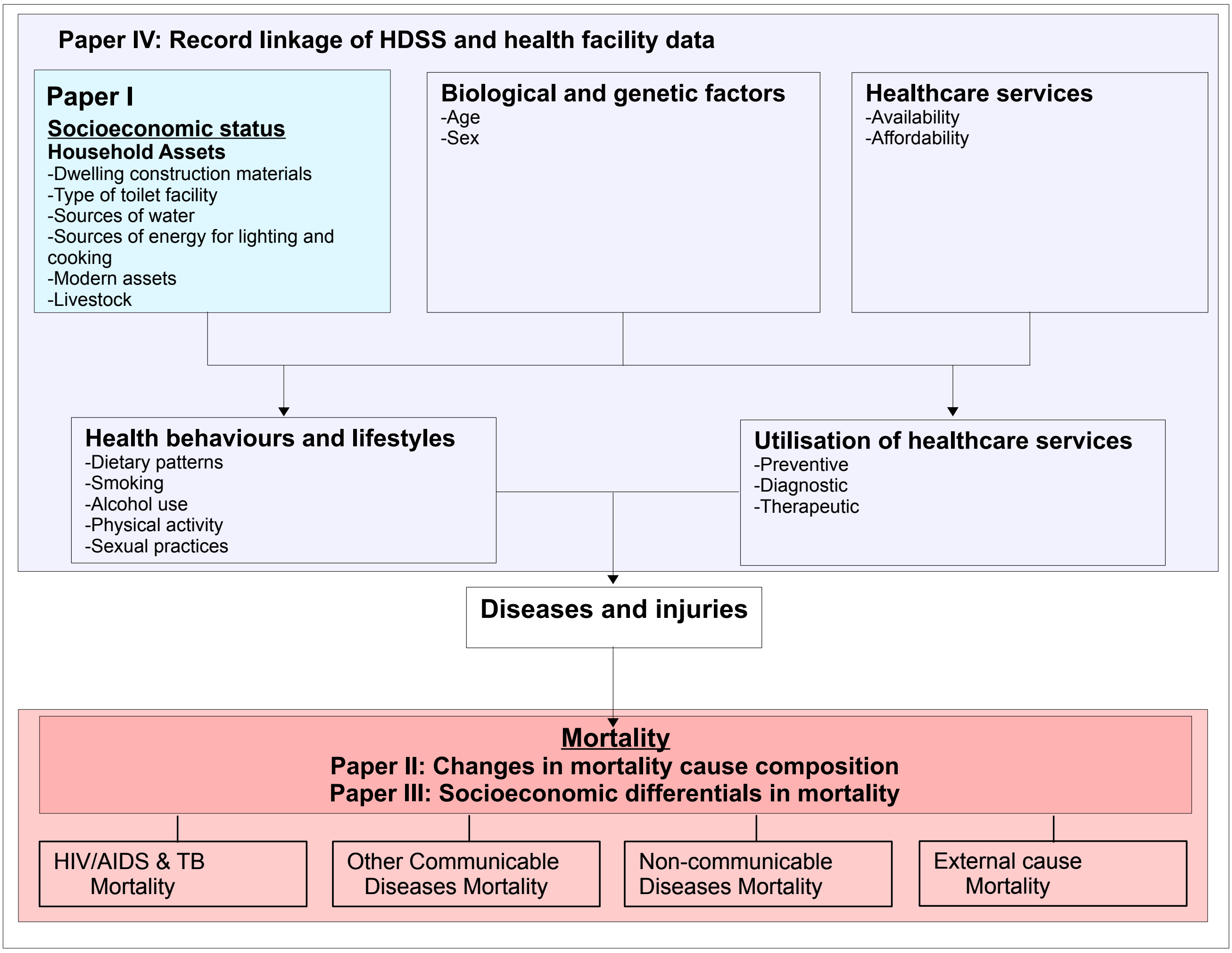

Paper IV: Record linkage of HDSS and health facility data
Paper I
Socioeconomic status
Household Assets
-Dwelling construction materials
-Type of toilet facility
-Sources of water
-Sources of energy for lighting and cooking
-Modern assets
-Livestock
Biological and genetic factors
-Age
-Sex
Healthcare services
-Availability
-Affordability
Health behaviours and lifestyles
-Dietary patterns
-Smoking
-Alcohol use
-Physical activity
-Sexual practices
Utilisation of healthcare services
-Preventive
-Diagnostic
-Therapeutic
Diseases and injuries
Mortality
Paper II: Changes in mortality cause composition
Paper III: Socioeconomic differentials in mortality
HIV/AIDS & TB Mortality
Other Communicable Diseases Mortality
Non-communicable Diseases Mortality
External cause Mortality

Figure 1.2: Thesis papers in relation to determinants of health conceptual framework

Chapter 2

Data and Methods

2.1 Study Setting

The research upon which this thesis is based was conducted in the Agincourt HDSS. The study area was established in 1992 and is located in a predominantly rural Sub-district of Bushbuckridge, Ehlanzeni District, Mpumalanga province, in north-eastern South Africa. From 1992 to 2006, the study area encompassed 21 villages spread over 402 km^2 of semi-arid scrubland (Kahn et al., 2007b). Then in 2007 the study area was extended to 26 villages. Another five villages were added in response to an expanding trials and evaluation portfolio between 2010 and 2012 (Kabudula et al., 2016). Presently, the HDSS covers 30 contiguous villages spread over 450 km^2 and is following a study population of some 115,000 people in 21,000 households (Figure 2.1). The population is largely Shangaan-speaking and almost a third is made up of former Mozambican refugees who arrived in the area in the early to mid-1980s, and their descendants.

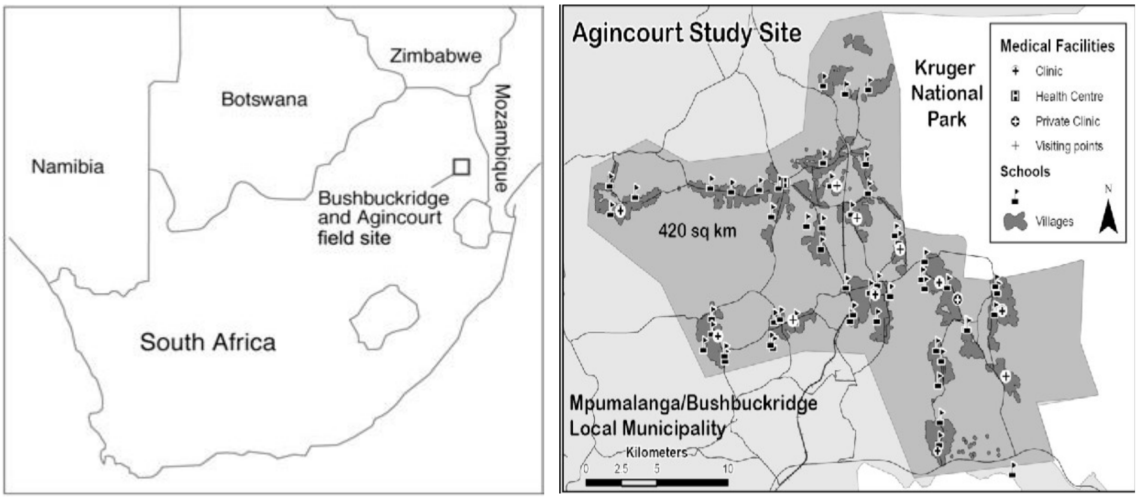


Figure 2.1: Location of Agincourt HDSS study area, Mpumalanga province, South Africa

2.2 Study design

The studies in the first three papers (**Paper I, II** & **III**) fall into a prospective longitudinal observational study design and the data on which they are based on was collected on a regular basis from the whole population of the Agincourt HDSS study area since 1992. The study in **Paper IV** is methodological and utilises surveillance data and health facility-based data.

2.3 Data

The specific data elements used in the study in each paper are presented in Table 2.1 and the data collection process is described below.

2.3.1 Health and socio-demographic surveillance system data

The description of the data collection process for the longitudinal health and socio-demographic data closely follows that reported by (Kahn et al., 2007b). Following the establishment of the study area, a baseline census was conducted in 1992. Each household in the study area was visited and every resident was registered. Since then, data have been collected on birth, death, in- and out-migration events at approximately 15-18-month intervals between 1993 and 1999 and on strictly annual basis since 1999. Data pertinent to studying other aspects of health and population dynamics have also been collected in the form of special modules. These modules have been collecting data at the individual as well as household level. Most of the modules were introduced from 2000 and each module is repeated at specific time intervals. A list of some of the modules and the years in which data were collected is provided in Figure 2.2. The papers of this thesis mainly use data from the household assets module.

The household assets module collects data on household asset indicators that include construction materials and structure of the main dwelling, type of toilet facilities, sources of water and energy, ownership of modern assets and livestock. The module was introduced in 2001 and was repeated every 2 years between 2001 and 2013 (Kahn et al., 2012). To assess changes in the asset indicators over the period 2001 to 2013, **Paper I** uses only the data collected from households in the original 21 villages.

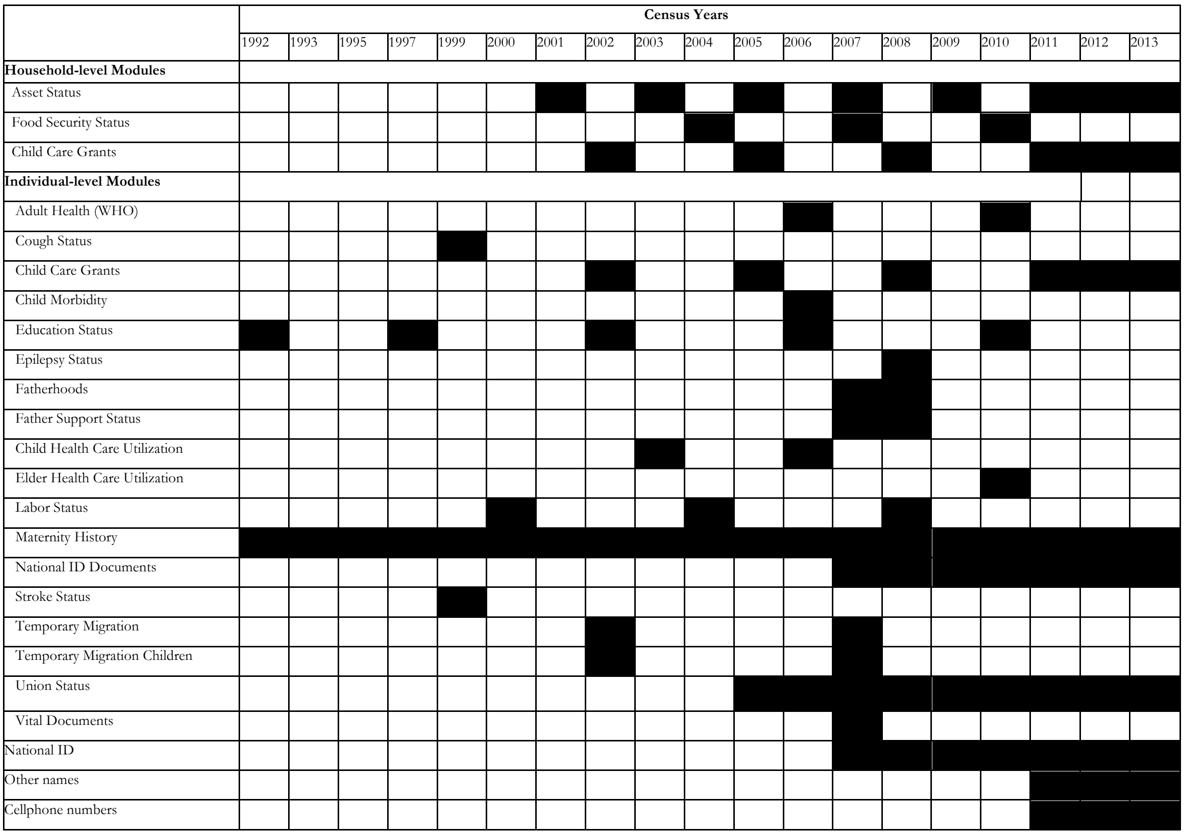
In the case of death events, in addition to collecting data on place of death, name of hospital if death occurred at the hospital and whether or not the death was registered, data has also been collected on all symptoms and signs preceding the death through verbal autopsy (VA) interviews. For each recorded death, the VA data have been collected from the closest caregiver of the deceased between one month and one year of the death. According to (Kahn et al., 2007b), the questionnaire that was used in collecting the VA data until 2011 was a modification of the one previously used in Niakhar, Senegal. From 2012

Table 2.1: Data for thesis papers

Study	Data
Paper I: Assessing changes in household wealth in the Agincourt sub-district from 2001 to 2013.	Household characteristics such as water source, toilet facilities, and construction materials and consumer durables such as ownership of radio, television, refrigerator, bicycle, motorcycle and car collected from 2001 to 2013
Paper II: Assessing changes in the cause composition of overall mortality in the Agincourt sub-district between 1993 and 2013	- Dates of birth, and sex of all individuals who lived in the study area since 1992 - Dates of death and likely cause of death of all individuals who died in the study area since 1992 - Dates of in-migration of all individuals who in-migrated into the study area since 1992 - Dates of out-migration of all individuals who out-migrated from households in the study area since 1992
Paper III: Assessing changes in socioeconomic differentials in mortality in the Agincourt sub-district from 2001 to 2013.	- All the data used in Papers I & II from 2000 to 2013
Paper IV: Assessing the feasibility of linking population data with health facility data in order to measure socioeconomic differentials in access to health services in the Agincourt sub-district.	- Identifying information such as name, surname, date of birth of individuals in the Agincourt HDSS - Identifying information such as name, surname, date of birth of all health facility attendees

a new VA questionnaire was introduced. The items on the new questionnaire were aligned to the WHO 2012 VA standards.

The VA data have been used to assign the likely cause of death to each recorded death. Until recently, the approach for assigning cause of death had mainly been that two doctors independently reviewed the VA data on each death and assigned a probable cause. If the two doctors assigned the same cause then that cause was assigned to the particular death as its likely cause. In the case where the two doctors assigned different causes, the doctors discussed the case in an effort to reach a consensus. The consensus diagnosis was then assigned to the particular death as its likely cause. Where consensus was not reached,



	Census Years																		
	1992	1993	1995	1997	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Household-level Modules																			
Asset Status							■		■		■		■		■		■	■	■
Food Security Status										■			■			■			
Child Care Grants								■			■			■			■	■	■
Individual-level Modules																			
Adult Health (WHO)												■				■			
Cough Status					■														
Child Care Grants								■			■			■			■	■	■
Child Morbidity												■							
Education Status	■			■				■				■				■			
Epilepsy Status														■					
Fatherhoods													■	■					
Father Support Status													■	■					
Child Health Care Utilization									■			■							
Elder Health Care Utilization																■			
Labor Status						■				■				■					
Maternity History	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
National ID Documents													■	■	■	■	■	■	■
Stroke Status					■														
Temporary Migration								■					■						
Temporary Migration Children								■					■						
Union Status											■	■	■	■	■	■	■	■	■
Vital Documents													■						
National ID													■	■	■	■	■	■	■
Other names																	■	■	■
Cellphone numbers																	■	■	■

Figure 2.2: Agincourt HDSS special census modules

a third doctor reviewed the data and assigned a probable cause. If the cause assigned by the third doctor matched with the cause assigned by one of the first two doctors then that cause was assigned to the particular death as its likely cause. In the case where all three doctors assigned different causes, the cause of death for the particular death got coded as “ill-defined”. In the recent years, a computer-based probabilistic model called InterVA has been used to assign causes of death from the VA data collected in the study area (Byass et al., 2010, 2011). All the papers assessing cause of death patterns in this thesis have used the InterVA probabilistic model (version 4.03) to assign causes of death.

Respondents to the household interviews have mostly been senior responsible adult members of the households. The norm has been to carry out a maximum of two visits per household to find the best respondent following which a neighbour has been used as a proxy informant for basic vital status information. In cases where the best respondent including a neighbour could not be found, collection of data on vital events has been done during the subsequent round unless the household permanently out-migrated.

Several quality control processes are put in place to ensure that the Agincourt HDSS data is of good quality. Firstly, fieldworkers check all the forms that they complete on daily basis. Secondly, supervisors check a random selection of the forms collected by the fieldworkers in their time. Thirdly, a team of specialised "quality checkers" thoroughly checks all forms for completeness and errors. The quality checkers send correct forms to the data room for data entry and return forms with errors to the field for correction. Lastly, validation checks embedded in the data entry system flag and filter out implausible and inconsistent data during data entry. Forms identified with errors during data entry are also returned to the field for correction.

During the early years the Agincourt HDSS data was entered and stored in a FoxPro data management system. The data management system was changed to Microsoft Access 95 along with improvements in the data model in 1996. It was thereafter upgraded to Microsoft Access 97 in 1999 with the data model brought up to the standard of the INDEPTH population reference data model (Benzler et al., 1998). In 2002 the data management system was upgraded from the Microsoft Access platform to Microsoft SQL Server platform. Currently the operational database is hosted on Microsoft SQL Server 2008 R2. A custom-made data-entry program, which mirrors the format of the data collection forms, sits on top of the SQL Server database. In the SQL Server database, the data is stored in a set of tables containing data fitted into predefined categories. Each table contains one or more data categories in columns and each row contains a unique instance of data for the categories defined by the columns. The main table, the "Individuals" table, stores key information on all individuals that have ever been in the surveillance population; the "Residences" table stores information on individual residence episodes within the surveillance area; a "Memberships" table stores information on entry and exit from a particular household and the relationship between each individual and the head of the household, and there is a separate table for each vital event (for example death) and special module (such as the household asset survey). The database is secured using two levels of user-access control: a password to log onto the operating system, and another password to log onto the database. Backups of the data are produced weekly and a copy is kept in another office located 40 km from the field office where data entry is conducted.

2.3.2 Health facility data

The health facility data used in the study in **Paper IV** came from the Agincourt Health Centre. The Agincourt Health Centre is one of eight local health facilities within the Agincourt HDSS study area. An electronic patient registration system was piloted at the health centre from 2008 to 2010. Clerks captured personal identifying information and fingerprints from all individuals that used the health centre. For the purpose of this study only the identifiers that were collected in both the health facilities and the Agincourt HDSS were used. The fingerprints were used to construct a gold standard dataset of matched record pairs that were used to evaluate the performance of various record linkage scenarios based on conventional personal identifiers.

2.4 Data Analysis

2.4.1 Measuring household SES

Household SES in **Paper I** and **III** was measured by a household asset index. The index was constructed from information on household asset indicators such as construction materials and structure of the main dwelling, type of toilet facilities, sources of water and energy, ownership of livestock and modern assets such as radio, television, refrigerator, bicycle, motorcycle and car. In **Paper I**, three indices were constructed using different analytic approaches. The indices were the (i) absolute index, (ii) principal component analysis (PCA) index, and (iii) multiple correspondence analysis (MCA) index. Only the absolute index was used in **Paper III** because of its comparability across time without the need to pool the data and that its performance is comparable to the other indices in assessing household SES using household assets as illustrated in **Paper I**. The analytic approaches used to construct the three indices are described in detail in **Paper I**.

2.4.2 Assessing changes in household SES

In **Paper I**, the method of relative distributions developed by Mark Handcock and Martina Morris (Handcock and Morris, 1999, 1998) was employed to assess changes in differentials in household wealth. The method quantifies differences between the distributions of a set of random measurements of an attribute of interest from a population at one time period and another set of random measurements of the same attribute from a different population or from the same population at a later time period. In the case of **Paper I** the attribute of interest was household wealth as measured by the asset index. Following Handcock and Morris (1999, 1998), we denoted Y_0 and Y as random variables representing asset index scores from the Agincourt HDSS population in 2001 and at a later time period respectively. Denoting the cumulative distribution functions (CDFs) of Y_0 and Y as $F_0(y)$ and $F(y)$ respectively, the relative distribution of Y to Y_0 was be obtained as

$$R = F_0(Y) \tag{2.1}$$

by transforming Y by the CDF of Y_0, F_0. With this transformation, R provided a measure of the relative rank of Y compared to Y_0. A uniform or flat probability density function (PDF) of R indicated that there were no differences between the distributions of Y_0 and Y. When there were differences between the distributions, the relative distribution "rose" or "fell" depending on the direction of the difference.

2.4.3 Data file for mortality analysis

Similar to an earlier study conducted in Agincourt by Clark et al. (2007), a person-year file was constructed containing one record for each year lived by each individual in the study population during the period 1993-2013. Attributes contained in each record consisted of Individual ID, sex, date of birth, date of death, age, calendar year, level of education completed, the most recent indicator of household SES and whether or not the person died within the year. If there was a death, the three probable causes of death generated by the InterVA model were included. All analyses regarding mortality and socioeconomic differences in mortality were conducted using this person-year file. The years covered in the analysis were split into the following time periods: 1993-1997, 1998-2000, 2001-2003, 2004-2007, 2008-2010 and 2011-2013 to contextualize the dynamics of the HIV epidemic and the roll out of prevention of mother to child transmission (PMTCT) and ART services. In addition, age was also categorized into the following commonly used age groups: 0-4, 5-14, 15-49, 50-64 and 65+.

2.4.4 Estimating mortality risk

In the studies reported in **Papers II** and **III**, discrete time event history analysis (Allison, 1984) was used to estimate the annual hazard of death as a function of independent variables. Annual hazards of death from all possible causes were estimated using binary logistic regression models. Estimates of the annual hazard of cause-specific mortality were obtained using multinomial logistic regression models. The binary logistic regression models took the form:

$$\log\frac{P_{ti}}{1-P_{ti}} = \alpha(t) + \beta' x_{ti} \tag{2.2}$$

where P_{ti} is the probability that an individual i dies at time t, given that the individual is still alive at time t, x_{ti} is a vector of covariates, $\alpha(t)$ is a function of time and β' is a vector of parameters to be estimated.

The estimated annual hazards of death, were then used to construct standard life tables to derive estimates of life expectancies at various ages including birth. The life tables were also used to estimate adult mortality rates (the probability of dying between ages 15 and 60 for those who survive to age 15 if subjected to age-specific mortality rates between those ages for the specified calendar year). *For complementary analyses reported only in this thesis, cause deleted life tables were constructed to assess years of life expectancy gained if selected causes were deleted.*

2.4.5 Assessing changes in the cause composition of overall mortality

In the study reported in **Paper II**, consistent with the standard approach used by the World Health Organisation (WHO) and other agencies, causes of death based on verbal autopsies were classified into three broad groups (Lopez et al., 2006). Group I consisted of communicable diseases, maternal, and perinatal conditions and nutritional deficiencies, group II of NCDs and group III of external causes.

Taking into account that: (i) the proportion of deaths attributed to each cause can range only from 0 to 1 ; and (ii) the set of proportions for all of the cause groups must sum to 1, the method suggested by Salomon and Murray (2002) was used to relate the distribution of mortality from different cause groups to the overall mortality levels. The method involves fitting cause specific mortality data with a set of simultaneous regression equations. First a vector of cause fractions: $P_i = (P_{i1}, \cdots, P_{iJ})$ is defined for each observation for J different cause groups. Thereafter, a $(J-1)$ vector Y_i, is generated by calculating the log ratios of each cause fraction relative to the fraction for cause J as shown in equation (2.3):

$$Y_{ij} = \ln\left(\frac{P_{ij}}{P_{iJ}}\right) \tag{2.3}$$

Then in the case of three cause groups, the distribution of mortality from different cause groups to the overall mortality levels is assessed using the following system of two equations:

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 \ln(M_i) + \epsilon_{i1} && (2.4) \\ Y_{i2} &= \gamma_0 + \gamma_1 \ln(M_i) + \epsilon_{i2} && (2.5) \\ & && (2.6) \end{aligned}$$

where Y_{i1} and Y_{i2} are the log-ratios as defined in Equation (2.3), M_i is the all-cause mortality rate and ϵ_{i1} and ϵ_{i2} are error terms. Estimates of the coefficients for the model were obtained using seemingly unrelated regression models. The motivation for this is that it accounts for correlations in the error terms and the full covariance structure of the coefficient.

2.4.6 Assessing socioeconomic differentials in mortality patterns

In **Paper III**, socioeconomic differences in the mortality indicators were quantified both in absolute and relative terms. Relative differences were estimated using the relative index of inequality (RII) (Wagstaff et al., 1991) and relative risk ratios (RRR) associated with household wealth quintile obtained from multinomial logistic regression models (Allison, 1982, 1984, 2010; Efron, 1988; Van Hook and Altman, 2013). Absolute differences were estimated using the slope index of inequality (SII) (Wagstaff et al., 1991).

Both RII and SII are regression based summary measures that take into account the whole SES distribution rather than only the two most extreme SES groups (e.g the richest and poorest SES quintiles) when quantifying the magnitude of SES differences in health outcomes. The RII provides the ratio and the SII provides the absolute difference of the mortality outcomes of those at the bottom of the SES hierarchy compared with those at the top of the hierarchy, estimated on the basis of the systematic association between mortality and SES for all groups (Mackenbach and Kunst, 1997). RII = 1 implies that mortality in the lower and higher ends of the socioeconomic continuum do not differ, RII > 1 implies greater mortality at the lower end, and RII < 1 implies greater mortality at the higher end. SII = 0 indicates that mortality at the lower and higher ends of the socioeconomic continuum do not differ, a positive SII indicates greater mortality at the lower end, and a negative SII indicates greater mortality at the higher end (Mackenbach and Kunst, 1997).

The RII and SII were calculated for socioeconomic differences in the probability of death from birth to age 5 years (${}_5q_0$), probability of death from age 15 to 60 years (${}_{45}q_{15}$), life expectancy at birth (e_0) and cause-specific mortality. The steps involved in calculating the RII and SII as described in **Paper III** follow those by Mackenbach and Kunst (1997), Korda et al. (2007) and Ernstsen et al. (2012). First, the population in each SES quintile was assigned a modified 'ridit' score representing the relative rank of that SES quintile in the cumulative distribution of household SES. The values of the modified ridit score ranged from 0 (richest) to 1 (poorest) and were calculated by arranging the household wealth quintiles in order from richest to poorest and assigning a cumulative proportion of the total population to each quintile. Thereafter, half the proportion of the total population in the fifth quintile was taken as the modified ridit score for the fifth quintile and half the proportion of the total population in the fourth quintile added to the proportion in the fifth quintile as the modified ridit score for the fourth quintile and so on. An illustrative example of the calculation of the modified ridit scores is presented in Table 2.2. Finally, estimates of the RII and SII were obtained using generalised linear models with a Gaussian distribution of the form:

$$g(Y) \quad = \quad \beta_0 + \beta_1 rscore + \epsilon \tag{2.7}$$

where Y is the mortality indicator (${}_5q_0$ or ${}_{45}q_{15}$ or e_0), $g(Y) = Y$ is an identity link function when the indicator of interest is SII or $g(Y) = log(Y)$ is a logarithm link function when the indicator of interest is RII , β_0 is a constant, β_1 is the beta or slope coefficient on modified ridit scores which expresses RII when the logarithm link function is used, and SII when the identity link function is used, *rscore* is the modified ridit score (representing the relative ranks of groups in the cumulative distribution of household SES) and ϵ is the error term.

As described in **Paper III**, since the data on household asset indicators used for calculating the household wealth index were collected in alternate years from 2001 to 2013, multiple imputation was used to minimise

Table 2.2: Calculation of modified ridit scores

SES quintile	Proportion distribution of population	Cumulative proportion of population	Modified ridit score
Fifth quintile (richest)	0.23	0.23	0.23/2 = 0.12
Fourth quintile	0.22	0.23 + 0.22 = 0.45	0.23 + (0.22/2) = 0.34
Third quintile	0.20	0.23 + 0.22 + 0.20 = 0.65	0.45 + (0.20/2) = 0.55
Second quintile	0.18	0.23 + 0.22 + 0.20 + 0.18 = 0.83	0.65 + (0.18/2) = 0.74
First quintile (poorest)	0.17	0.23+0.22+0.20+0.18+0.17 = 1.00	0.83 + (0.17/2) = 0.92

the loss of data due to missing values. Partial mean matching (based on the nearest two neighbours) was used to generate five imputed datasets and derive parameter estimates and SEs by averaging across the imputations and adjusting for variance. Similar to Houle et al. (2016), the imputations were generated from a household-year data set that includes counts of men, boys, women, and girls, Mozambicans and South Africans, individuals aged younger than 20 years, 20-59 years, and 60 years and older, and 1-2 year lags of household wealth index. The imputations were generated for households with missing assets information for the years where household assets were collected and for all households for the years where no household assets were collected.

2.4.7 Assessing the feasibility of record linkage of health facility data and demographic surveillance data

In **Paper IV**, deterministic and probabilistic record linkage approaches were used to link data from the Agincourt HDSS with data from health facilities. Deterministic record linkage designates a pair of records from two data sources as belonging to the same individual when they match on a unique identifier such as fingerprints, a social security or national identification number, or a set of conventional personal identifiers (e.g., the combination of first name, last name and date of birth) (Li et al., 2006; Machado, 2004; Maso et al., 2001; Victor and Mera, 2001). Probabilistic record linkage classifies a pair of records from two data sources as belonging to the same individual based on the statistical probability that common identifiers drawn from the two data sources belong to the same individual (Howe, 1998; Beauchamp et al., 2011; Cook et al., 2001; Jaro, 1989, 1995; Nitsch et al., 2006). Whereas deterministic linkage is most suitable when unique identifiers are available and the quality of the data are high, probabilistic linkage yields better results when unique identifiers are lacking or in situations where there is variation in reporting or transcription of personal identifiers (Li et al., 2006; Beauchamp et al., 2011; Clark, 2004; Pacheco et al., 2008; Rosman, 1996).

Probabilistic record linkage approaches involve estimating weights for different identifiers and using those weights to compute a composite score that determines whether a pair of records from two different data sources pertain to the same individual (Jaro, 1989; Cook et al., 2001; Victor and Mera, 2001). For each linking identifier i, the process involves first estimating the probability that the identifier agrees between two records given that the records pertain to the same individual denoted by m_i and the probability that the identifier agrees between two records given that the records do not pertain to the same individual denoted by u_i (Jaro, 1989; Cook et al., 2001; Nitsch et al., 2006). Once the parameters m_i and u_i have been estimated, a weight value of $\log_2(m_i/u_i)$ gets assigned to a record pair j if identifier i agrees and a weight value of $\log_2((1-m_i)/(1-u_i))$ if identifier i disagrees. The weights on all the identifiers are then summed to obtain a linkage score for each record pair. Scores for all record pairs are then analyzed to determine the optimum threshold value at which record pairs would be classified as matches or non-matches. In **Paper IV**, we followed established practice in probabilistic record linkage and used the expectation maximization (EM) algorithm based on the Fellegi-Sunter model (Fellegi and Sunter, 1969) to obtain estimates of the m_i and u_i parameters from the data (Winkler, 1988; Jaro, 1989; Grannis et al., 2002).

2.5 Ethics

Ethics approval was obtained from the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, Johannesburg, South Africa, for surveillance activities in the Agincourt HDSS (protocols M960720 and M110138), record linkage of health facilities registries with Agincourt HDSS database (protocol M071141) and for the analyses reported in this study (protocol M120488). Verbal informed consent was obtained at every surveillance visit from the head of the household or another eligible adult in the household. The person giving consent was noted in the household roster, and the details and date of the process were recorded by the responsible fieldworker. Further verbal informed consent was obtained for the collection of fingerprints and to link the Agincourt HDSS database record to any visits to Agincourt Health Centre, which is one of eight local health facilities within the Agincourt HDSS.

Chapter 3

Findings on changes in household SES in Agincourt, rural South Africa, 2001-2013

3.1 Measuring SES

People's SES in low- and middle-income settings is widely measured using a composite index known as a "wealth index" or "asset index" (Howe et al. 2012) which is constructed from information on household assets and dwelling characteristics (Ataguba et al., 2011; Barros et al., 2010; Gwatkin et al., 2007; Hong and Mishra, 2011; Hosseinpoor et al., 2006; Minujin and Delamonica, 2004; Nkonki et al., 2011; Uthman, 2009; Van de Poel et al., 2008; Ziraba et al., 2009). A number of theoretical reasons ranging from reliability to time and cost effectiveness (Balen et al., 2010; Howe et al., 2009, 2012; Montgomery et al., 2000; Sahn and Stifel, 2003) influence the preference of the asset index as a measure of SES over "direct" measures of SES such as income, expenditure, and financial assets (e.g., savings and pensions). The reasons include that the construction of the asset index requires information that is relatively easy and inexpensive to collect. In addition, in low- and middle-income settings household assets also provide a better proxy for a household's long-run wealth compared to information on income or expenditures. This is because information on income or expenditures is affected by seasonal variability in earnings, income from potentially multiple and diverse informal activities, high rates of self-employment, likely recall bias and misreporting.

One of the debates about the measurement of SES using household assets involves the analytic procedures used to construct the asset indices. **Paper I** contributes to this debate by constructing and comparing three asset indices that are widely reported in the literature namely the absolute index, Principal Components Analysis (PCA) index and Multiple Correspondence Analysis (MCA) index. Following Howe et al. (2008), the three indices are compared with each other using scatter plots and the percentage of households classified into the same and different SES quintiles. Then the agreement of classification of households into SES quintiles between indices is assessed using Kappa statistics which takes values between 0 (no agreement better than chance) and 1 (perfect agreement). Similar to Balen et al. (2010), further comparisons of the indices is done using the Spearman's rank correlation coefficient. The results show that the indices are reasonably comparable despite differences in the weights assigned to the asset items in the three indices. Pairwise comparisons of the values of the indices shown in Table 2 of **Paper I** and reproduced in Table 3.1 yield correlation coefficients of at least 0.95 and each pair of indices assigns at least 71% of households in the same SES quintile with Kappa statistics of at least 0.64. Movement is generally limited to one quintile, with less than 1% of households moving between two or more quintiles where a pair of indices places households in different quintiles.

Table 3.1: Movement of households between quintiles of absolute, PCA and MCA indices

Indices being compared	Correlation coefficient	Percent of households moving between quintiles			Kappa statistic
		Same quintile	One quintile	Two quintiles	
Absolute and PCA	0.9561	71.28	28.11	0.6	0.641
Absolute and MCA	0.9668	74.73	25.03	0.24	0.6841
PCA and MCA	0.9835	83.05	16.93	0.02	0.7881

3.2 Poverty transition: Changes in household SES

Paper I also documents temporal changes in household asset ownership and distributional changes in SES based on the absolute, PCA and MCA indices as means of assessing poverty transition in the Agincourt population. Figures 3.1 - 3.6, visually depict the percentage of households owning particular asset items over time from 2001 to 2013. As can clearly be seen from the figures, the proportion of households that owns assets associated with greater modern wealth has substantially increased over time. As reported in **Paper I**, the proportion of households with dwellings constructed with either brick or cement walls increased from 76% in 2001 to 98% in 2013 and the prevalence of tiles as roof and floor materials of dwellings increased respectively from 3% and 0.5% in 2001 to 15% and 14% in 2013. The use of electricity for lighting and cooking respectively increased from 69% and 4% of households in 2001 to 96% and 50% of households in 2013. Also prominent are increases in the proportions of households owning a stove, fridge, cellphone and car respectively from 41%, 40%, 37 % and 14% in 2001 to 85%, 86%, 98%

and 20% in 2013. In contrast, proportions of households that own asset items associated with traditional wealth such as animal drawn cart and livestock with the exception of chickens have remained persistently low. For example, the proportion of households owning animal drawn carts changed negligibly from 3% in 2001 to 1% in 2013. Similarly, the proportion of households owning cows or pigs only marginally changed from 15% in 2001 to 12% in 2013 for cows and from 4% in 2001 to 2 % in 2013 for pigs. In addition, ownership of goats decreased slightly from 13% in 2001 to 8% in 2013. A comparison of the prevalence of selected household asset indicators in 2013 in the Agincourt population and tribal areas in Mpumalanga province and South Africa show that for most asset indicators, the prevalence in the Agincourt population is comparable to that of aggregated tribal areas in Mpumalanga province (Figure 3.7).

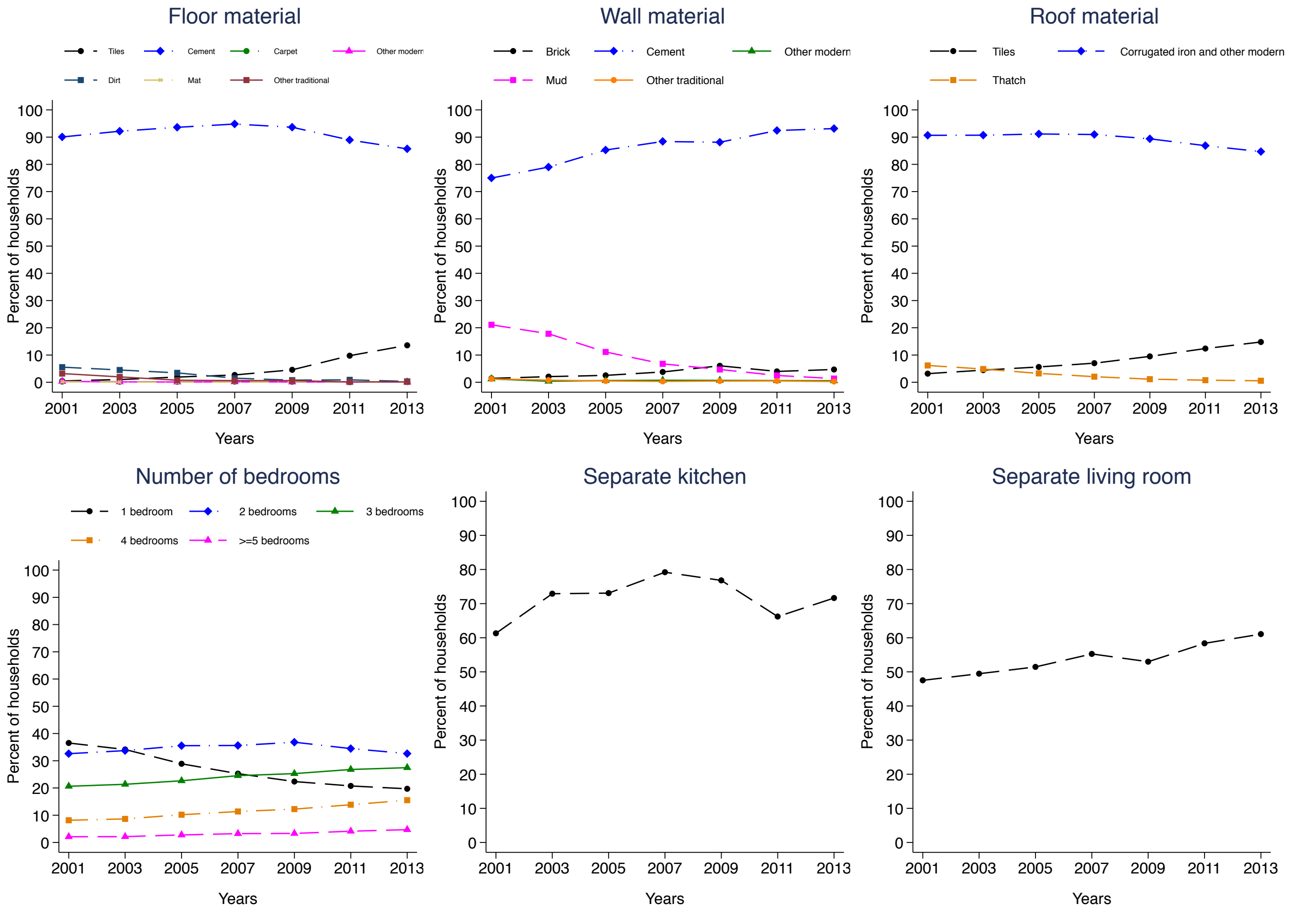


Figure 3.1: Dwelling structure

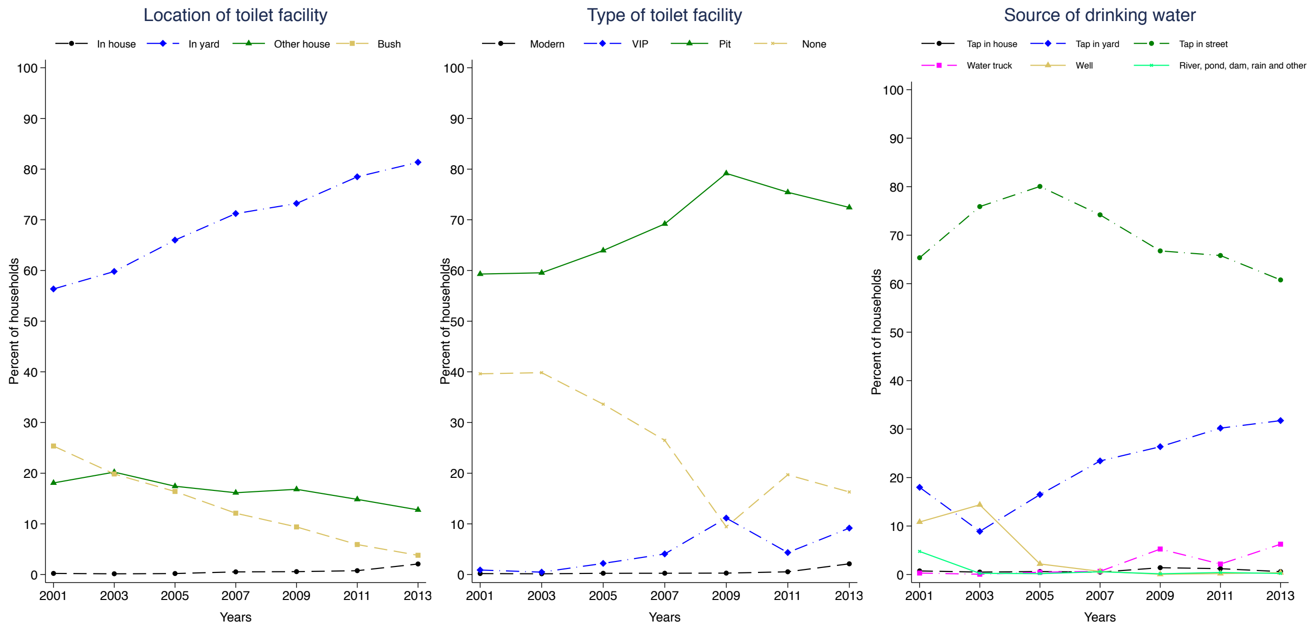


Figure 3.2: Water and Sanitation

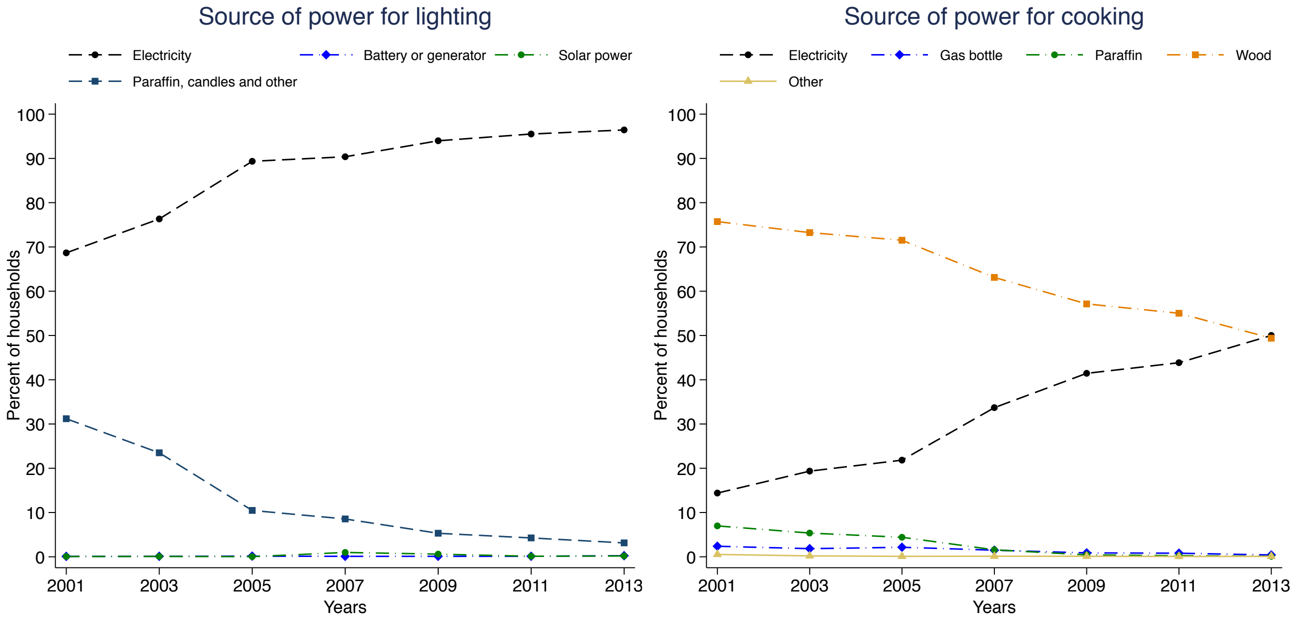


Figure 3.3: Power supply

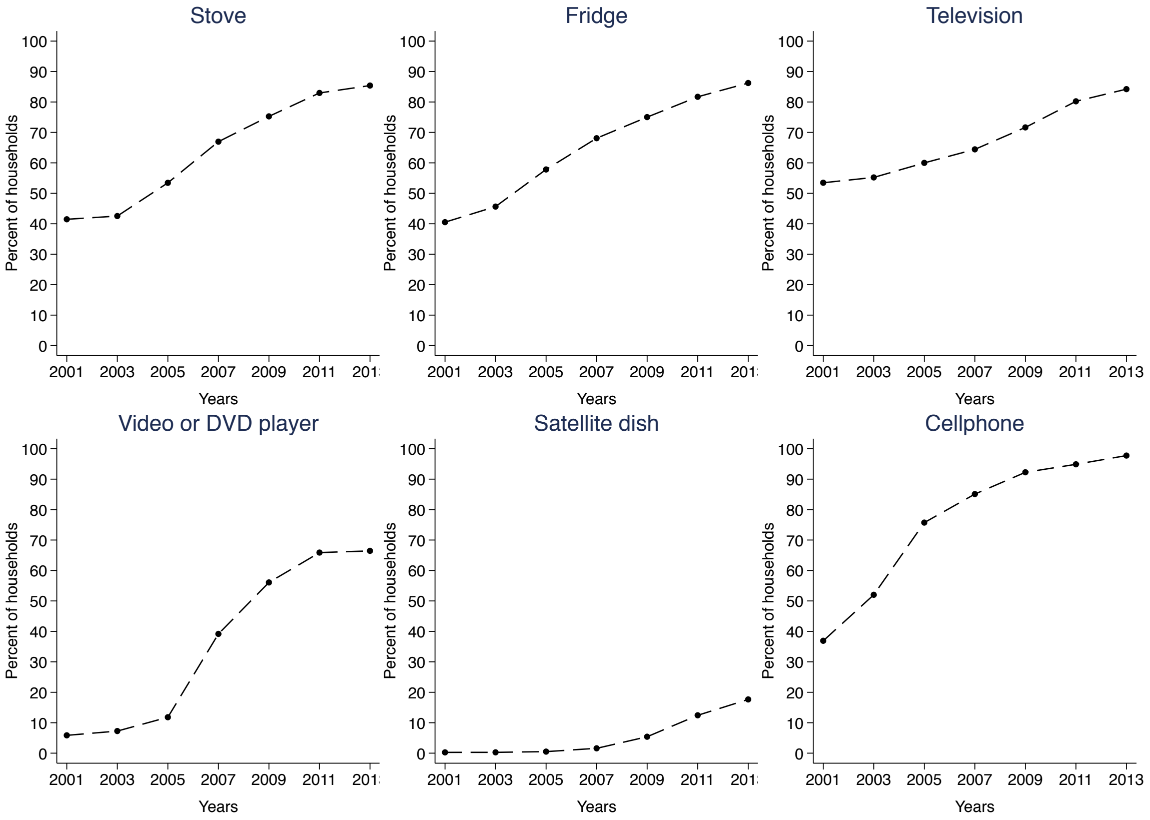


Figure 3.4: Appliances

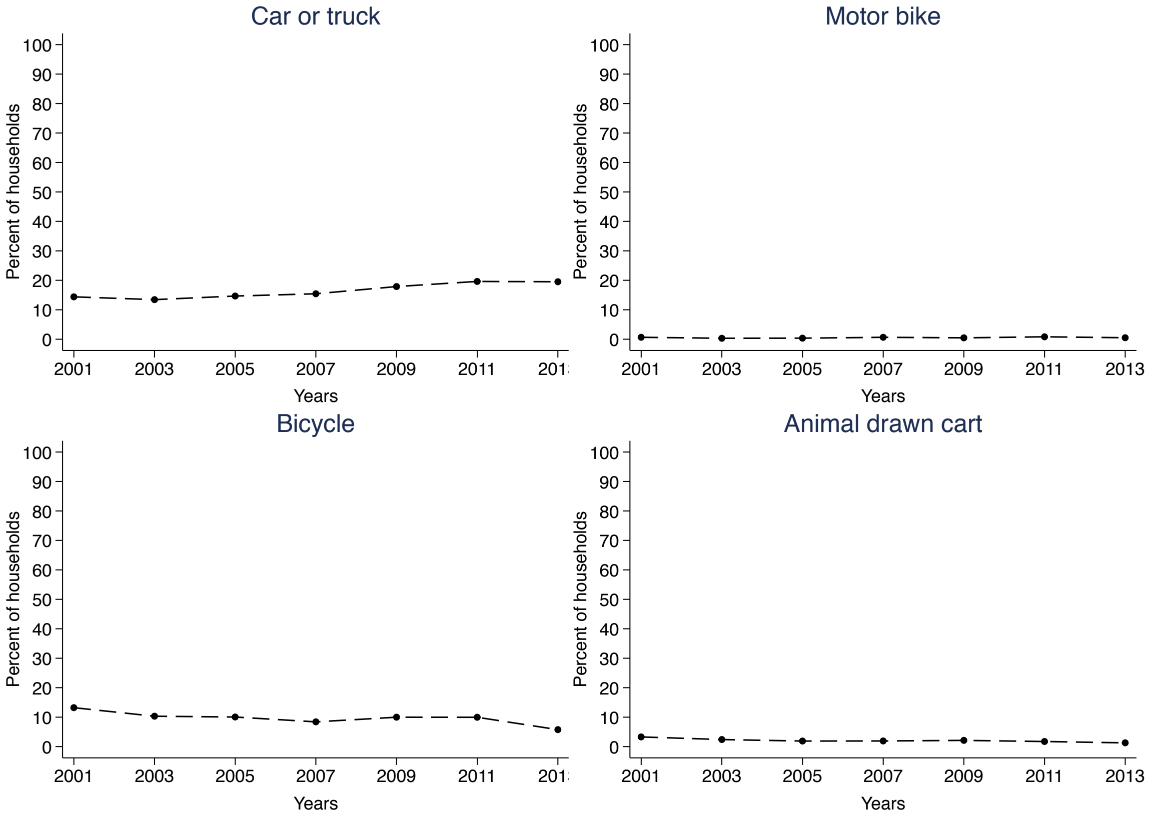


Figure 3.5: Transport

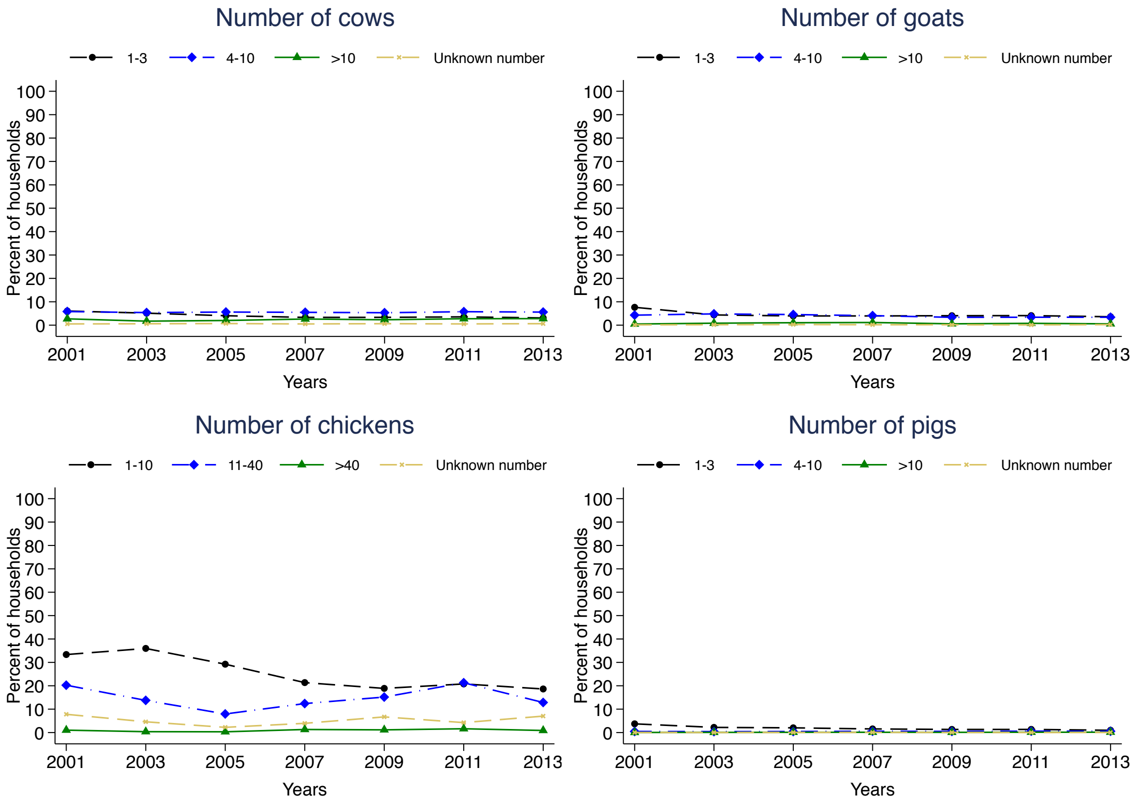


Figure 3.6: Livestock

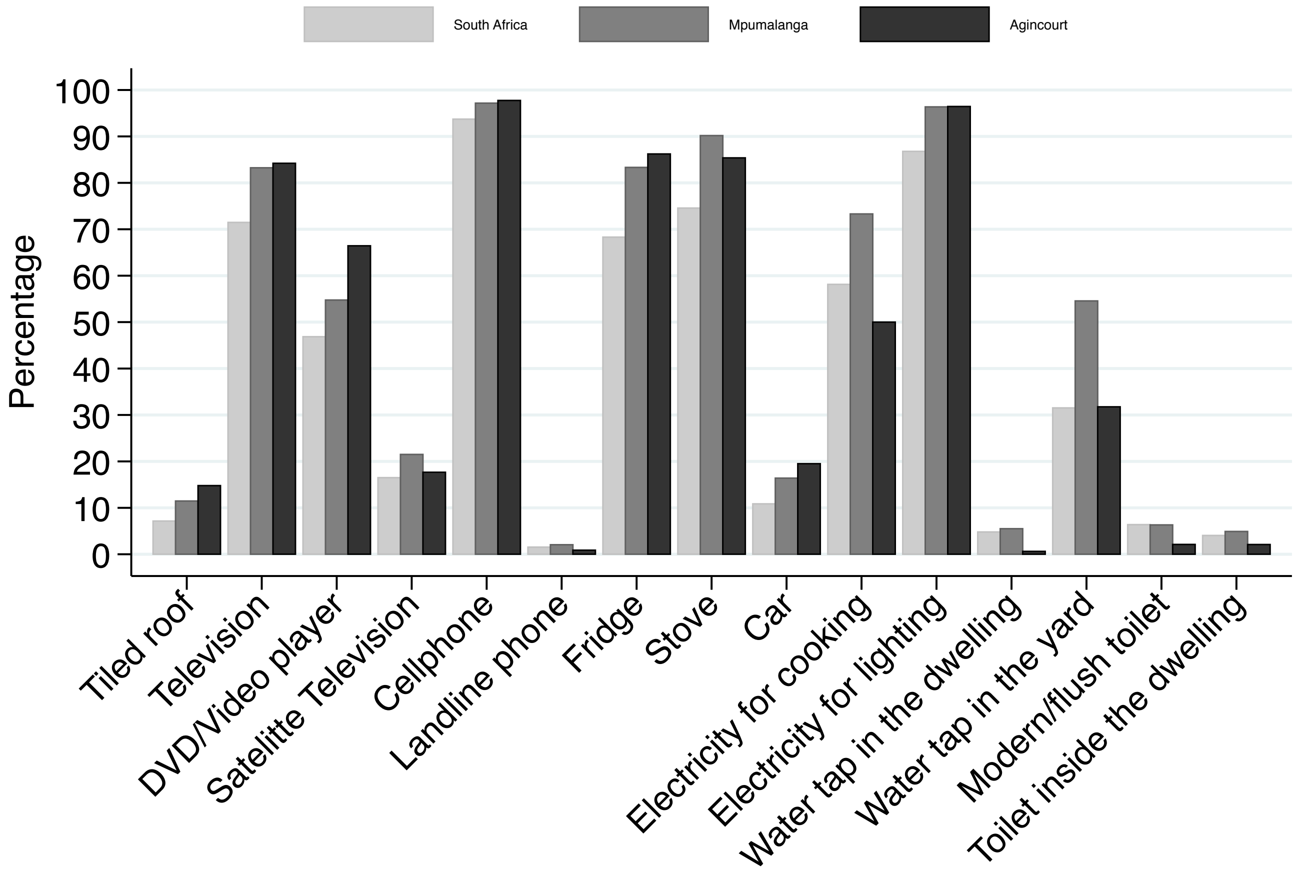


Figure 3.7: Prevalence of selected household asset indicators in 2013 in the Agincourt population and tribal areas in Mpumalanga province and South Africa
Note: Percentages for Mpumalanga province and South Africa have been computed from the South Africa - General Household Survey 2013 dataset (Statistics South Africa, 2014b).

In line with increases over time in the proportion of households that owns assets associated with greater modern wealth, distributional changes in SES reported in **Paper I** indicate that there has been positive location and shape shifts in the SES distribution between 2001 and 2013. Across all the three indices, from one time period to the next the mean and median values have persistently shifted to the right and the level of variability in the values of the indices, as depicted by the standard deviation values, has progressively declined. Complementary results presented in Table 3.2 in the form of a transition matrix based on the absolute index also reveal substantial and increasing upward mobility in SES between 2001 and 2013. This is evidenced with most cells on the diagonal (in bold and depicting households that remained in the same SES quintile between 2001 and 2013) containing less than 25% of the row percentage and the cells to the left of the diagonal (which contain households that moved into lower SES quintiles between 2001 and 2013) containing low values in contrast to the higher values contained in the cells to the right of the diagonal (which contain households that moved into higher SES quintiles between

2001 and 2013).

Table 3.2: Household SES quintile transition matrix between 2001 and 2013: Agincourt

			2013					
			Quintiles of SES					Number of household
			1	2	3	4	5	
2001	Quintiles of SES	1	**7.84**	18.9	25.71	26.94	20.61	2,450
		2	1.3	**9.55**	19.72	31.19	38.25	1,770
		3	0.24	4.47	**11.89**	30.65	52.75	1,253
		4	0.22	1.9	5.47	**23.1**	69.31	896
		5	0	0.75	1.79	15.05	**82.41**	671
Number of household			220	710	1,189	1,904	3,017	7,040

Chapter 4

Findings on mortality transition in Agincourt, rural South Africa, 1993-2013.

4.1 Mortality transition

This section presents findings on trends and levels in various sex, age and cause components of mortality in the Agincourt HDSS surveillance population over the period 1993-2013.

4.1.1 Overall mortality

As presented in Table 1 of **Paper II** and Table 4.1 in this thesis, a total of 13 472 (6445 in females and 7 027 in males) deaths were recorded in 1 604 085 (833 468 in females and 770 617 in males) person-years of follow-up in the Agincourt HDSS surveillance population over the period 1993-2013. The numbers in the tables show that both males and females experienced steady increases in overall mortality from the 1993-1997 time period until the 2004-2007 time period. From base averages of 5.6 and 4.4 deaths per 1000 person-years for males and females respectively during the 1993-1997 time period, overall mortality steadily increased and reached averages of 12.8 and 10.6 deaths per 1000 person years for males and females respectively during the 2004-2007 time period before starting to decline steadily. From the peak levels during the 2004-2007 time period, overall mortality declined to averages of 8.2 and 7.5 deaths per 1000 person years for males and females respectively during the 2011-2013 time period. Year by year changes in the levels of overall mortality over the period 1993-2013 are presented in Figure 1 and Table 2 of **Paper II**.

Table 4.1: Person years and number of deaths by time period and cause of death categories, Agincourt, South Africa, 1993-2013

	Indicator	1993-1997	1998-2000	2001-2003	2004-2007	2008-2010	2011-2013	1993-2013
Females	Person years	174518	108599	110608	155062	138883	145799	833468
	Number of deaths	773	677	1019	1651	1229	1096	6445
	HIV/AIDS & TB	200	229	507	828	483	300	2547
	Other Communicable	138	103	124	210	212	233	1020
	Non Communicable	245	203	233	386	417	448	1932
	Injuries	48	30	38	46	31	35	228
	Indeterminate	55	49	49	104	54	45	356
	VA not done	87	63	68	77	32	35	362
Males	Person years	161119	101311	102972	143188	127695	134331	770617
	Number of deaths	900	708	1115	1833	1363	1108	7027
	HIV/AIDS & TB	241	239	472	748	511	287	2498
	Other Communicable	120	96	122	230	258	205	1031
	Non Communicable	232	175	237	426	357	391	1818
	Injuries	131	79	119	161	105	132	727
	Indeterminate	47	21	45	80	56	42	291
	VA not done	129	98	120	188	76	51	662

4.1.2 Mortality risk by age

One way of presenting mortality risk by age is with survival curves that show the proportions of a synthetic birth cohort that survive to later ages. Figure 4.1 presents survival curves that show the decline by age in the survivors in a synthetic birth cohort in the Agincourt HDSS population by sex and time period. The curves are constructed from values of survival probabilities obtained from a set of abridged life tables constructed for each time period separately for males and females. Table 4.2 shows the values represented on the curves for selected ages. The proportion surviving to age 5 declined progressively between the 1993-1997 and 2001-2003 time periods. The trend reversed thereafter and by the 2011-2013 time period the proportion surviving to age 5 had reached almost the level of the 1993-1997 time period. The proportions surviving to ages 50 and 65 declined progressively from the 1993-1997 time period to the 2004-2007 time period. The trend reversed thereafter but the levels during the 2011-2013 time period were lower than the baseline values during the 1993-1997 time period. Although the trends in males mirror the trends in females, consistently males had lower proportions surviving to later ages compared to females. The curves based on the 2011-2013 time period survival probabilities indicate that 74% of newborn females in the Agincourt HDSS population can expect to reach age 50, the corresponding percentage for males is 71%. In addition, 62% of newborn females can expect to survive to age 65 while only 48% of the newborn males can expect to reach that age.

Figure 4.2 compares the distribution by age of deaths recorded in the Agincourt surveillance population between 1997 and 2013 and the deaths that occurred in South Africa between 1997 and 2013 and were registered by the Department of Home Affairs. The data for deaths that occurred in South Africa were

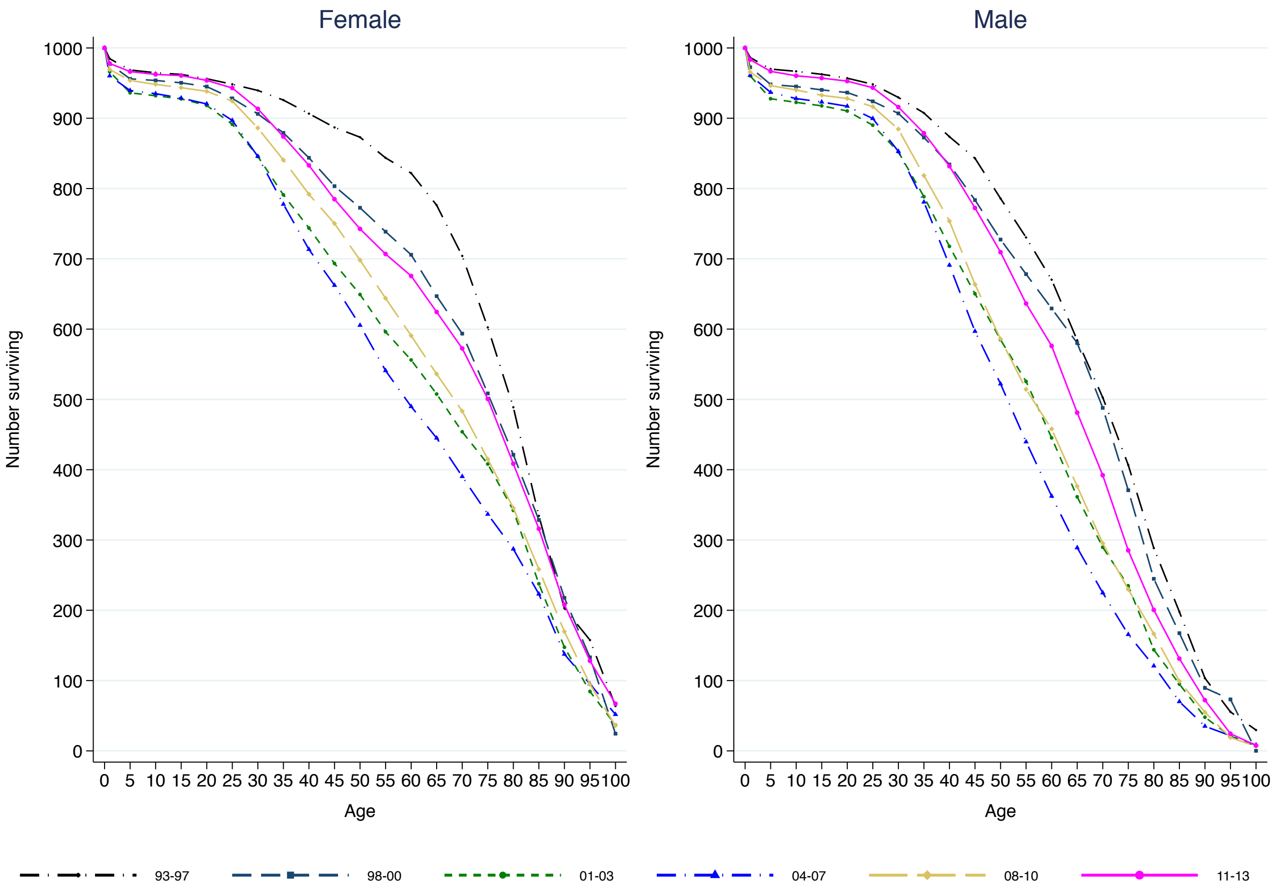


Figure 4.1: Survivors of the synthetic birth cohort of 1000 by age, sex, and time period: Agincourt 1993-2013

Table 4.2: Survivors of the synthetic birth cohort of 1000 to ages 5, 50 and 65

	Age	1993-1997	1998-2000	2001-2003	2004-2007	2008-2010	2011-2013
Females	5	969	956	936	939	954	966
	50	873	773	649	605	698	742
	65	776	647	508	445	536	624
Males	5	970	948	928	937	946	967
	50	786	727	585	522	586	709
	65	584	580	361	289	376	481

compiled by Statistics South Africa and archived by DataFirst. Although the percentages of deaths in each age group are not exactly equal, the general pattern of distribution by age of deaths recorded in the Agincourt surveillance population between 1997 and 2013 is similar to that of deaths that occurred in South Africa between 1997 and 2013. Both data sources reveal that between 1997 and 2013 most of the deaths have been occurring in the 15-49 age group. From both data sources, the percentage of deaths in the 15-49 age group increased steadily between 1997 and 2005. After 2005 the percentage of deaths in the 15-49 age group have been declining progressively. However, the percentages are still higher than in the other age groups. Also notable in Figure 4.2 are increases in the percentage of deaths in the 65+ age group following declines in the percentages of deaths in the 15-49 age group in both data sources.

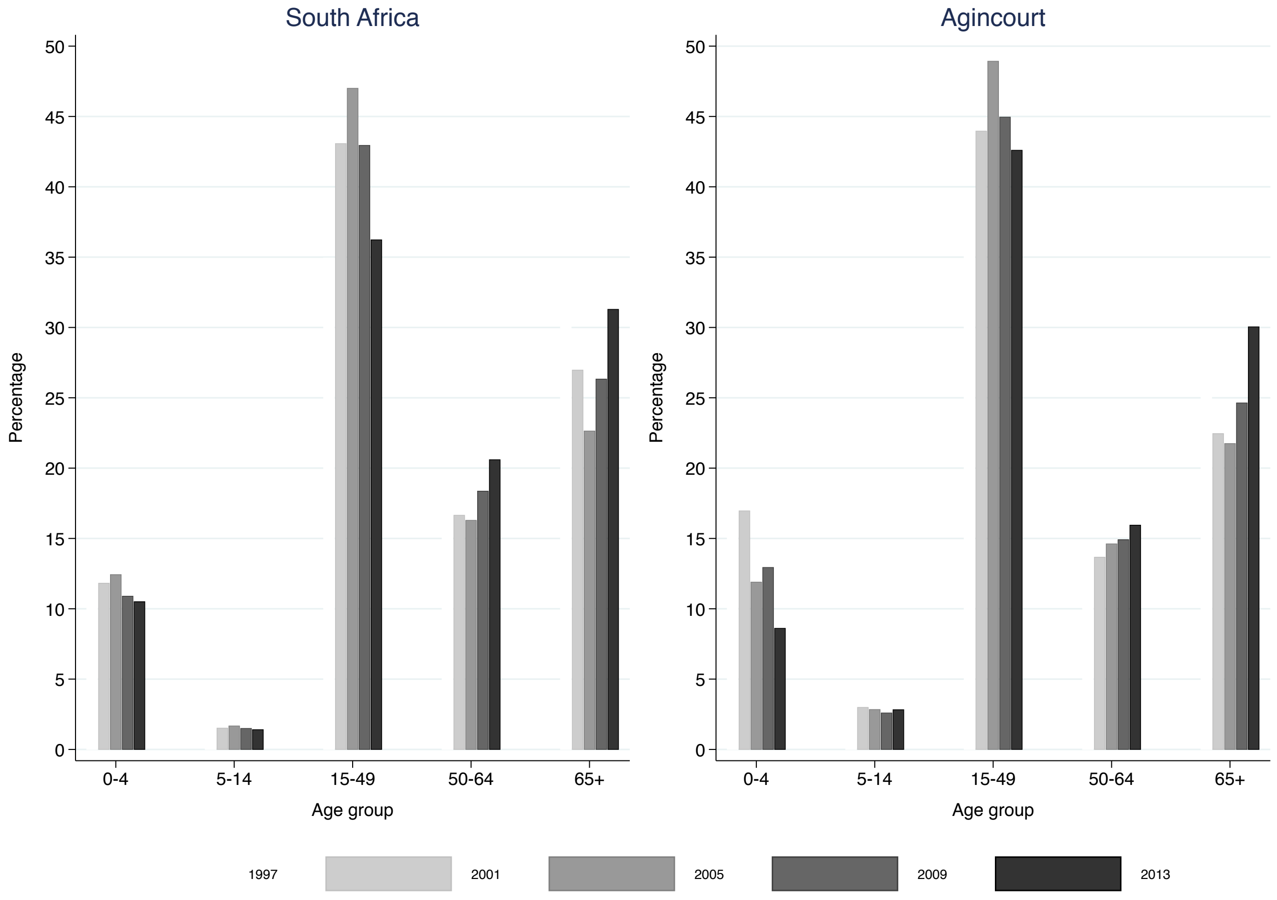


Figure 4.2: Percentage distribution of deaths by age and time period in Agincourt compared with South Africa

Note: Percentages for South Africa are based on mortality and cause of death data for deaths that occurred in South Africa between 1997 and 2013 and were registered by the Department of Home Affairs. The data were compiled by Statistics South Africa and archived by DataFirst (https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/central)

4.1.3 Life expectancy trends

Life expectancies at birth and at selected ages from 1993-2013 are presented in Table 4.3 and visually depicted in Figure 4.3. Figure 1 and Table 2 of **Paper II** also present year by year changes in the values of life expectancies at birth. Trends in life expectancy at birth follow a similar pattern to trends in overall mortality for both males and females. Life expectancy at birth for females dropped from about 74 years during the 1993-1997 time period to about 59 years in the 2004-2007 time period (a loss of 15 years) and returned to around 69 years by the 2011-2013 time period. For males, life expectancy at birth dropped from about 67 years during the 1993-1997 time period to about 51 years during the 2004-2007 time period (a loss of 16 years) and increased to around 61 years by the 2011-2013 time period. Throughout all time periods the years of life expected at all ages for males is persistently lower than that for females. As shown in Table 4.8 life expectancy at birth for males in each time period has been at least 6 years lower than in females.

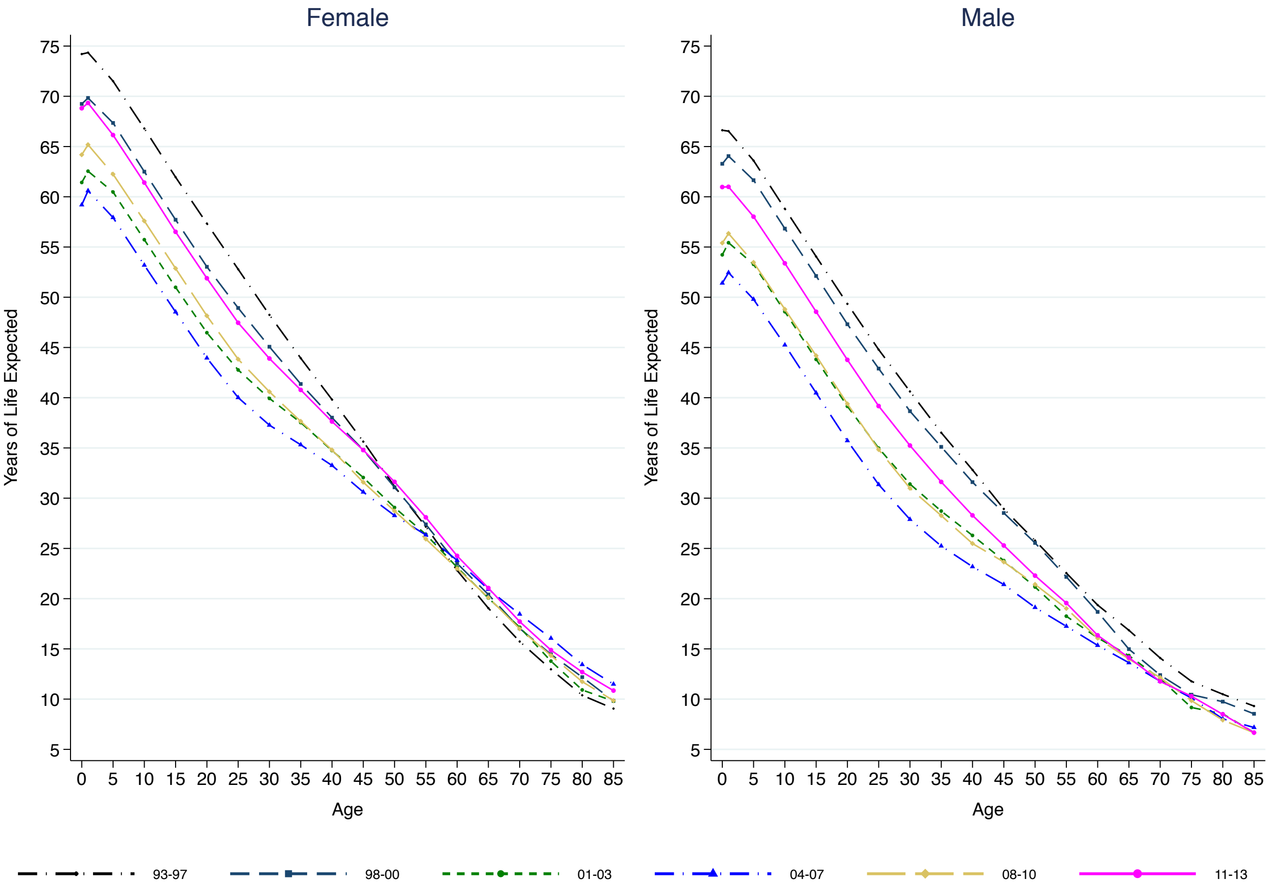


Figure 4.3: Years of life expected at selected ages, Agincourt 1993-2013

Figure 4.4 compares estimates of life expectancy at birth in the Agincourt surveillance population with estimates of life expectancy at birth from other selected rural demographic surveillance sites in Southern

Table 4.3: Years of life expected at selected ages

	Age	1993-1997	1998-2000	2001-2003	2004-2007	2008-2010	2011-2013
Females	0	74.20	69.24	61.43	59.19	64.18	68.81
	5	71.51	67.35	60.47	57.92	62.24	66.14
	15	61.95	57.71	50.98	48.53	52.87	56.51
	50	31.16	31.08	29.07	28.27	28.72	31.62
	65	19.05	20.41	20.06	20.93	20.11	21.05
Males	0	66.62	63.28	54.22	51.41	55.40	60.97
	5	63.60	61.65	53.28	49.79	53.46	58.02
	15	54.05	52.11	43.81	40.47	44.18	48.55
	50	25.78	25.54	21.16	19.12	21.42	22.29
	65	16.85	14.97	14.33	13.62	13.98	14.13

Table 4.4: Female-male difference in life expectancy at selected ages

Age	1993-1997	1998-2000	2001-2003	2004-2007	2008-2010	2011-2013
0	7.6	6.0	7.2	7.8	8.8	7.8
5	7.9	5.7	7.2	8.1	8.8	8.1
15	7.9	5.6	7.2	8.1	8.7	8.0
50	5.4	5.5	7.9	9.1	7.3	9.3
65	2.2	5.4	5.7	7.3	6.1	6.9

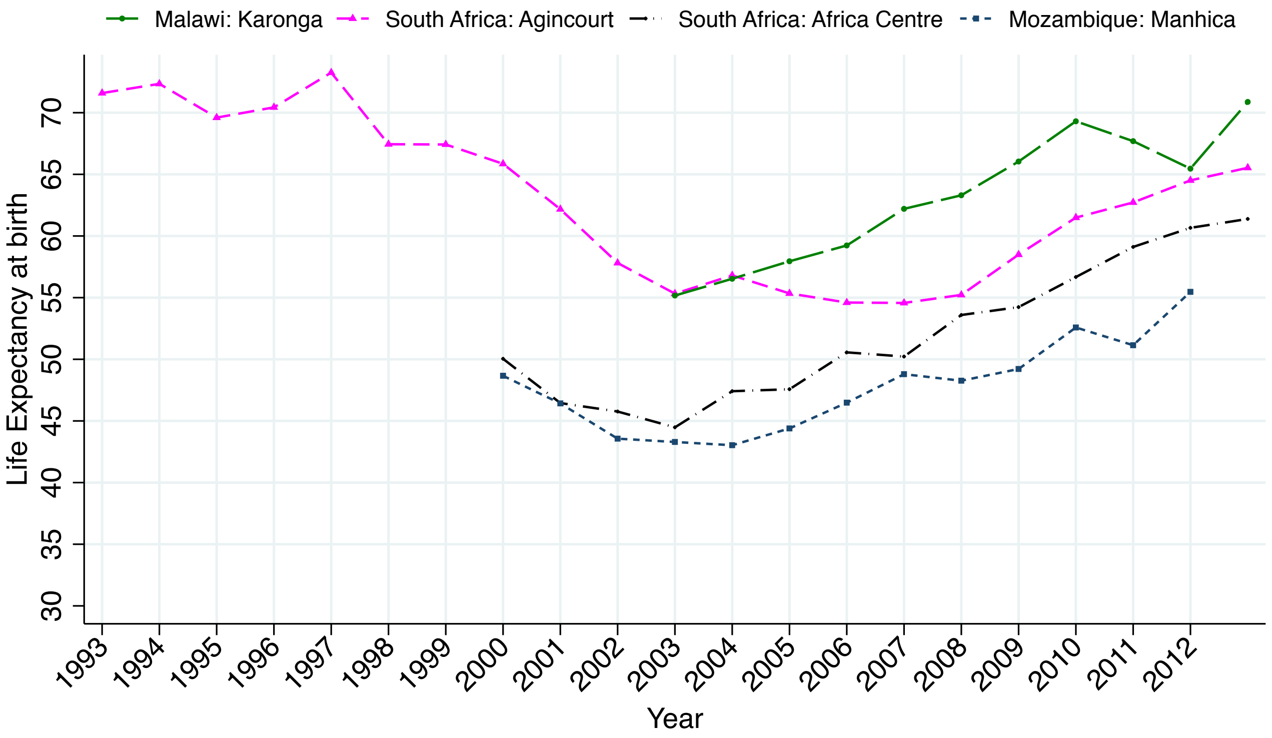


Figure 4.4: Trends in life expectancy at birth in Agincourt and selected demographic surveillance sites in Southern Africa

Africa downloaded from the INDEPTH Population Statistics (INDEPTHStats) website (http://www.indepth-ishare.org/indepthstats/). The data from the other sites is only available from the early 2000s. As Figure 4.4 shows, life expectancy in Agincourt has consistently been higher compared to the Africa Centre site in rural KwaZulu-Natal province, South Africa as well as the Manhiça site in rural southern Mozambique, but consistently lower than the Karonga site in rural northern Malawi. In addition, while life expectancy had started to rise by 2004 in the other rural sites, in Agincourt increases in life expectancy only started to emerge from 2008.

4.1.4 Distribution of deaths by cause categories

Paper II reports estimates of the predicted summed annual probability of dying per 1 000 person-years by year and broad cause of death categories that include HIV/AIDS and TB, other communicable diseases, NCDs, and injuries. Overall, the probability of dying from HIV/AIDS and TB has persistently been the highest throughout the period 1993-2013. The estimates show that the annual probability of dying from HIV/AIDS and TB followed an upward trend from 1993 to around 2004-2005 and has been declining since 2007 but the level in 2013 is still higher than the level in 1993. The prime contribution of HIV/AIDS and TB to the mortality burden in the Agincourt population is also shown in Table 4.5 which lists the leading 5 causes of death for all ages and sexes combined by time period. As shown in Table 4.5, the proportion of HIV/AIDS and TB deaths among all deaths increased from 26% during the 1993-1997 time period to 46% during the 2001-2003 time period and remained relatively stable at 45% in the 2004-2007 time period before taking a progressive downward trend afterwards to 27% in the 2011-2013 time period.

The estimates in **Paper II** show that from the 2004-2007 time period the probability of dying from HIV/AIDS and TB in children under the age of 5 years has been following a downward trend and the level in the 2011-2013 time period is almost the same as the level in the 1993-1997 time period. As shown in Table 4.7, apart from HIV/AIDS and TB, acute respiratory infection and diarrhoeal diseases have persistently featured in the top three causes of death among children under the age of five in the Agincourt population for over two decades. As can be seen from Table 4.7, following the impressive progressive reduction in the importance of HIV/AIDS and TB as a cause of death in children under the age of 5 years, acute respiratory infection has become the leading cause of death in this age group. Deaths from acute respiratory infection constituted 29% of all deaths to children under the age of 5 years during the 2008-2010 time period and 19% during the 2011-2013 time period.

Tables 4.7-4.10 present complementary results to those presented in **Paper II** on the evolution of the distribution of deaths by cause categories, sex and time period for ages 5-14, 15-49, 50-64 and 65+. As shown in these tables, throughout all time periods HIV/AIDS and TB feature as one of the top three single causes of death in all age groups for both males and females. Of course, the prominence of

HIV/AIDS and TB as a cause of death in all time periods is more pronounced in the 15-49 age group followed by the 50-64 age group. In the 15-49 age group, HIV/AIDS and TB mortality accounted for 50% of the deaths in males and 70% of the deaths in females in the peak period of HIV/AIDS and TB mortality in 2004-2007. Although HIV/AIDS and TB mortality has steadily been declining in recent years, 32% of the deaths in males and 41% of the deaths in females in the 2011-2013 time period were HIV/AIDS and TB related. The results also provide evidence that NCDs such as neoplasms, stroke, cardiac failure and diabetes have been significant causes of death in the adult age groups along with HIV/AIDS and TB.

More complementary results on the contribution of HIV/AIDS and TB to mortality at various ages are displayed in Figure 4.5 which compares mortality rates from all causes to those from HIV/AIDS and TB as well as causes other than HIV/AIDS and TB. The figure shows that during the peak period of HIV/AIDS and TB mortality in 2004-2007 in the adult ages from 20 to 54 years mortality rates of HIV/AIDS and TB were higher than those for other causes of death.

Table 4.5: Top five causes of death among all individuals by time period, Agincourt 1993-2013

Rank	1993-1997		1998-2000		2001-2003		2004-2007		2008-2010		2011-2013	
	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)
	All Causes	1673(100%)	All Causes	1385(100%)	All Causes	2134(100%)	All Causes	3484(100%)	All Causes	2592(100%)	All Causes	2204(100%)
1	HIV/TB	441(26%)	HIV/TB	468(34%)	HIV/TB	979(46%)	HIV/TB	1576(45%)	HIV/TB	994(38%)	HIV/TB	587(27%)
2	Digestive neoplasms	112(7%)	Acute respiratory infection	94(7%)	Acute respiratory infection	109(5%)	Acute respiratory infection	242(7%)	Acute respiratory infection	274(11%)	Acute respiratory infection	264(12%)
3	Acute respiratory infection	108(6%)	Digestive neoplasms	76(5%)	Digestive neoplasms	79(4%)	Digestive neoplasms	113(3%)	Stroke	127(5%)	Asthma	138(6%)
4	Assault	72(4%)	Stroke	44(3%)	Assault	59(3%)	Stroke	98(3%)	Digestive neoplasms	96(4%)	Stroke	99(4%)
5	Diarrhoeal diseases	70(4%)	Assault	42(3%)	Diarrhoeal diseases	55(3%)	Diabetes mellitus	96(3%)	Other cardiac	80(3%)	Digestive neoplasms	97(4%)

Table 4.6: Top five causes of death among children 0-4 years old by time period, Agincourt 1993-2013

Rank	1993-1997		1998-2000		2001-2003		2004-2007		2008-2010		2011-2013	
	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)
	All Causes	300(100%)	All Causes	263(100%)	All Causes	348(100%)	All Causes	461(100%)	All Causes	340(100%)	All Causes	227(100%)
1	HIV/TB	73(24%)	HIV/TB	77(29%)	HIV/TB	130(37%)	HIV/TB	134(29%)	Acute respiratory infection	100(29%)	Acute respiratory infection	44(19%)
2	Diarrhoeal diseases	53(18%)	Acute respiratory infection	48(18%)	Acute respiratory infection	60(17%)	Acute respiratory infection	116(25%)	HIV/TB	70(21%)	HIV/TB	36(16%)
3	Acute respiratory infection	43(14%)	Diarrhoeal diseases	27(10%)	Diarrhoeal diseases	44(13%)	Diarrhoeal diseases	47(10%)	Diarrhoeal diseases	32(9%)	Diarrhoeal diseases	23(10%)
4	Neonatal pneumonia	11(4%)	Birth asphyxia	7(3%)	Birth asphyxia	10(3%)	Neonatal pneumonia	22(5%)	Neonatal pneumonia	27(8%)	Malaria	19(8%)
5	Other neonatal causes	8(3%)	Malaria	6(2%)	Neonatal pneumonia	9(3%)	Malaria	11(2%)	Malaria	13(4%)	Neonatal pneumonia	19(8%)

Table 4.7: Top five causes of death among individuals aged 5-14 years by time period, Agincourt 1993-2013

	Rank	1993-1997		1998-2000		2001-2003		2004-2007		2008-2010		2011-2013	
		Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)
Males		All Causes	36(100%)	All Causes	24(100%)	All Causes	31(100%)	All Causes	52(100%)	All Causes	43(100%)	All Causes	30(100%)
	1	HIV/TB	7(19%)	HIV/TB	4(17%)	HIV/TB	10(32%)	HIV/TB	19(37%)	HIV/TB	14(33%)	Road traffic accident	4(13%)
	2	Accidental drowning	5(14%)	Malaria	3(12%)	Road traffic accident	6(19%)	Road traffic accident	6(12%)	Acute respiratory infection	13(30%)	Acute respiratory infection	4(13%)
	3	Acute respiratory infection	5(14%)	Road traffic accident	2(8%)	Accidental fire	2(6%)	Acute respiratory infection	4(8%)	Road traffic accident	4(9%)	HIV/TB	4(13%)
	4	Road traffic accident	3(8%)	Accidrmtal drowning	2(8%)	Intentional self-harm	2(6%)	Asthma	4(8%)	Meningitis and encephalitis	2(5%)	Other transport accident	3(10%)
	5	Malaria	2(6%)	Intentional self-harm	2(8%)	Acute respiratory infection	2(6%)	Malaria	2(4%)	Other infectious diseases	2(5%)	Acute abdomen	2(7%)
Females		All Causes	32(100%)	All Causes	17(100%)	All Causes	27(100%)	All Causes	42(100%)	All Causes	32(100%)	All Causes	18(100%)
	1	HIV/TB	6(19%)	Accidental fire	3(18%)	HIV/TB	10(37%)	HIV/TB	9(21%)	HIV/TB	18(56%)	HIV/TB	7(39%)
	2	Road traffic accident	4(12%)	Malaria	3(18%)	Road traffic accident	2(7%)	Malaria	5(12%)	Acute respiratory infection	4(12%)	Acute respiratory infection	4(22%)
	3	Malaria	4(12%)	HIV/TB	2(12%)	Acute respiratory infection	2(7%)	Acute respiratory infection	4(10%)	Asthma	2(6%)	Malaria	2(11%)
	4	Acute respiratory infection	3(9%)	Asthma	2(12%)	Other infectious diseases	2(7%)	Accidrmtal drowning	2(5%)	Malaria	1(3%)	Other infectious diseases	1(6%)
	5	Epilepsy	3(9%)	Other infectious diseases	2(12%)	Asthma	1(4%)	Diarrhoeal diseases	2(5%)	Liver cirrhosis	1(3%)	Road traffic accident	1(6%)

Table 4.8: Top five causes of death among individuals aged 15-49 years by time period, Agincourt 1993-2013

	Rank	1993-1997		1998-2000		2001-2003		2004-2007		2008-2010		2011-2013	
		Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)
Males		All Causes	308(100%)	All Causes	263(100%)	All Causes	505(100%)	All Causes	893(100%)	All Causes	652(100%)	All Causes	475(100%)
	1	HIV/TB	85(28%)	HIV/TB	110(42%)	HIV/TB	258(51%)	HIV/TB	444(50%)	HIV/TB	306(47%)	HIV/TB	154(32%)
	2	Assault	45(15%)	Road traffic accident	25(10%)	Assault	41(8%)	Assault	53(6%)	Acute respiratory infection	47(7%)	Acute respiratory infection	54(11%)
	3	Road traffic accident	29(9%)	Assault	25(10%)	Road traffic accident	17(3%)	Road traffic accident	51(6%)	Assault	40(6%)	Road traffic accident	44(9%)
	4	Digestive neoplasms	15(5%)	Digestive neoplasms	12(5%)	Intentional self-harm	17(3%)	Acute respiratory infection	38(4%)	Road traffic accident	38(6%)	Digestive neoplasms	25(5%)
	5	Acute respiratory infection	10(3%)	Other neoplasms	8(3%)	Digestive neoplasms	16(3%)	Digestive neoplasms	22(2%)	Digestive neoplasms	19(3%)	Assault	20(4%)
Females		All Causes	194(100%)	All Causes	261(100%)	All Causes	485(100%)	All Causes	811(100%)	All Causes	517(100%)		480(100%)
	1	HIV/TB	87(45%)	HIV/TB	143(55%)	HIV/TB	338(70%)	HIV/TB	568(70%)	HIV/TB	312(60%)	HIV/TB	198(41%)
	2	Assault	10(5%)	Acute respiratory infection	13(5%)	Digestive neoplasms	14(3%)	Acute respiratory infection	31(4%)	Acute respiratory infection	34(7%)	Acute respiratory infection	53(11%)
	3	Digestive neoplasms	9(5%)	Respiratory neoplasms	8(3%)	Reproductive neoplasms	10(2%)	Digestive neoplasms	15(2%)	Asthma	18(3%)	Asthma	28(6%)
	4	Intentional self-harm	9(5%)	Digestive neoplasms	8(3%)	Road traffic accident	9(2%)	Obstetric haemorrhage	12(1%)	Reproductive neoplasms	15(3%)	Digestive neoplasms	19(4%)
	5	Road traffic accident	8(4%)	Breast neoplasms	7(3%)	Acute respiratory infection	7(1%)	Stroke	12(1%)	Respiratory neoplasms	10(2%)	Reproductive neoplasms	14(3%)

Table 4.9: Top five causes of death among individuals aged 50-64 years by time period, Agincourt 1993-2013

	Rank	1993-1997		1998-2000		2001-2003		2004-2007		2008-2010		2011-2013	
		Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)
Males		All Causes	158(100%)	All Causes	84(100%)	All Causes	191(100%)	All Causes	326(100%)	All Causes	214(100%)	All Causes	214(100%)
	1	HIV/TB	52(33%)	HIV/TB	26(31%)	HIV/TB	84(44%)	HIV/TB	144(44%)	HIV/TB	83(39%)	HIV/TB	62(29%)
	2	Digestive neoplasms	14(9%)	Digestive neoplasms	8(10%)	Digestive neoplasms	14(7%)	Digestive neoplasms	23(7%)	Acute respiratory infection	27(13%)	Acute respiratory infection	25(12%)
	3	Road traffic accident	9(6%)	Assault	7(8%)	Stroke	9(5%)	Acute respiratory infection	15(5%)	Digestive neoplasms	11(5%)	Digestive neoplasms	19(9%)
	4	Other cardiac	8(5%)	Stroke	5(6%)	Acute respiratory infection	9(5%)	Stroke	14(4%)	Diabetes mellitus	9(4%)	Asthma	18(8%)
	5	Assault	7(4%)	Respiratory neoplasms	4(5%)	Assault	7(4%)	Respiratory neoplasms	13(4%)	Other cardiac	8(4%)	Acute abdomen	9(4%)
Females		All Causes	94(100%)	All Causes	83(100%)	All Causes	125(100%)	All Causes	231(100%)	All Causes	175(100%)	All Causes	124(100%)
	1	HIV/TB	21(22%)	HIV/TB	21(25%)	HIV/TB	50(40%)	HIV/TB	113(49%)	HIV/TB	72(41%)	HIV/TB	36(29%)
	2	Digestive neoplasms	9(10%)	Digestive neoplasms	10(12%)	Digestive neoplasms	10(8%)	Respiratory neoplasms	11(5%)	Stroke	16(9%)	Acute respiratory infection	17(14%)
	3	Other cardiac	6(6%)	Other cardiac	6(7%)	Diabetes mellitus	9(7%)	Diabetes mellitus	9(4%)	Digestive neoplasms	13(7%)	Asthma	11(9%)
	4	Diabetes mellitus	5(5%)	Stroke	6(7%)	Other cardiac	6(5%)	Digestive neoplasms	9(4%)	Acute respiratory infection	9(5%)	Digestive neoplasms	5(4%)
	5	Respiratory neoplasms	4(4%)	Acute respiratory infection	5(6%)	Other neoplasms	6(5%)	Other neoplasms	8(3%)	Other cardiac	9(5%)	Acute abdomen	5(4%)

Table 4.10: Top five causes of death among individuals aged 65+ years by time period, Agincourt 1993-2013

	Rank	1993-1997		1998-2000		2001-2003		2004-2007		2008-2010		2011-2013	
		Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)	Cause	n (%)
Males		All Causes	251(100%)	All Causes	195(100%)	All Causes	205(100%)	All Causes	327(100%)	All Causes	272(100%)	All Causes	274(100%)
	1	HIV/TB	64(25%)	HIV/TB	57(29%)	HIV/TB	56(27%)	HIV/TB	82(25%)	HIV/TB	69(25%)	HIV/TB	46(17%)
	2	Digestive neoplasms	35(14%)	Digestive neoplasms	24(12%)	Other cardiac	15(7%)	Diabetes mellitus	26(8%)	Digestive neoplasms	26(10%)	Acute respiratory infection	36(13%)
	3	Other cardiac	16(6%)	Acute respiratory infection	10(5%)	Digestive neoplasms	14(7%)	Other cardiac	26(8%)	Stroke	22(8%)	Other cardiac	20(7%)
	4	Respiratory neoplasms	15(6%)	Other neoplasms	10(5%)	Respiratory neoplasms	14(7%)	Digestive neoplasms	26(8%)	Acute respiratory infection	21(8%)	Stroke	19(7%)
	5	Stroke	13(5%)	Respiratory neoplasms	9(5%)	Other neoplasms	12(6%)	Respiratory neoplasms	21(6%)	Respiratory neoplasms	19(7%)	Digestive neoplasms	18(7%)
Females		All Causes	300(100%)	All Causes	195(100%)	All Causes	217(100%)	All Causes	341(100%)	All Causes	347(100%)	All Causes	362(100%)
	1	HIV/TB	46(15%)	HIV/TB	28(14%)	HIV/TB	43(20%)	HIV/TB	63(18%)	Stroke	62(18%)	Stroke	53(15%)
	2	Digestive neoplasms	30(10%)	Stroke	22(11%)	Stroke	17(8%)	Diabetes mellitus	36(11%)	HIV/TB	50(14%)	HIV/TB	44(12%)
	3	Other cardiac	23(8%)	Other cardiac	15(8%)	Diabetes mellitus	16(7%)	Stroke	32(9%)	Other cardiac	39(11%)	Asthma	37(10%)
	4	Stroke	23(8%)	Digestive neoplasms	14(7%)	Chronic obstructive pulmonary dis	12(6%)	Other cardiac	28(8%)	Diabetes mellitus	25(7%)	Other cardiac	34(9%)
	5	Acute respiratory infection	23(8%)	Other neoplasms	13(7%)	Other cardiac	12(6%)	Chronic obstructive pulmonary dis	19(6%)	Asthma	20(6%)	Acute respiratory infection	27(7%)

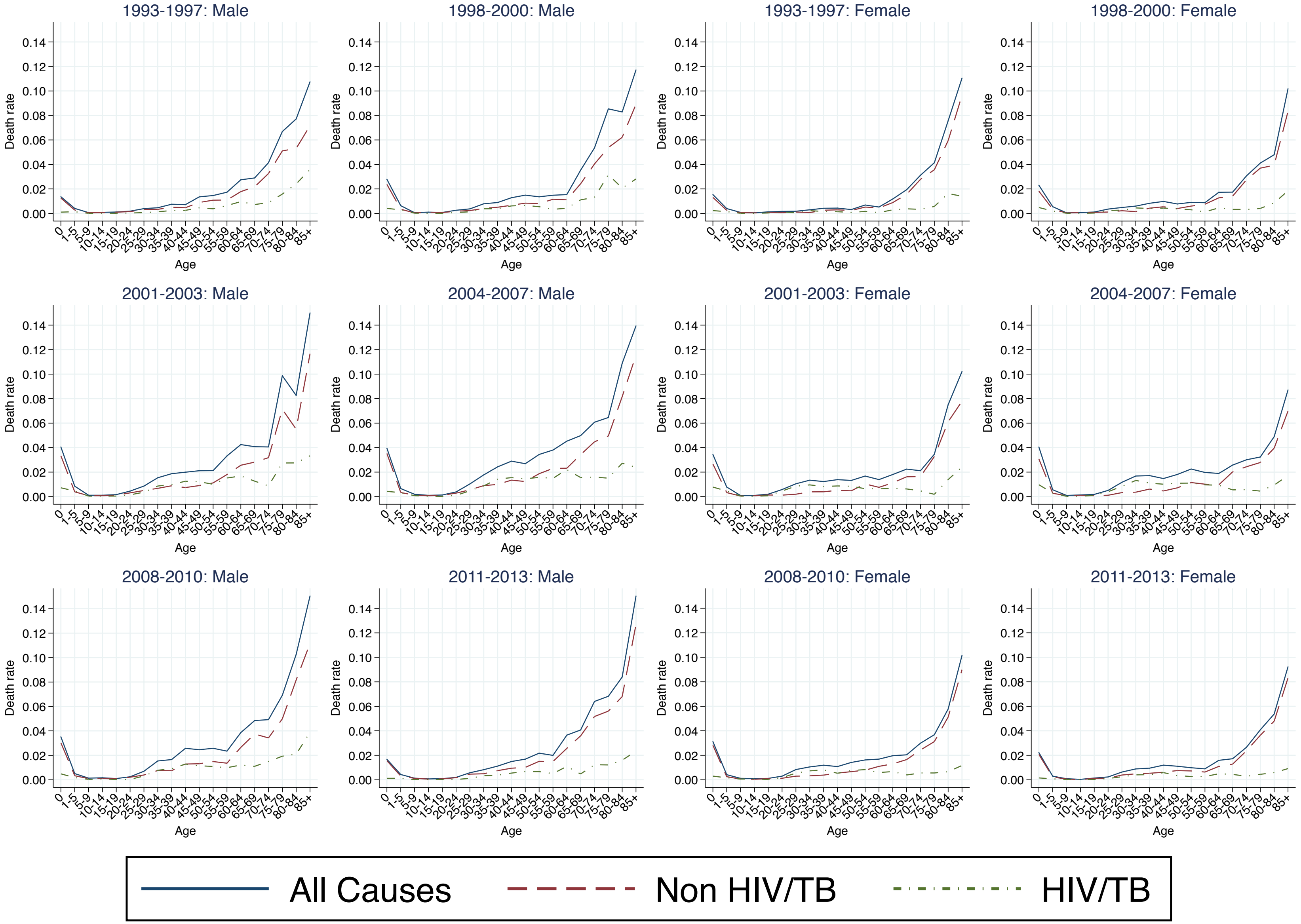


Figure 4.5: Mortality rates with and without HIV/TB by age, sex and time period: Agincourt 1993-2013

4.1.5 HIV-cause-deleted life expectancy

Another important way of assessing the effect of a particular disease on overall mortality is by examining the potential increases in life expectancy at various ages that would result if deaths due to that particular disease were eliminated. Owing to the significant contribution of HIV/AIDS and TB to the mortality profiles of the Agincourt population in the last two decades, another complementary analysis estimated the potential increases in life expectancy at various ages that would result if deaths due to HIV/AIDS and TB were eliminated. The results are presented in Figures 4.6-4.8 and Table 4.11 and they indicate that the elimination of HIV/AIDS and TB would have added 11 and 15.7 years to male and female life expectancies at birth, respectively, during the peak period of HIV/AIDS and TB mortality in 2004-2007. Following the reduction in HIV/AIDS and TB mortality, in the 2011-2013 time period the elimination of HIV/AIDS and TB would have added 5.5 years to male and 6.5 years to female life expectancy at birth. As shown in Figure 4.7, the HIV-cause deleted life expectancy at birth in the period 2011-2013 is lower than that in the period 1993-1997 for both males and females. This indicates the existence of competing mortality risks that are affecting life expectancy alongside HIV/AIDS and TB. The findings of the analysis of the leading causes of death show that the main competing risks are NCDs, especially in older females (Table 4.10), and injuries and accidents particularly in middle aged males (Table 4.8). The findings are also consistent with the finding that in the 15-49 year age group trends in NCD mortality mirrored trends in HIV/AIDS mortality (Kabudula et al., 2014b) .

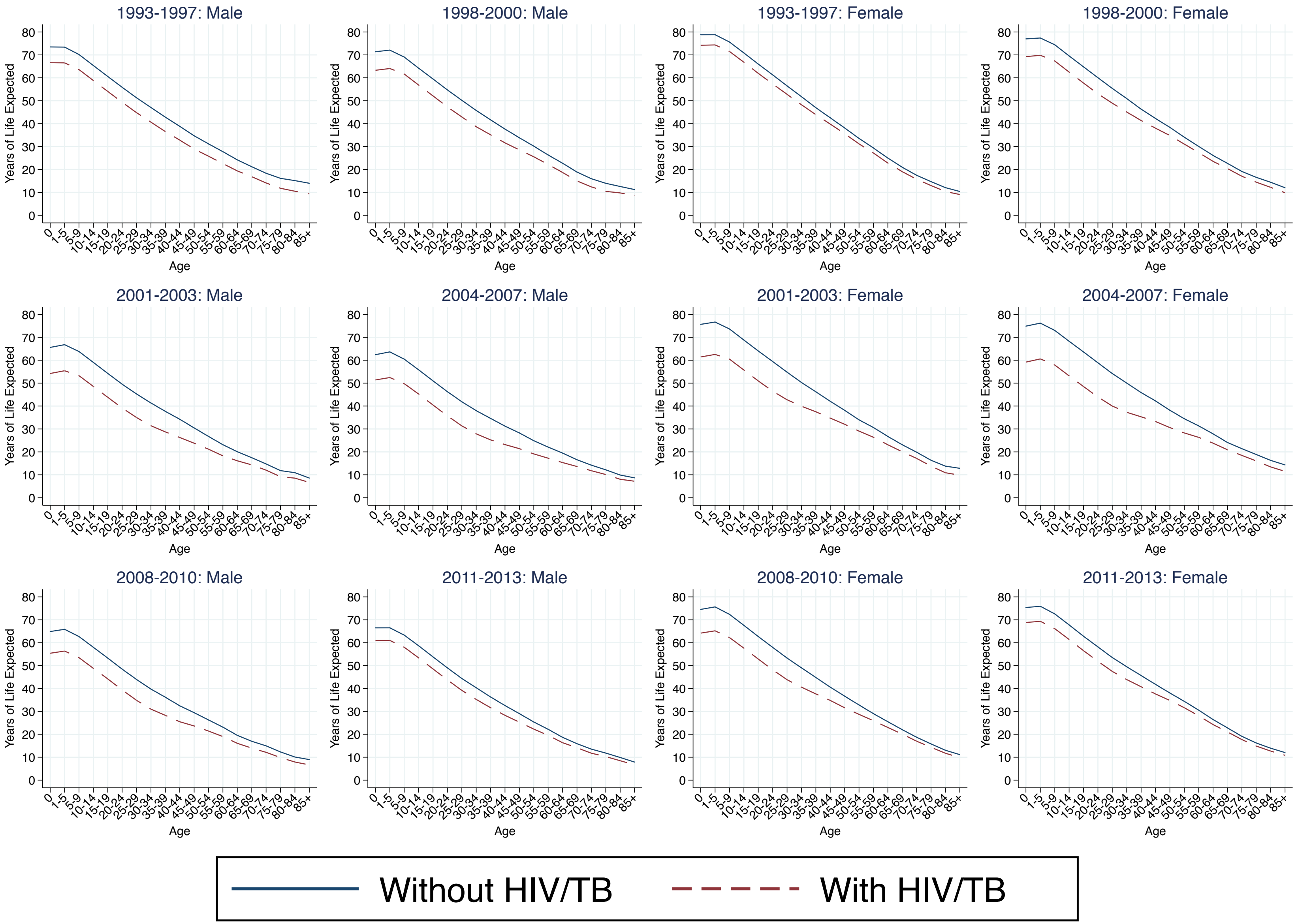


Figure 4.6: Life expectancy at selected ages with and without HIV/TB by sex and time period: Agincourt 1993-2013

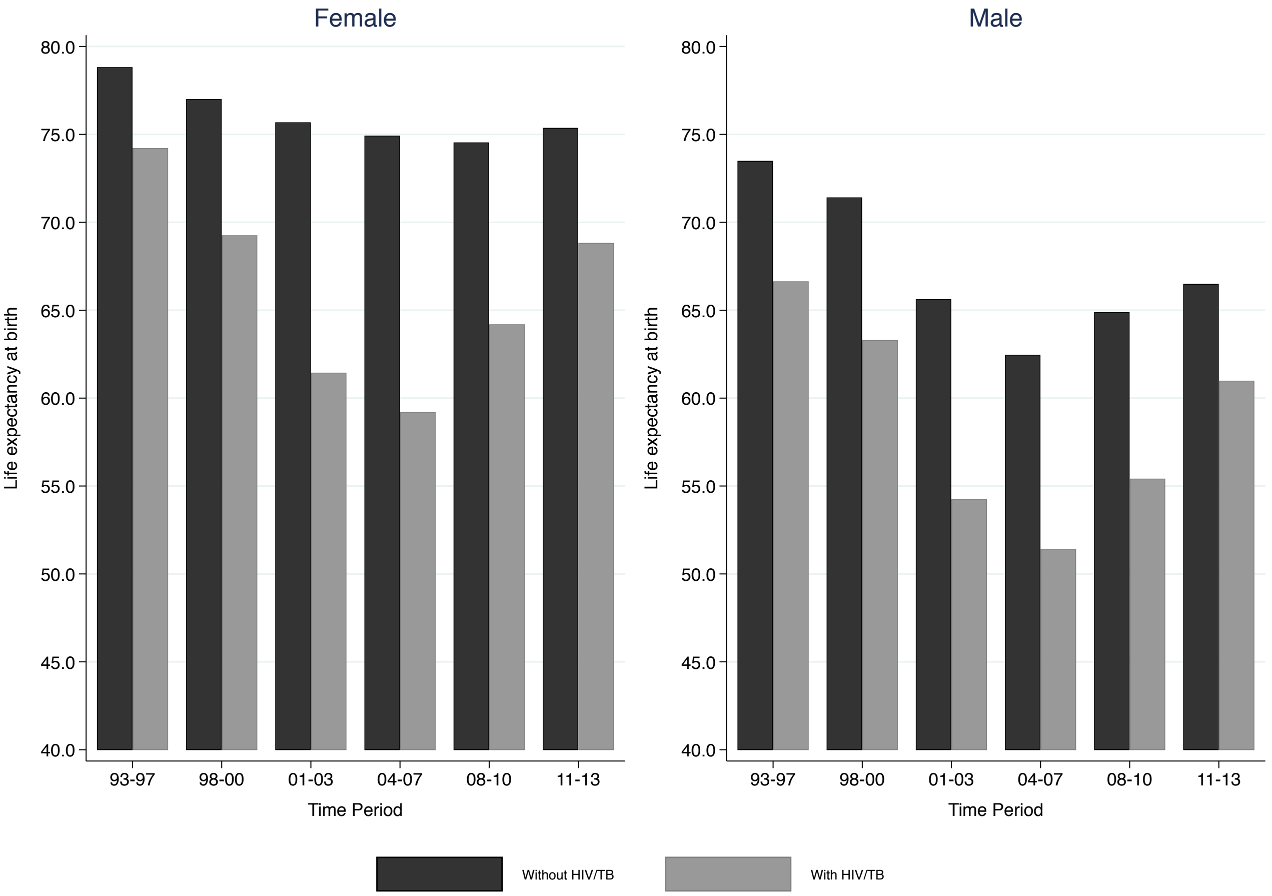


Figure 4.7: HIV-cause-deleted life expectancy at birth

Table 4.11: Potential years gained in life expectancy at selected ages if HIV/AIDS and TB causes were deleted

	Age	1993-1997	1998-2000	2001-2003	2004-2007	2008-2010	2011-2013
Females	0	4.6	7.7	14.2	15.7	10.3	6.5
	5	4.1	7.1	13.2	15.1	10.1	6.4
	15	4.0	7.1	13.1	15.1	9.8	6.3
	50	2.4	3.0	4.9	6.2	4.1	2.8
	65	1.9	2.2	3.1	3.2	1.9	1.8
Males	0	6.9	8.1	11.4	11.0	9.5	5.5
	5	6.6	7.4	10.5	10.7	9.2	5.3
	15	6.5	7.4	10.4	10.5	9.1	5.3
	50	5.4	4.6	5.6	5.8	4.9	3.1
	65	4.3	3.9	3.1	2.9	3.0	1.8

4.1.6 Progression of the epidemiological transition

One of the main propositions of the theory of the epidemiological transition is that over time mortality and disease patterns in human populations transition from very high and fluctuating mortality concentrated at younger ages and largely caused by infectious diseases and nutritional deficiencies to relatively stable low

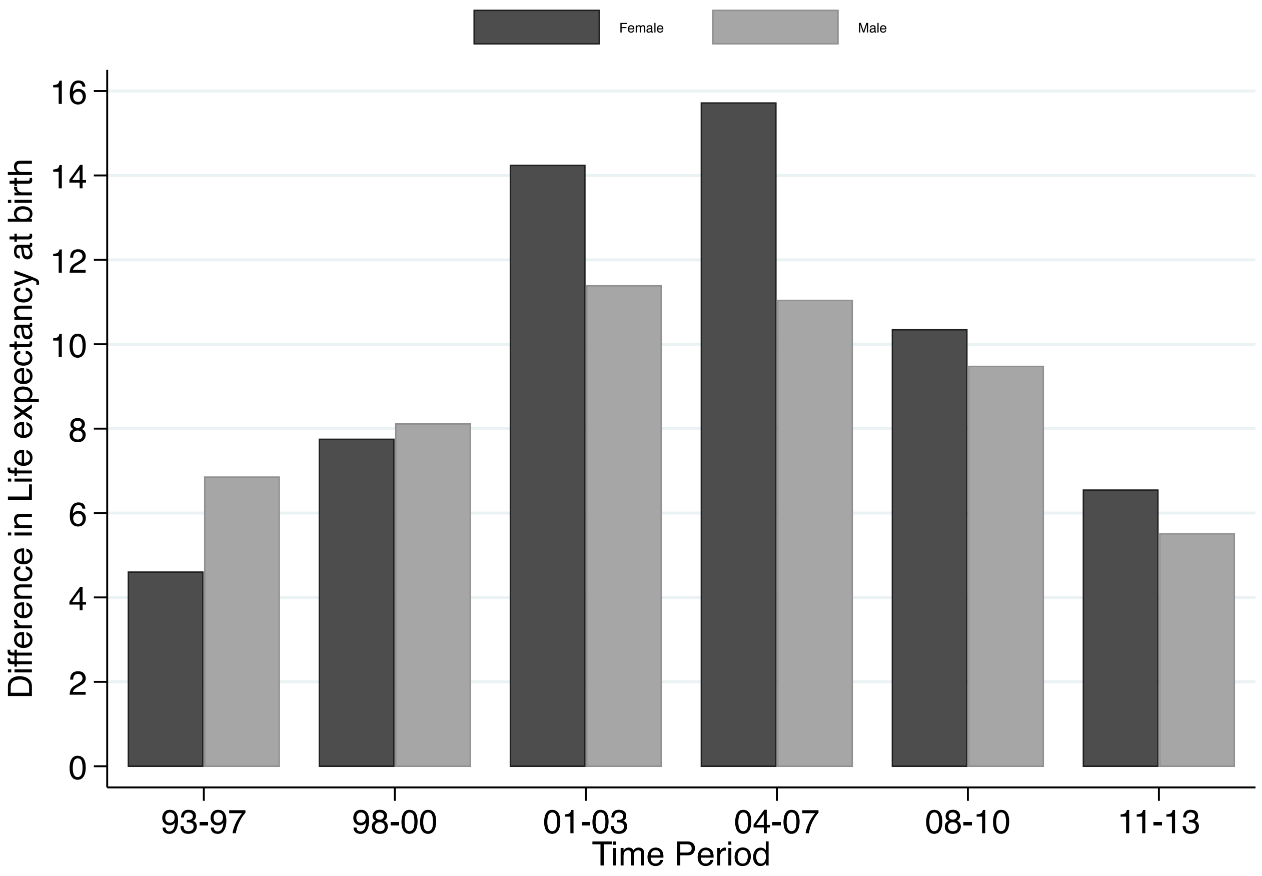


Figure 4.8: Years of life expectancy gained at birth if HIV/AIDS and TB causes were deleted by sex, and time period: Agincourt 1993-2013

mortality concentrated at older ages and largely caused by NCDs and injuries Omran (1971). **Paper II** further assesses the progression of the epidemiological transition in the Agincourt population by relating overall mortality levels to changes in the cause of death composition over the period 1993-2013. The results clearly show that the Agincourt population has experienced a protracted epidemiological transition characterised by reversals in changes in mortality and cause of death composition over the two decades, 1993-2013. Reversals in the epidemiological transition which began in the early 1990s (Kahn et al., 2007a; Kabudula et al., 2014b) due to increases in mortality attributable to HIV/AIDS and TB continued until around 2004-2007. The transition started to move in the positive direction, with falling overall mortality and standard (as predicted by the classic theory of the epidemiological transition) changes to the cause of death distribution thereafter. Thus, the progression of the epidemiological transition in the Agincourt population follows the "counter" and "protracted" transition pattern that Frenk et al. (1989) proposed for populations in LMICs based on experiences in Mexico and South America.

Figure 4.9 compares the percentage distribution of deaths due to communicable diseases (Group I), NCDs (Group II) and injuries (Group III) by year of death in the Agincourt surveillance population, South Africa and Africa Centre surveillance population. The figure shows that the overall changes over time in the percentage distribution of deaths due to the three broad categories of causes in the Agincourt surveillance population follow a similar pattern to the changes in the South African population although the percentages are different.

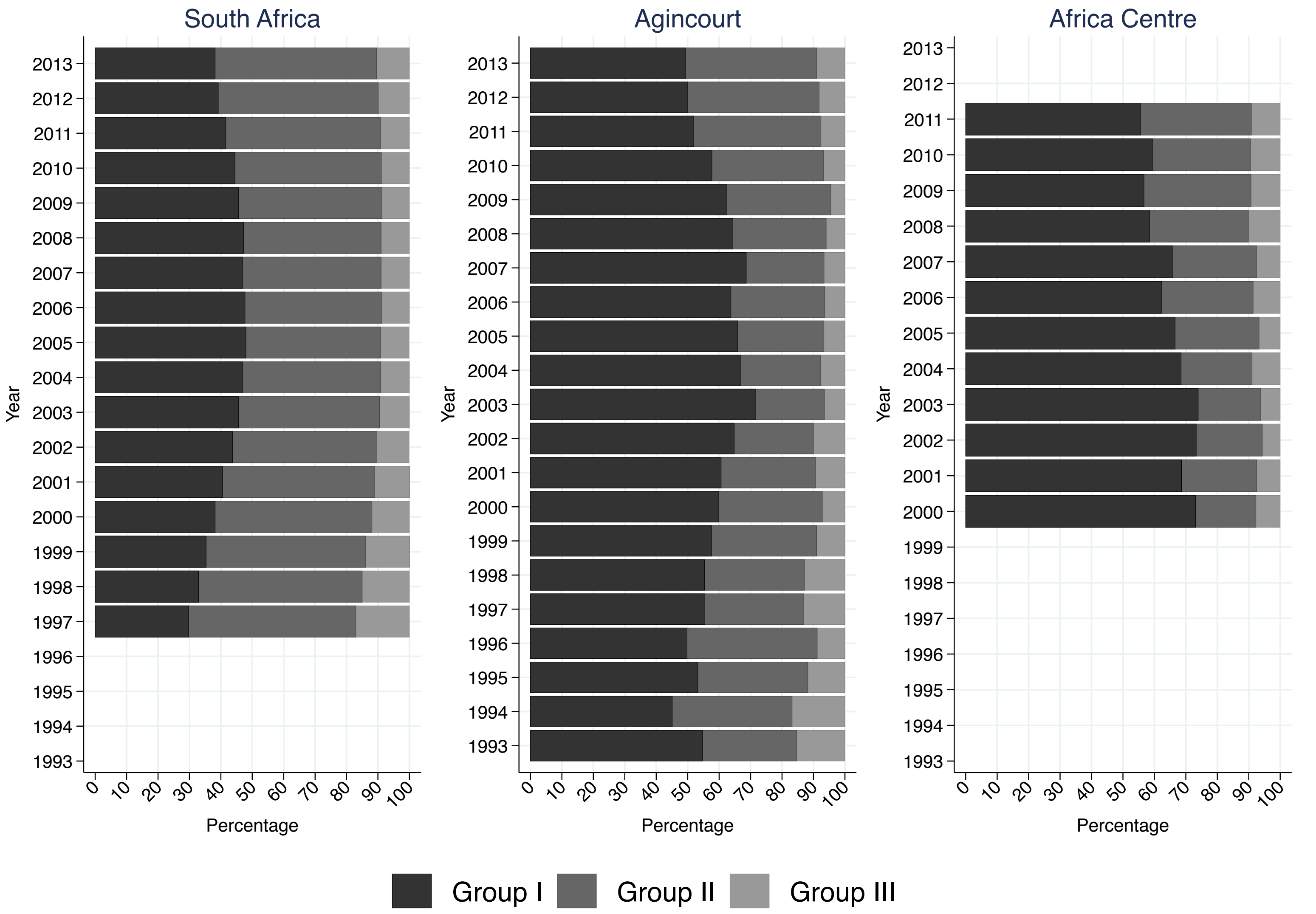


Figure 4.9: Cause of death distribution in South Africa, Agincourt and Africa Centre

Note: Percentages for South Africa come from a report by Statistics South Africa (Statistics South Africa, 2014a). Percentages for Africa Centre have been computed from the INDEPTH Network Cause-Specific Mortality - Release 2014 dataset (INDEPTH Network, 2014).

Chapter 5

Socioeconomic disparities in mortality indicators

Paper III assesses trends in socioeconomic differentials relating to changes in mortality patterns in the Agincourt population over the period 2001-2013. Emphasis is given to mortality in children under the age of five years and in adults aged 15-59 years because they were heavily impacted by the HIV/AIDS epidemic which was the major driver of the changes in mortality and cause of deaths patterns in the Agincourt study population.The paper documents significant socioeconomic gradients in under-five mortality (probability of death from birth to age 5 years), adult mortality (probability of death from age 15 to 60 years) and life expectancy at birth, with better outcomes favouring individuals from the highest SES households (Table 1 in the appendix of **Paper III** and the associated Figure 2 in the same paper). In 2001-2003, under-five mortality was 90.95 (95% CI: 73.23-108.66) per 1 000 person-years in the poorest households and 53.98 (95% CI: 40.53-67.43) per 1 000 person-years in the richest households. Although under-five mortality had substantially reduced among both the poorest and richest households by 2011-2013, the difference remained significant (42.81, 32.57-53.05 in the poorest households *vs* 19.46, 12.26-26.67 in the richest households). An inverse socioeconomic gradient in adult mortality was significant for females from 2001 to 2007 and for males throughout the period 2001-2013. In females, adult mortality remained higher in the poorest households (440.31, 350.56-530.05 in 2008-2010 and 325.96, 266.43-378.64 in 2011-2013) in comparison to the richest households (322.57, 266.50-378.64 in 2008-2010 and 265.01, 224.25-305.78 in 2011-2013), but the difference was not statistically significant. Life expectancy at birth was significantly lowest in the poorest wealth quintile compared to the richest wealth quintile for females in the periods 2001-2003 and 2004-2007, but the differences progressively narrowed to non-significant levels from 2008-2010. The lowest life expectancy at birth was also seen in males in the poorest quintile compared to the richest quintile and significant differences persisted throughout the period 2001-2013.

Figure 5.1 visually shows the results of socioeconomic differences in summary mortality indicators as measured by RIIs presented in Table 3 of **Paper III**. In all the time periods, the RIIs for under-five mortality and adult mortality were greater than 1, indicating higher mortality at the lower end of the socioeconomic continuum. The RIIs for under-five mortality decreased between 2001-2003 and 2004-2007 and steadily increased afterwards, but the differences over time were not significant. Differences over time were also not significant for the RIIs in adult mortality for both males and females. The notable difference between males and females is that the RIIs for adult mortality decreased steadily from 1.81(95% CI: 1.33-2.45) in 2001-2003 to 1.29(95% CI: 1.16-1.44) in 2011-2013 for females while for males it fluctuated between 1.33 and 1.54. The RIIs for life expectancy at birth show that relative socioeconomic inequalities narrowed over time for both males and females, but with larger magnitudes and significant differences over time in females and not in males.

Table 3 of **Paper III** also shows that all SIIs for under-five mortality and adult mortality were positive, indicating higher mortality at the lower end of the socioeconomic continuum. Similar to the RIIs, the SIIs for under-five mortality decreased between 2001-2003 and 2004-2007 and steadily increased afterwards, but the differences over time were also not significant. The SIIs for adult mortality steadily declined with significant differences over time in females, but fluctuated over time and showed no significant differences over time in males. The SIIs for life expectancy at birth also decreased steadily with significant differences over time in females, but fluctuated over time with no significant differences over time in males.

Estimates of socioeconomic differentials in the risk of dying from specific groups of causes of death revealed a persistent strong inverse gradient between HIV/AIDS and TB mortality and household SES as shown in Figure 3 and Table 4 of **Paper III** and Figure 5.2 in this thesis. This inverse gradient continued even as ART became more widely available at no cost through the public health system. The results in **Paper III** also indicate that deaths from NCDs were increasing among individuals from poor households but the socioeconomic gradient between household SES and mortality from this cause of death category was yet to reach statistically significant levels. Another finding of **Paper III** is that deaths due to external causes also remained an important cause of mortality for males, but did not significantly vary by household SES.

All in all, the results in **Paper III** provide further insights into how the epidemiological transition has unfolded in the Agincourt population between 2001 and 2013. Specifically, with the large proportion of the mortality burden being borne by the poorer sector of the population, the findings show polarisation of the ongoing epidemiological transition. This was proposed by Frenk et al. (1989) based on experiences in Mexico as one of the key elements of the epidemiological transition in populations in LMICs.

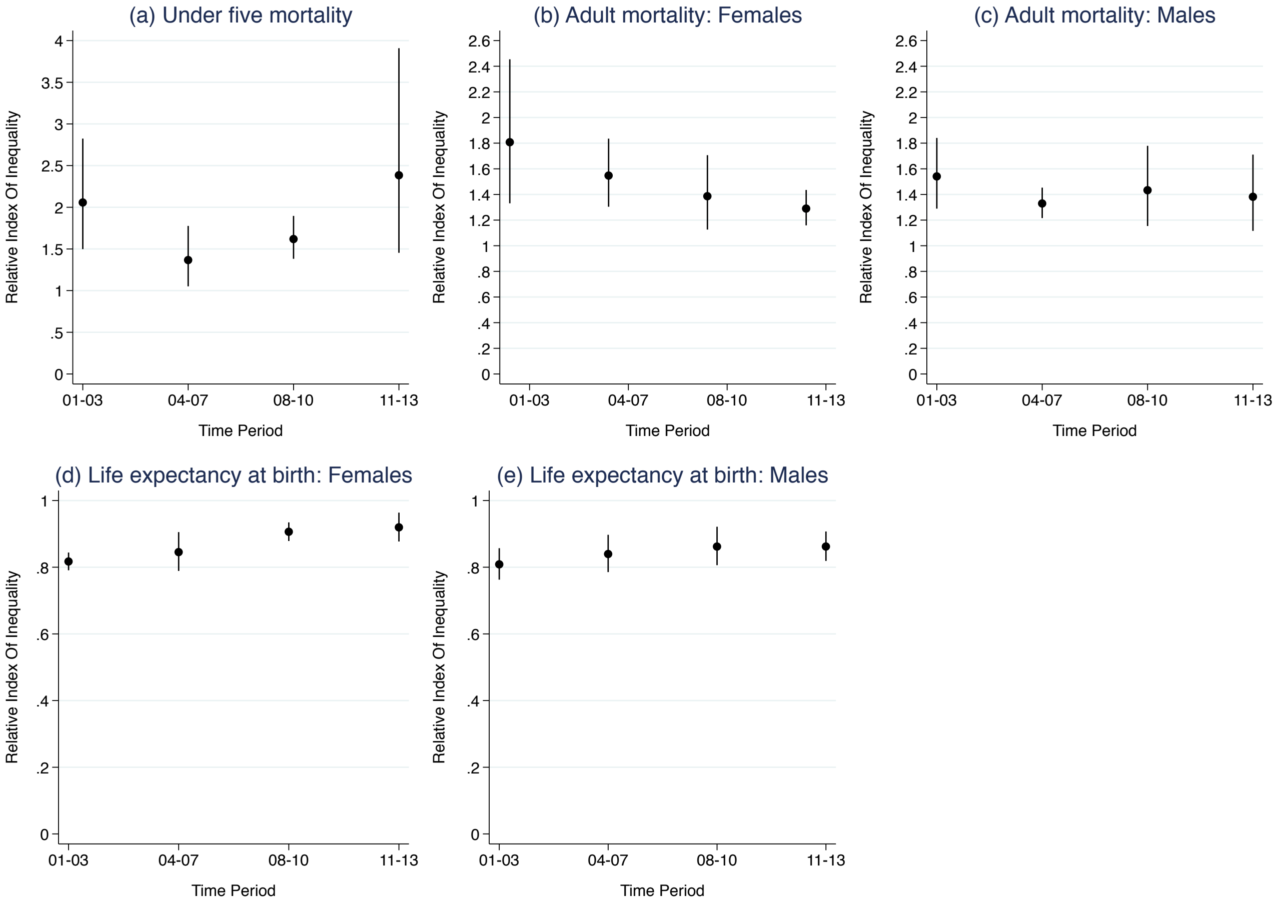


Figure 5.1: Relative inequalities in summary mortality indicators: Agincourt 2001-2013

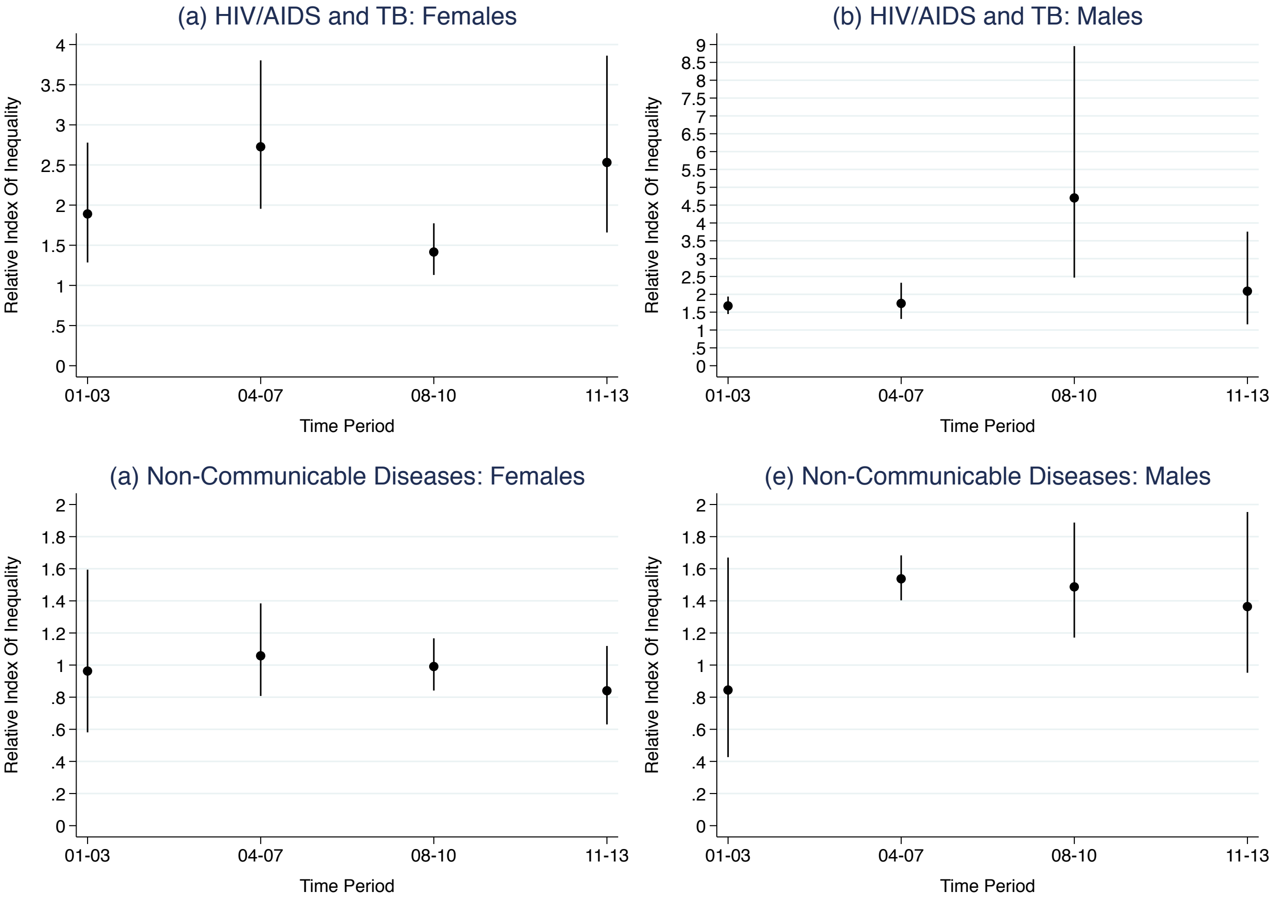


Figure 5.2: Relative inequalities in mortality by cause of death and sex: Agincourt 2001-2013

Chapter 6

Feasibility of using record linkage to enhance the utility of HDSS for studying utilisation of health services and its effects on mortality

Paper IV evaluates the coverage and quality of record linkage of data from the Agincourt surveillance population and patient records from one of the health facilities that serve the surveillance population. The linkage of the HDSS data with data from health facilities is important because once it is established it will make possible population-based investigations of the effect of socioeconomic disparities in the utilisation of healthcare services; and the effect of this on mortality risk (as well as morbidity or disability). Furthermore, it will also help identify improvements and gaps in clinical care provided, including patient adherence to medications and their availability.

The process followed in **Paper IV** involves creating a gold standard dataset of records matched by means of fingerprints and using it to evaluate the performance of 20 different record linkage scenarios with conventional personal identifiers. The various record linkage scenarios presented in Table 1 of Paper IV are distinguishable in a number of ways. At first, only personal identifiers that are routinely collected in health facilities (first name, surname, date of birth, sex and village) are used in one set of scenarios whereas an extended set of identifiers (that includes another household member's names, National ID number and telephone number) are used in another set of scenarios. Secondly, some scenarios use purely probabilistic methods of record linkage, whereas in others records are first matched deterministically using National ID number or telephone number and first name, and probabilistic record linkage is applied to the remaining records. Thirdly, different string comparison metrics are used for comparing names from the two data sources. At last, some scenarios are purely automated while others involve clerical review of a subset of record pairs.

The results show that the hybrid approach of first matching records deterministically using National ID number or telephone number and first name and thereafter applying probabilistic record linkage techniques that include another household member's first name as a matching variable in addition to routinely collected identifiers to the remaining records yield the best linkage results. One notable finding from the results is that another household member's first name substantially improves the linkage results. On the contrary adding another household member's surname as a matching variable in the probabilistic approach negatively affects the linkage results since it is often the same as that of the person to be linked and does not add much new information. With regard to string comparison metrics, the results show that the use of a combination of Soundex, Double Metaphone and a Jaro-Winkler score above 0.9 produce best linkage results. In the fully automated record linkage scenarios using a set of personal identifiers that are routinely collected in health facilities the best scenario (scenario S6 in Table 1 of Paper IV) yields a sensitivity of 75.28% and positive predictive value (PPV) of 90.89%. The sensitivity and PPV increase to 83.63% and 95.07%, respectively in the best fully automated record linkage scenario that use an extended set of identifiers and apply a hybrid deterministic-probabilistic approach (scenario S16 in Table 1 of Paper IV). When clerical review is performed on 10% of the record pairs around the matching score threshold of scenario S16, the sensitivity and PPV increase further to 84.27% and 96.86%, respectively.

The sensitivity and PPV reported in Paper IV are fully comparable to those reported in other published record linkage studies. For example, a study by Campbell et al. (2008), which compared the results of de-duplicating the client database of the Washington State's Division of Alcohol and Substance Abuse using three different record linkage algorithms, found sensitivity of 79.1%, 96.7% and 94.1% respectively with corresponding PPV of 97.4%, 96.1% and 94.8% at the optimal thresholds. More recently, a study by Paixão et al. (2017), which compared the accuracy of different linkage methods for assessing the risk of stillbirth due to dengue in pregnancy in Brazil found that the optimal method had sensitivity of 83.7% and PPV of 95.2%.

An assessment of the association between individual characteristics and the likelihood of successful linkage between the health facility and the population surveillance records found that in all record linkage scenarios records belonging to women (compared to men), non-South Africans (compared to native South Africans), poorly educated and older individuals were less likely to be matched. Nevertheless, the distribution of socio-demographic background characteristics in the gold standard dataset is similar to the distribution of socio-demographic background characteristics in the dataset generated via record linkage on conventional personal identifiers.

Chapter 7

Discussion and conclusions

7.1 General discussion

A clear understanding of the scale and patterning of mortality and diseases, their determinants and implications in different sub-populations is continuously needed in order to formulate and implement locally-relevant policies and programmes to improve population health. However, effective responses to this need for populations in most resource-poor settings are hampered by lack of comprehensive reliable population-based data on health risks, exposures and outcomes. One of the viable alternatives of addressing this data scarcity has been the establishment of HDSSs in some of the resource-poor settings. HDSSs enumerate populations in geographically well-defined areas and prospectively collect detailed information on core components of population change (births, deaths, and migrations) as well as complementary information on health, social and economic indicators (Bangha et al., 2010; Sankoh, 2010; Sankoh and Byass, 2012; Ye et al., 2012). In this thesis we have used data from one of the longest running HDSSs in Southern Africa to investigate the extent of changes in the cause composition of overall mortality and the socioeconomic patterning of the mortality changes that occurred in a poor rural population in northeast South Africa over the period 1993-2013. We have also assessed the feasibility of applying record linkage techniques to integrate data from HDSS and health facilities in order to enhance the utility of HDSS data for studying mortality and disease patterns and their determinants and implications.

7.1.1 Implications of findings on changes in SES

The results in **Paper I** show that the wide-ranging reforms introduced in South Africa by the post-apartheid government such as the provision of free basic services such as electricity (50 kWh per household per month), water, sanitation and housing and non-contributory social grants to previously disadvantaged populations (Bhorat and van der Westhuizen, 2013; Collinson, 2010; Lund, 2002) have reduced levels of absolute poverty and improved living conditions of the Agincourt population. However, the population still exhibits marked variations in levels of socioeconomic development. Over the 13-year period (2001-2013), the poorest segment of the population experienced little or no improvement in their SES. It is therefore important to identify households with chronically poor individuals and implement and evaluate targeted interventions that could improve their SES. At best, current strategies are only partially effective.

7.1.2 Changes in mortality patterns in Agincourt in the context of the epidemiological transition framework

The availability of mortality and cause of death information by age and sex for a well-defined rural population over an extended period (1993-2013) places the Agincourt surveillance population in a unique position to contribute to the understanding of the nature and pace of the epidemiological transition in poor rural sub-Saharan African settings. As shown in a study examining cause-specific mortality trends in selected populations in sub-Saharan Africa 1992-2012 - using data from selected INDEPTH HDSS sites - the data from Agincourt HDSS covered the longest time period and provided greater detail on the speed and complexity of disease changes underway across resource-poor African settings (Santosa and Byass, 2016). An exceptional feature of Agincourt HDSS data is that the period they cover includes the start of the HIV/AIDS epidemic, its rapid spread and growth, introduction of ART and widespread rollout of ART coupled with broader demographic, socioeconomic, technological, political, and cultural changes.

Papers II & III update and extend measures of the trends in mortality and cause of death profiles for the Agincourt surveillance population that were reported in earlier studies by among others Houle et al. (2014b); Kabudula et al. (2014b); Byass et al. (2010); Tollman et al. (2008); Kahn et al. (2007a). **Paper II** also operationalises the epidemiological transition using a statistical framework that allows characterisation of its progress by relating overall mortality levels to changes in the cause composition and statistically testing for changes and differentials. Together, these papers allow further critiquing of the applicability and universality of the propositions of the epidemiological transition theory and some of its subsequent refinements across diverse places and contexts.

The findings from **Papers II & III** indicate that some of the propositions in Omran's original epidemiological transition theory and its subsequent revised versions manifest in the Agincourt surveillance population. The empirical evidence from the Agincourt surveillance population clearly supports Omran's third proposition that "during the epidemiologic transition the most profound changes in health and disease patterns are obtained among children and young women". As per the findings in **Paper II**, overall mortality progressively increased from the early 1990s until around 2004-2007 then steadily declined in the subsequent time period. For the period that is characterised by falling overall mortality, the decline has been large for females compared to males. In addition, the mortality decline has progressed rapidly among children (aged 0-4 years) followed by young adults (aged 15-49 years) compared to older individuals. Therefore, the Agincourt surveillance population adds to the evidence base from many countries in support of Omran's third proposition documented in a review by Santosa et al. (2014). The findings from Agincourt also support the fourth proposition that in the less developed countries transitions are triggered by "medical progress, organised health care, and disease control programs that are usually internationally assisted and financed, and thus largely independent of the socioeconomic level of the country." This is evident in the declines in overall mortality experienced in the population in the recent years. Although the Agincourt population experienced significant improvements in ownership of household assets associated with modern wealth as documented in **Paper I**, the declines in mortality in recent years are mainly a product of the widespread availability and uptake of ART and PMTCT services that are successfully reducing the number of deaths attributable to HIV/AIDS and TB. The greater mortality burden (particularly from HIV/AIDs and TB) among the socioeconomically disadvantaged individuals reported in **Paper III** is also in line with the third proposition in Omran's revised version of the epidemiological transition theory published in 1998 (Omran, 1998). The proposition states that "during the transition profound inequalities in health and disease changes occur according to age, gender, race or ethnic origin, social class, indigenous status and geopolitical locales within or among countries". The co-occurrence of HIV/AIDS and NCDs also conforms to some of the features of the "age of triple health burden" in Omran's revised version of the epidemiological transition theory, in which populations in LMICs are expected to experience simultaneously three kinds of health burdens. These health burdens include: (i) continued existence of communicable diseases, perinatal and maternal morbidity and mortality, malnutrition, poor sanitation, persistent problems of poverty, low literacy, overpopulation and limited access to health care and safe water, particularly in rural areas, (ii) the emergence of a new set of heath problems whereby degenerative diseases such as heart diseases, stroke, cancer and metabolic disorders, stress and man-made diseases gradually increase and (iii) ill-prepared health systems and physicians and other health professionals in effectively dealing with the emerging health problems.

One major deviation from the classical epidemiologic transition theory that has been supported by some of the available empirical evidence is against the second proposition which states that "during the transition, a long-term shift occurs in mortality and disease patterns whereby pandemics of infection are gradually displaced by degenerative and man-made diseases as the chief form of morbidity and primary cause of death" in an orderly manner along a linear and unidirectional path distinguished by distinct stages. In some LMIC settings, changes in mortality and disease patterns have revealed that downward trends in mortality can be reversed, changes in mortality and disease patterns can be partial and different types of diseases can occur simultaneous in the same population (Santosa et al., 2014; Caselli et al., 2002; Frenk et al., 1989; Kahn et al., 2007a; Moser et al., 2005; Masquelier et al., 2014). The findings of reversals in mortality declines from the mid-1990s to around 2004-2007 and the concurrent occurrence of communicable and NCDs in the Agincourt population reported in **Paper II** further confirm that the second proposition of the epidemiologic transition theory is not universally applicable across diverse places and contexts. The changes in mortality and disease patterns in the Agincourt population documented in **Papers II & III** manifest elements of "counter-transition" and "protracted transition" which Frenk et al. (1989) proposed as key features of the epidemiological transition in populations in LMIC settings based on data from Mexico and South America. The element of counter-transition is evidenced by the reversals in the mortality declines and disease patterns, and protracted transition by the partial changes in mortality and disease patterns and the co-occurrence of HIV/AIDS and NCDs in older adults. Also contrary to the proposition that populations arrive to the "age of triple health burden" after completing the "age of pestilence and famine" followed by the "age of receding pandemics", the Agincourt surveillance population manifested features of the "age of triple health burden" before successfully completing the "age of receding pandemics."

7.1.3 Using record linkage to enhance the utility of HDSS for studying utilisation of health services and its effects on mortality

While the population-based longitudinal information on mortality and cause of death by age and sex collected through the rigorous data-collection procedures of a health and socio-demographic surveillance system in Agincourt makes it possible to assess some features of the progression of the epidemiological transition that is unfolding in this rural South African population, the utility of the data can be enhanced by integrating it with data from other sources. The value of integrating HDSS data with other data sources has been acknowledged by the INDEPTH network in its recently proposed new generation of population surveillance systems known as Comprehensive Health and Epidemiological Surveillance Systems (CHESS) (Sankoh, 2015). **Paper IV** reports findings from one of the pioneering studies on record linkage of population-based HDSS data and other data sources. The findings demonstrated the feasibility of record linkage of data from the Agincourt HDSS and data from local health facilities using conventional identifiers.

The findings from this pioneer work have facilitated other record linkage studies in Agincourt and elsewhere. In 2012 we successfully executed record linkage of Agincourt HDSS mortality data and the South African civil and vital registration system hosted at Statistics South Africa (Kabudula et al., 2014a) to study the level of agreement between VA cause of death information from the Agincourt HDSS and cause of death information from the civil registration death certification (Joubert et al., 2014). In addition, in 2014 we implemented a real time record linkage system in all public health facilities within the Agincourt HDSS study area to link information on clinical and treatment visits of HIV and chronic care patients to the information in the Agincourt HDSS database. This linkage work is on-going and has already facilitated two clustered randomised trials aimed at understanding and improving population health in Agincourt. One of the trials aims to test the hypothesis that the introduction of clinic-based lay health workers to assist nurses with the management of patients with chronic diseases in rural primary care clinics will improve the management of hypertension at the population level by improving diagnosis, retention in care and adherence to treatment by individuals with hypertension (Thorogood et al., 2014). The other trial aims to evaluate a community mobilisation intervention to improve engagement in HIV testing and care (Lippman et al., 2017). The real-time record linkage methods and approaches derived from the work in **Paper IV** have also been adopted in other ALPHA Network member sites (Network for Analysing Longitudinal Population based HIV/AIDS data on Africa) (see for example Rentsch et al. (2017b) and Rentsch et al. (2017a)).

7.1.4 Implications of the changing mortality patterns for health policy and practice

As findings from **Paper II** suggest, the epidemiological transition in rural South Africa will continue to be protracted in the near future, especially in adults of middle age. Furthermore, concentration of mortality will shift towards older age categories and the mortality burden from cardiovascular and other chronic NCDs will likely become more prominent as more people living with HIV/AIDS access ART and attain prolonged survival. The healthcare system therefore needs realignment for it to concurrently cater for multiple disease conditions.

Individuals from the poorest households in the resource-poor settings of rural South Africa also bear a disproportionately high mortality burden from the long-standing HIV/AIDS epidemic as supported by the finding of persistent SES disparities in HIV/AIDS and TB mortality over the period 2001-2013 despite wide availability of free ART in recent years (**Paper III**). Therefore, further reductions in HIV/AIDS and TB mortality and improvement of the overall health status of rural South African populations require new strategies to increase the uptake of HIV/AIDS testing and ART among individuals with a low SES.

7.1.5 Limitations of the data and methods

We acknowledge a number of limitations of the data and methods used in this thesis. First, we used data from one defined geographic region in rural South Africa. Therefore, as previously noted, the findings may not generalise to other settings (Houle et al., 2014b). Nevertheless, the comparability of the Agincourt population and the South African population in the general pattern of distribution by age of deaths, and the Agincourt population, the Africa Centre population and the South African population in the overall pattern of changes over time in the percentage distribution of deaths due to communicable diseases (Group I), NCD (Group II) and external causes (Group III) suggests that our findings may reflect the mortality and cause of death transition experienced by other rural populations in South Africa although the magnitudes may be different. There is also some indication that the general pattern of changes in the level of overall mortality in the Agincourt population may be similar to that of other rural populations in Southern Africa as evidenced by the finding of an upward trend in life expectancy in recent years in both the Agincourt population and some selected rural populations in Southern Africa. Secondly, updates of vital events in the Agincourt HDSS occur once a year. As a result, some still births, neonatal and infant deaths may not be recorded particularly when birth and death occur between consecutive household visits (Kahn et al., 2012). This bias is minimal in recent years because since 2000 names of the most recent child born to each woman appear on the pre-populated household roster and since 2006 there is careful probing for pregnancies and births since the last recorded child by asking about pregnancy status of every woman of childbearing age (Kahn et al., 2012). Thirdly, as pointed out by Kahn et al. (2012), the use of proxy respondents when updating the household roster and vital events (common to all HDSS data collection methods) may reduce the accuracy of some individual-level information (e.g. dates of birth, migration and death). Fourthly, SES is measured using household wealth indices constructed from information on ownership of household assets. By no means is this the only way to measure SES. Since the indices do not include other factors associated with social exclusion such as gender, education, occupation and ethnic background, the findings in this thesis may provide only a partial view of the multi-dimensional concept of poverty, inequality and inequity and associated disparities in mortality indicators. Fifthly, the data we use also do not include an individual measure of HIV seroprevalence and access to HIV care and treatment services. This made it difficult to determine the magnitude of excess HIV/AIDS-related mortality among individuals from poor households resulting specifically from increased risk of infection and lack of access to HIV care and treatment services. However, building on findings from **Paper IV**, future analyses using record linked data between the Agincourt HDSS and health facilities will unravel these aspects. Lastly, in addition to InterVA, there exist several other computer based tools that use VA information to assign causes of death in a standardised, automated, faster and more consistent manner than a physician would do such as Tariff (James et al., 2011) and InSilicoVA (McCormick et al., 2016). Because we only used the InterVA tool, the sensitivity of our findings to different computer-based tools for assigning causes of death is not known. However, we do not expect the use of a different tool to significantly alter our findings as the causes of death derived from the InterVA tool were found to be not substantially different from those generated by physician coding (Byass et al., 2011).

7.1.6 Future research areas

The investigations reported in this thesis can be extended in many ways. For example, leveraging the record linked data between the Agincourt HDSS and health facilities, future studies could investigate the effects of SES on utilisation of healthcare services and how that in return directly influences mortality risk. The record linkage of the HDSS and health facilities data would also facilitate monitoring and evaluation of progress towards universal health coverage which the South African government has committed to

(Ataguba et al., 2014). Furthermore, record linkage of the HDSS data and data from other sources such as school attendance registers in addition to data from health facilities will enhance the utility of the HDSS data further and facilitate other investigations of population-based health outcomes and their social determinants. Future studies could also investigate the sensitivity of findings on socioeconomic differentials in mortality risk to different computer-based automated methods of assigning causes of death. The investigation on socioeconomic changes in the population could also be extended to include other indicators of SES such as education and occupation. Also, as the findings in **Paper I** have shown, in the later years (∼ 2013) the variability in ownership of some of the household assets (e.g. cellphones) has reduced compared to the earlier year (2001). Future studies need to revise the household items included in the computation of SES indices by removing items that no longer discriminate between poor and rich households, and introducing new items that are associated with modern markers of wealth such as computers and smartphones. As expressed in **Chapter 1**, one important factor that influences disease risk and subsequent mortality patterns among the black population in South Africa is the oscillatory migration lifestyle. This thesis did not investigate the contribution of the widespread oscillatory migration lifestyle to the mortality patterns in the Agincourt population. This is an area that will be examined in future. Lastly, comparative analysis of mortality changes and associated socioeconomic differentials in the Agincourt population and other rural populations under extensive health and demographic surveillance in uMkhanyakude in KwaZulu-Natal province and Dikgale in Limpopo province can provide details of the generalisability of the findings reported in this thesis to other rural settings.

7.2 Conclusions

The investigations in this thesis have shown that the population in Agincourt, a rural region of South Africa, adjacent to southern Mozambique, experienced complex and rapidly evolving socioeconomic and health transitions over the period 1993-2013. Findings from the investigations highlight the importance of assessing mortality patterns and their risk factors at the local level in order to inform locally relevant public health responses. The finding that the poorest people experienced little or no improvement in their SES over the period 2001-2013 calls for strategies to identify the chronically poorest individuals and target them with interventions that can improve their SES and take them out of the vicious circle of poverty. As the findings have shown that the Agincourt population is experiencing a protracted epidemiological transition, with multiple stages overlapping and changes incomplete, continued progress in reducing preventable mortality and improving health across the life course will be affected by the intersection and interaction of HIV/AIDS and ART, non-communicable disease risk factors and complex social and behavioural changes. The finding that HIV/AIDS and tuberculosis mortality in the Agincourt population were associated with disparities in SES that remained unchanged over the period 2001-13 despite widespread availability and provision of free ART at public health facilities from around 2009 suggests that individuals from the poorest households continue to bear a disproportionately high burden of increased mortality and shortened lives related to the longstanding HIV/AIDS epidemic. These findings on mortality patterns and associated socioeconomic differentials bring to the fore the need for integrated health-care planning and programme delivery strategies to increase access to and uptake of HIV testing, linkage to care and ART, and prevention and treatment of NCDs among the poorest individuals to reduce the inequalities in cause-specific and overall mortality. These findings from a local, rural setting over an extended period also contribute to the evidence base to inform further refinement and advancement of health and epidemiological transition theory. Furthermore, the ability to link the HDSS data with data from health facilities will make it possible to conduct population-based investigations of the effect of socioeconomic disparities in the utilisation of healthcare services on mortality risk.

Bibliography

Adler, N. and Newman, K. (2002). Socioeconomic disparities in health: pathways and policies. *Health Affairs*, 21(2):60–76.

Alicandro, G., Frova, L., Sebastiani, G., Boffetta, P., and La Vecchia, C. (2017). Differences in education and premature mortality: a record linkage study of over 35 million italians. *European journal of public health*.

Allison, P. (1984). *Event history analysis: Regression for longitudinal event data*, volume 46. Sage.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13(1):61–98.

Allison, P. D. (2010). *Survival analysis using SAS: A practical guide*. SAS Institute.

Ataguba, J. E., Akazili, J., McIntyre, D., et al. (2011). Socioeconomic-related health inequality in south africa: evidence from general household surveys. *International Journal for Equity in Health*, 10(1):48.

Ataguba, J. E., Day, C., and McIntyre, D. (2014). Monitoring and evaluating progress towards universal health coverage in South Africa. *PLoS Medicine*, 11(9):e1001686.

Balen, J., McManus, D. P., Li, Y.-S., Zhao, Z.-Y., Yuan, L.-P., Utzinger, J., Williams, G. M., Li, Y., Ren, M.-Y., Liu, Z.-C., et al. (2010). Comparison of two approaches for measuring household wealth via an asset-based index in rural and peri-urban settings of hunan province, China. *Emerging Themes in Epidemiology*, 7(7).

Bangha, M., Diagne, A., Bawah, A., and Sankoh, O. (2010). Monitoring the millennium development goals: the potential role of the indepth network. *Global Health Action*, 3.

Barros, F. C., Victora, C. G., Scherpbier, R., and Gwatkin, D. (2010). Socioeconomic inequities in the health and nutrition of children in low/middle income countries. *Revista de Saúde Pública*, 44(1):1–16.

Beauchamp, A., Tonkin, A. M., Kelsall, H., Sundararajan, V., English, D. R., Sundaresan, L., Wolfe, R., Turrell, G., Giles, G. G., and Peeters, A. (2011). Validation of de-identified record linkage to ascertain hospital admissions in a cohort study. *BMC Medical Research Methodology*, 11(1):42.

Benzler, J., Herbst, K., and MacLeod, B. (1998). A data model for demographic surveillance systems.

Bhorat, H. and van der Westhuizen, C. (2013). Non-monetary dimensions of well-being in South Africa, 1993–2004: A post-apartheid dividend? *Development Southern Africa*, 30(3):295–314.

Bradshaw, D., Groenewald, P., Laubscher, R., Nannan, N., Nojilana, B., Norman, R., Pieterse, D., Schneider, M., Bourne, D., Timaeus, I., et al. (2003). Initial burden of disease estimates for South Africa, 2000. *South African Medical Journal*, 93(9):682–688.

Bradshaw, D., Laubscher, R., Dorrington, R., Bourne, D., and Timaeus, I. (2004). Unabated rise in number of adult deaths in South Africa. *South African Medical Journal*, 94(4):278–279.

Bradshaw, D., Nannan, N., Groenewald, P., Joubert, J., Laubscher, R., Nijilana, B., Norman, R., Pieterse, D., and Schneider, M. (2005). Provincial mortality in South Africa, 2000-priority-setting for now and benchmark for the future. *South African Medical Journal*, 95(7):496–503.

Braveman, P., Egerter, S., and Williams, D. R. (2011). The social determinants of health: coming of age. *Annual Review of Public Health*, 32:381–398.

Braveman, P. and Gruskin, S. (2003). Defining equity in health. *Journal of Epidemiology and Community Health*, 57(4):254–258.

Byass, P., Kahn, K., Fottrell, E., Collinson, M. a., and Tollman, S. M. (2010). Moving from data on deaths to public health policy in Agincourt, South Africa: Approaches to analysing and understanding verbal autopsy findings. *PLoS Medicine*, 7(8):e1000325.

Byass, P., Kahn, K., Fottrell, E., Mee, P., Collinson, M., and Tollman, S. (2011). Using verbal autopsy to track epidemic dynamics: the case of HIV-related mortality in South Africa. *Population Health Metrics*, 9(46):1–8.

Campbell, K., Deck, D., and Krupski, A. (2008). Record linkage software in the public domain: a comparison of link plus, the link king, and abasic'deterministic algorithm. *Health Informatics Journal*, 14(1):5.

Carolina, M. and Gustavo, L. (2003). Epidemiological transition: Model or illusion? a look at the problem of health in Mexico. *Social Science & Medicine*, 57(3):539–550.

Caselli, G., Mesle, F., and Vallin, J. (2002). Epidemiologic transition theory exceptions. *Genus*, 58(1):9–51.

Clark, D. (2004). Practical introduction to record linkage for injury research. *Injury Prevention*, 10(3):186.

Clark, S., Collinson, M., Kahn, K., Drullinger, K., and Tollman, S. (2007). Returning home to die: Circular labour migration and mortality in South Africa. *Scandinavian Journal of Public Health*, 35 Suppl 69:35–44.

Collinson, M. A. (2010). Striving against adversity: the dynamics of migration, health and poverty in rural South Africa. *Global Health Action*, 3.

Commission on Social Determinants of Health (2007). *A conceptual framework for action on the social determinants of health*. Geneve: World Health Organization.

Cook, L., Olson, L., and Dean, J. (2001). Probabilistic record linkage: relationships between file sizes, identifiers, and match weights. *Methods of Information in Medicine*, 40(3):196–203.

Coovadia, H., Jewkes, R., Barron, P., Sanders, D., and McIntyre, D. (2009). The health and health system of South Africa: historical roots of current public health challenges. *The Lancet*, 374(9692):817–834.

Drevenstedt, G., Crimmins, E., Vasunilashorn, S., and Finch, C. (2008). The rise and fall of excess male infant mortality. *Proceedings of the National Academy of Sciences*, 105(13):5016–5021.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83(402):414–425.

Embrett, M. G. and Randall, G. (2014). Social determinants of health and health equity policy research: exploring the use, misuse, and nonuse of policy analysis theory. *Social Science & Medicine*, 108:147–155.

Ernstsen, L., Strand, B. H., Nilsen, S. M., Espnes, G. A., and Krokstad, S. (2012). Trends in absolute and relative educational inequalities in four modifiable ischaemic heart disease risk factors: repeated cross-sectional surveys from the nord-trøndelag health study (hunt) 1984–2008. *BMC public health*, 12(1):266.

Eyawo, O., Franco-Villalobos, C., Hull, M. W., Nohpal, A., Samji, H., Sereda, P., Lima, V. D., Shoveller, J., Moore, D., Montaner, J. S., et al. (2017). Changes in mortality rates and causes of death in a population-based cohort of persons living with and without hiv from 1996 to 2012. *BMC infectious diseases*, 17(1):174.

Fellegi, I. and Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210.

Frenk, J., Bobadilla, J., Sepúlveda, J., and Cervantes, M. (1989). Health transition in middle-income countries: new challenges for health care. *Health Policy and Planning*, 4(1):29.

Frenk, J., Bobadilla, J., Stern, C., Frejka, T., and Lozano, R. (1991). Elements for a theory of the health transition. *Health Transition Review*, 1(1):21–38.

Graham, H. (2004). Social determinants and their unequal distribution: clarifying policy understandings. *The Milbank Quarterly*, 82(1):101–124.

Grannis, S., Overhage, J., and McDonald, C. (2002). Analysis of identifier performance using a deterministic linkage algorithm. In *Proceedings of the AMIA Symposium*, pages 305–309. American Medical Informatics Association.

Groenewald, P., Bradshaw, D., Daniels, J., Zinyakatira, N., Matzopoulos, R., Bourne, D., Shaikh, N., and Naledi, T. (2010). Local-level mortality surveillance in resource-limited settings: a case study of Cape Town highlights disparities in health. *Bulletin of the World Health Organization*, 88(6):444–451.

Gwatkin, D., Rutstein, S., Johnson, K., Suliman, E., Wagstaff, A., and Amouzou, A. (2007). *Socio-economic differences in health, nutrition, and population within developing countries*. Washington, DC: World Bank.

Handcock, M. and Morris, M. (1998). Relative distribution methods. *Sociological Methodology*, 28(1):53–97.

Handcock, M. and Morris, M. (1999). *Relative distribution methods in the social sciences*. Springer.

Harper, S., Lynch, J., Burris, S., and Davey Smith, G. (2007). Trends in the black-white life expectancy gap in the United States, 1983-2003. *The Journal of the American Medical Association*, 297(11):1224–1232.

Harron, K., Dibben, C., Boyd, J., Hjern, A., Azimaee, M., Barreto, M. L., and Goldstein, H. (2017). Challenges in administrative data linkage for research. *Big Data & Society*, 4(2):2053951717745678.

Harron, K., Goldstein, H., and Dibben, C. (2015). *Methodological developments in data linkage*. John Wiley & Sons.

Herbst, A., Mafojane, T., Newell, M., et al. (2011). Verbal autopsy-based cause-specific mortality trends in rural KwaZulu-Natal, South Africa, 2000-2009. *Population Health Metrics*, 9(1):47.

Herbst, A. J., Cooke, G. S., Bärnighausen, T., KanyKany, A., Tanser, F., and Newell, M.-L. (2009). Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa. *Bulletin of the World Health Organization*, 87(10):754–762.

Holman, C. J., Bass, A. J., Rouse, I. L., and Hobbs, M. S. (1999). Population-based linkage of health records in western australia: development of a health services research linked database. *Australian and New Zealand Journal of Public Health*, 23(5):453–459.

Hong, R. and Mishra, V. (2011). Effect of wealth inequality on chronic under-nutrition in Cambodian children. *Journal of Health, Population and Nutrition*, 24(1):89–99.

Hosegood, V., Vanneste, A., and Timæus, I. (2004). Levels and causes of adult mortality in rural South Africa: the impact of AIDS. *AIDS*, 18(4):663.

Hosseinpoor, A. R., Van Doorslaer, E., Speybroeck, N., Naghavi, M., Mohammad, K., Majdzadeh, R., Delavar, B., Jamshidi, H., and Vega, J. (2006). Decomposing socioeconomic inequality in infant mortality in Iran. *International Journal of Epidemiology*, 35(5):1211–1219.

Houle, B., Clark, S. J., Gómez-Olivé, F. X., Kahn, K., and Tollman, S. M. (2014a). The unfolding counter-transition in rural South Africa: mortality and cause of death, 1994–2009. *PLoS One*, 9(6):e100420.

Houle, B., Clark, S. J., Gómez-Olivé, F. X., Kahn, K., and Tollman, S. M. (2014b). The unfolding counter-transition in rural South Africa: mortality and cause of death, 1994–2009. *PLoS One*, 9(6):e100420.

Houle, B., Pantazis, A., Kabudula, C., Tollman, S., and Clark, S. J. (2016). Social patterns and differentials in the fertility transition in the context of hiv/aids: evidence from population surveillance, rural south africa, 1993–2013. *Population health metrics*, 14(1):1.

Howe, G. R. (1998). Use of computerized record linkage in cohort studies. *Epidemiologic Reviews*, 20(1):112–121.

Howe, L. D., Galobardes, B., Matijasevich, A., Gordon, D., Johnston, D., Onwujekwe, O., Patel, R., Webb, E. A., Lawlor, D. A., and Hargreaves, J. R. (2012). Measuring socio-economic position for epidemiological studies in low-and middle-income countries: a methods of measurement in epidemiology paper. *International Journal of Epidemiology*, page dys037.

Howe, L. D., Hargreaves, J. R., Gabrysch, S., and Huttly, S. R. (2009). Is the wealth index a proxy for consumption expenditure? a systematic review. *Journal of Epidemiology and Community Health*, 63(11):871–877.

INDEPTH Network (2014). Indepth network cause-specific mortality - release 2014. http://www.indepth-ishare.org/index.php/catalog/48.

Ingle, S., Margaret, M., Uebel, K., Timmerman, V., Kotze, E., Bachmann, M., Sterne, J. A., Egger, M., and Fairall, L. (2010). Outcomes in patients waiting for antiretroviral treatment in the free state province, south africa: prospective linkage study. *AIDS (London, England)*, 24(17):2717.

James, S. L., Flaxman, A. D., and Murray, C. J. (2011). Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31.

Jaro, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420.

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

Johnson, L. F., Dorrington, R. E., Laubscher, R., Hoffmann, C. J., Wood, R., Fox, M. P., Cornell, M., Schomaker, M., Prozesky, H., Tanser, F., et al. (2015). A comparison of death recording by health centres and civil registration in south africans receiving antiretroviral treatment. *Journal of the International AIDS Society*, 18(1).

Joubert, J., Bradshaw, D., Kabudula, C., Rao, C., Kahn, K., Mee, P., Tollman, S., Lopez, A. D., and Vos, T. (2014). Record-linkage comparison of verbal autopsy and routine civil registration death certification in rural north-east South Africa: 2006–09. *International Journal of Epidemiology*, 43(6):1945–1958.

Jutte, D. P., Roos, L. L., and Brownell, M. D. (2011). Administrative record linkage as a tool for public health research. *Annual Review of Public Health*, 32:91–108.

Kabudula, C. W., Houle, B., Collinson, M. A., Kahn, K., Tollman, S., and Clark, S. (2016). Assessing changes in household socioeconomic status in rural south africa, 2001–2013: A distributional analysis using household asset indicators. *Social Indicators Research*, pages 1–27.

Kabudula, C. W., Joubert, J. D., Tuoane-Nkhasi, M., Kahn, K., Rao, C., Gómez-Olivé, F. X., Mee, P., Tollman, S., Lopez, A. D., Vos, T., et al. (2014a). Evaluation of record linkage of mortality data between a health and demographic surveillance system and national civil registration system in South Africa. *Population Health Metrics*, 12(1):23.

Kabudula, C. W., Tollman, S., Mee, P., Ngobeni, S., Silaule, B., Gómez-Olivé, F. X., Collinson, M., Kahn, K., and Byass, P. (2014b). Two decades of mortality change in rural northeast South Africa. *Global Health Action*, 7.

Kahn, K., Collinson, M., Gómez-Olivé, F., Mokoena, O., Twine, R., Mee, P., Afolabi, S., Clark, B., Kabudula, C., Khosa, A., et al. (2012). Profile: Agincourt health and socio-demographic surveillance system. *International Journal of Epidemiology*, 41(4):988–1001.

Kahn, K., Garenne, M., Collinson, M., and Tollman, S. (2007a). Mortality trends in a new South Africa: Hard to make a fresh start. *Scandinavian Journal of Public Health*, 35(69 suppl):26.

Kahn, K., Tollman, S., Collinson, M., Clark, S., Twine, R., Clark, B., Shabangu, M., Gomez-Olive, F., Mokoena, O., and Garenne, M. (2007b). Research into health, population and social transitions in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System. *Scandinavian Journal of Public Health*, 35(69 suppl):8–20.

Karim, S., Churchyard, G., Karim, Q., and Lawn, S. (2009). HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. *The Lancet*, 374(9693):921–933.

Korda, R. J., Butler, J. R., Clements, M. S., and Kunitz, S. J. (2007). Differential impacts of health care in Australia: trend analysis of socioeconomic inequalities in avoidable mortality. *International Journal of Epidemiology*, 36(1):157–165.

Lee, P. and Paxman, D. (1997). Reinventing public health. *Annual Review of Public Health*, 18(1):1–35.

Li, B., Quan, H., Fong, A., and Lu, M. (2006). Assessing record linkage between health care and vital statistics databases using deterministic methods. *BMC Health Services Research*, 6(1):48.

Lippman, S. A., Pettifor, A., Rebombo, D., Julien, A., Wagner, R. G., Dufour, M.-S. K., Kabudula, C. W., Neilands, T. B., Twine, R., Gottert, A., et al. (2017). Evaluation of the Tsima community mobilization intervention to improve engagement in HIV testing and care in South Africa: study protocol for a cluster randomized trial. *Implementation Science*, 12(1):9.

Lopez, A., Mathers, C., Ezzati, M., Jamison, D., and Murray, C. (2006). Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *The Lancet*, 367(9524):1747–1757.

Lund, F. (2002). Social security and the changing labour market: Access for non-standard and informal workers in South Africa. *Social Dynamics*, 28(2):177–206.

Lyons, R., Jones, K., John, G., Brooks, C., Verplancke, J.-P., Ford, D., Brown, G., and Leake, K. (2009). The sail databank: linking multiple health and social care datasets. *BMC Medical Informatics and Decision Making*, 9(1):3.

Machado, C. (2004). A literature review of record linkage procedures focusing on infant health outcomes. *Cadernos de Saúde Pública*, 20:362–371.

Mackenbach, J. and Kunst, A. (1997). Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from europe. *Social Science & Medicine*, 44(6):757–771.

Mandacaru, P. M. P., Andrade, A. L., Rocha, M. S., Aguiar, F. P., Nogueira, M. S. M., Girodo, A. M., Pedrosa, A. A. G., de Oliveira, V. L. A., Alves, M. M. M., Paixão, L. M. M. M., et al. (2017). Qualifying information on deaths and serious injuries caused by road traffic in five brazilian capitals using record linkage. *Accident Analysis & Prevention*, 106:392–398.

Marmot, M. (2005). Social determinants of health inequalities. *The Lancet*, 365(9464):1099–1104.

Marmot, M., Friel, S., Bell, R., Houweling, T. A., Taylor, S., on Social Determinants of Health, C., et al. (2008). Closing the gap in a generation: health equity through action on the social determinants of health. *The Lancet*, 372(9650):1661–1669.

Maso, L., Braga, C., and Franceschi, S. (2001). Methodology used for software for automated linkage in italy (sali). *Computers and Biomedical Research*, 34(6):395.

Masquelier, B., Waltisperger, D., Ralijaona, O., Pison, G., and Ravélo, A. (2014). The epidemiological transition in Antananarivo, Madagascar: an assessment based on death registers (1900–2012). *Global Health Action*, 7.

Mayosi, B., Flisher, A., Lalloo, U., Sitas, F., Tollman, S., and Bradshaw, D. (2009). The burden of non-communicable diseases in South Africa. *The Lancet*, 374(9693):934–947.

Mayosi, B. M. and Benatar, S. R. (2014). Health and health care in South Africa – 20 years after Mandela. *New England Journal of Medicine*, 371(14):1344–1353.

Mayosi, B. M., Lawn, J. E., Van Niekerk, A., Bradshaw, D., Karim, S. S. A., Coovadia, H. M., team, L. S. A., et al. (2012). Health in South Africa: changes and challenges since 2009. *The Lancet*, 380(9858):2029–2043.

McCormick, T. H., Li, Z. R., Calvert, C., Crampin, A. C., Kahn, K., and Clark, S. J. (2016). Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049.

McMichael, A., McKee, M., Shkolnikov, V., and Valkonen, T. (2004). Mortality trends and setbacks: global convergence or divergence? *The Lancet*, 363(9415):1155–1159.

Meara, E., Richards, S., and Cutler, D. (2008). The gap gets bigger: changes in mortality and life expectancy, by education, 1981-2000. *Health Affairs*, 27(2):350–360.

Minujin, A. and Delamonica, E. (2004). Socio-economic inequalities in mortality and health in the developing world. *Demographic Research*, pages 329–54.

Montgomery, M. R., Gragnolati, M., Burke, K. A., and Paredes, E. (2000). Measuring living standards with proxy variables. *Demography*, 37(2):155–174.

Moser, K., Shkolnikov, V., and Leon, D. (2005). World mortality 1950-2000: divergence replaces convergence from the late 1980s. *Bulletin of the World Health Organization*, 83(3):202–209.

Mosley, W. H. and Chen, L. C. (1984). An analytical framework for the study of child survival in developing countries. *Population and Development Review*, 10(0):25–45.

Moure-Eraso, R., Flum, M., Lahiri, S., Tilly, C., and Massawe, E. (2007). A review of employment conditions as social determinants of health part II: the workplace. *New Solutions: A Journal of Environmental and Occupational Health Policy*, 16(4):429–448.

Msemburi, W., Pillay-van Wyk, V., Dorrington, R., Neethling, I., Nannan, N., Groenewald, P., and Bradshaw, D. (2016). *Second national burden of disease study for South Africa: Cause-of-death profile for South Africa, 1997–2010.* South African Medical Research Council, Cape Town.

Murray, C. and Lopez, A. (1997). Mortality by cause for eight regions of the world: Global burden of disease study. *The Lancet*, 349(9061):1269–1276.

National Department of Health (2015). The National Antenatal Sentinel HIV prevalence Survey, South Africa, 2013.

Nitsch, D., Morton, S., DeStavola, B., Clark, H., and Leon, D. (2006). How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950s study. *BMC Medical Research Methodology*, 6(1):15.

Nkonki, L. L., Chopra, M., Doherty, T. M., Jackson, D., Robberstad, B., et al. (2011). Explaining household socio-economic related child health inequalities using multiple methods in three diverse settings in South Africa. *International Journal for Equity in Health*, 10(1):13.

Nojilana, B., Bradshaw, D., Pillay-van Wyk, V., Msemburi, W., Laubscher, R., Somdyala, N. I., Joubert, J. D., Groenewald, P., and Dorrington, R. E. (2016). Emerging trends in non-communicable disease mortality in South Africa, 1997-2010. *South African Medical Journal*, 106(5):477–484.

Nolen, L., Braveman, P., Dachs, J., Delgado, I., Gakidou, E., Moser, K., Rolfe, L., Vega, J., and Zarowsky, C. (2005). Strengthening health information systems to address health equity challenges. *Bulletin of the World Health Organization*, 83(8):597–603.

Olshansky, S. and Ault, A. (1986). The fourth stage of the epidemiologic transition: the age of delayed degenerative diseases. *The Milbank Memorial Fund Quarterly*, 64:355–391.

Omran, A. (1971). The epidemiologic transition: a theory of the epidemiology of population change. *The Milbank Memorial Fund Quarterly*, 49(4):509–538.

Omran, Abdel, R. (1998). The epidemiologic transition theory revisited thirty years later. *World Health Statistics Quarterly*, 51(2/3/4):99–119.

Pacheco, A., Saraceni, V., Tuboi, S., Moulton, L., Chaisson, R., Cavalcante, S., Durovni, B., Faulhaber, J., Golub, J., King, B., et al. (2008). Validation of a hierarchical deterministic record-linkage algorithm using data from 2 different cohorts of human immunodeficiency virus-infected persons and mortality databases in brazil. *American Journal of Epidemiology.*

Paixão, E. S., Harron, K., Andrade, K., Rodrigues, L. C., Teixeira, M. G., Maria da Conceição, N. C., and Fiaccone, R. L. (2017). Evaluation of record linkage of two large administrative databases in a middle income country: stillbirths and notifications of dengue during pregnancy in brazil. *BMC medical informatics and decision making*, 17(1):108.

Patton, G., Coffey, C., Sawyer, S., Viner, R., Haller, D., Bose, K., Vos, T., Ferguson, J., and Mathers, C. (2009). Global patterns of mortality in young people: a systematic analysis of population health data. *The Lancet*, 374(9693):881–892.

Pillay-van Wyk, V., Msemburi, W., Laubscher, R., Dorrington, R. E., Groenewald, P., Glass, T., Nojilana, B., Joubert, J. D., Matzopoulos, R., Prinsloo, M., et al. (2016). Mortality trends and differentials in South Africa from 1997 to 2012: second national burden of disease study. *The Lancet Global Health*, 4(9):e642–e653.

Raphael, D. (2006). Social determinants of health: present status, unanswered questions, and future directions. *International Journal of Health Services*, 36(4):651–677.

Rentsch, C. T., Kabudula, C. W., Catlett, J., Beckles, D., Machemba, R., Mtenga, B., Masilela, N., Michael, D., Natalis, R., Urassa, M., et al. (2017a). Point-of-contact interactive record linkage (pirl): A software tool to prospectively link demographic surveillance and health facility data. *Gates Open Research*, 1.

Rentsch, C. T., Reniers, G., Kabudula, C., Machemba, R., Mtenga, B., Harron, K., Mee, P., Michael, D., Natalis, R., Urassa, M., et al. (2017b). Point-of-contact interactive record linkage (pirl) between demographic surveillance and health facility data in rural tanzania. *International Journal for Population Data Science*, 2(1).

Reppermund, S., Srasuebkul, P., Heintze, T., Reeve, R., Dean, K., Emerson, E., Coyne, D., Snoyman, P., Baldry, E., Dowse, L., et al. (2017). Cohort profile: a data linkage cohort to examine health service profiles of people with intellectual disability in new south wales, australia. *BMJ open*, 7(4):e015627.

Rogers, R., Hummer, R., and Krueger, P. (2005). Adult mortality. In Poston, D. and Micklin, M., editors, *Handbook of population*, pages 283–309. Springer.

Rosman, D. (1996). The feasibility of linking hospital and police road crash casualty records without names. *Accident Analysis & Prevention*, 28(2):271–274.

Sahn, D. E. and Stifel, D. (2003). Exploring alternative measures of welfare in the absence of expenditure data. *Review of Income and Wealth*, 49(4):463–489.

Salomon, J. and Murray, C. (2002). The epidemiologic transition revisited: compositional models for causes of death by age and sex. *Population and Development Review*, 28(2):205–228.

Sankoh, O. (2010). Global health estimates: stronger collaboration needed with low-and middle-income countries. *PLoS Medicine*, 7(11):e1001005.

Sankoh, O. (2015). Chess: an innovative concept for a new generation of population surveillance. *The Lancet Global Health*, 3(12):e742.

Sankoh, O. and Byass, P. (2012). The INDEPTH Network: Filling vital gaps in global epidemiology. *International Journal of Epidemiology*, 41(3):579–588.

Santosa, A. and Byass, P. (2016). Diverse empirical evidence on epidemiological transition in low-and middle-income countries: Population-based findings from indepth network data. *PloS one*, 11(5):e0155753.

Santosa, A., Wall, S., Fottrell, E., Högberg, U., and Byass, P. (2014). The development and experience of epidemiological transition theory over four decades: a systematic review. *Global Health Action*, 7.

Schröders, J., Wall, S., Hakimi, M., Dewi, F. S. T., Weinehall, L., Nichter, M., Nilsson, M., Kusnanto, H., Rahajeng, E., and Ng, N. (2017). How is indonesia coping with its epidemic of chronic noncommunicable diseases? a systematic review with meta-analysis. *PloS one*, 12(6):e0179186.

Sengayi, M., Spoerri, A., Egger, M., Kielkowski, D., Crankshaw, T., Cloete, C., Giddy, J., and Bohlius, J. (2016). Record linkage to correct under-ascertainment of cancers in hiv cohorts: The sinikithemba hiv clinic linkage project. *International journal of cancer*, 139(6):1209–1216.

Shisana, O., Rehle, T., Simbayi, L., Zuma, K., Jooste, S., Zungu, N., Labadarios, D., and Onoya, D. (2014). *South African national HIV prevalence, incidence and behaviour survey, 2012.* Cape Town: HSRC Press.

Siegel, J. S. (2011). *The demography and epidemiology of human health and aging.* Springer Science & Business Media.

Simmons, R., Ciancio, B., Kall, M., Rice, B., and Delpech, V. (2013). Ten-year mortality trends among persons diagnosed with hiv infection in england and wales in the era of antiretroviral therapy: Aids remains a silent killer. *HIV medicine*, 14(10):596–604.

Slobbe, L. C., Wong, A., Verheij, R. A., Oers, H. A., and Polder, J. J. (2017). Determinants of first-time utilization of long-term care services in the netherlands: an observational record linkage study. *BMC health services research*, 17(1):626.

Spijker, J. and Llorens, A. (2009). Mortality in Catalonia in the context of the third, fourth and future phases of the epidemiological transition theory. *Demographic Research*, 20(8):129–168.

Statistics South Africa (2014a). Mortality and causes of death in South Africa, 2013: Findings from death notification.

Statistics South Africa (2014b). South Africa - General Household Survey 2013.

Teng, A. M., Atkinson, J., Disney, G., Wilson, N., and Blakely, T. (2017). Changing socioeconomic inequalities in cancer incidence and mortality: Cohort study with 54 million person-years follow-up 1981–2011. *International journal of cancer*, 140(6):1306–1316.

Thorne, K., Johansen, A., Akbari, A., Williams, J., and Roberts, S. (2016). The impact of social deprivation on mortality following hip fracture in england and wales: a record linkage study. *Osteoporosis International*, 27(9):2727–2737.

Thorogood, M., Goudge, J., Bertram, M., Chirwa, T., Eldridge, S., Gomez-Olive, F. X., Limbani, F., Musenge, E., Myakayaka, N., Tollman, S., et al. (2014). The Nkateko health service trial to improve hypertension management in rural South Africa: study protocol for a randomised controlled trial. *Trials*, 15(1):435.

Tollman, S., Kahn, K., Sartorius, B., Collinson, M., Clark, S., and Garenne, M. (2008). Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. *The Lancet*, 372(9642):893–901.

Tollman, S. M., Kahn, K., Garenne, M., and Gear, J. S. (1999). Reversal in mortality trends: evidence from the Agincourt field site, South Africa, 1992-1995. *AIDS*, 13(9):1091–1097.

United Nations, Department of Economic and Social Affairs, Population Division (2011). World population prospects: The 2010 revision, cd-rom edition.

Uthman, O. A. (2009). Decomposing socio-economic inequality in childhood malnutrition in Nigeria. *Maternal & Child Nutrition*, 5(4):358–367.

Van de Poel, E., Hosseinpoor, A. R., Speybroeck, N., Van Ourti, T., and Vega, J. (2008). Socioeconomic inequality in malnutrition in developing countries. *Bulletin of the World Health Organization*, 86(4):282–291.

Van Hook, J. and Altman, C. E. (2013). Using discrete-time event history fertility models to simulate total fertility rates and other fertility measures. *Population research and policy review*, 32(4):585–610.

Victor, T. and Mera, R. (2001). Record linkage of health care insurance claims. *Journal of the American Medical Informatics Association*, 8(3):281–288.

Wagstaff, A., Paci, P., and Van Doorslaer, E. (1991). On the measurement of inequalities in health. *Social Science & Medicine*, 33(5):545–557.

Winkler, W. (1988). Using the EM algorithm for weight computation in the Fellegi-Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods, American Statistical Association*, volume 667, page 671.

Ye, Y., Wamukoya, M., Ezeh, A., Emina, J., and Sankoh, O. (2012). Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa? *BMC Public Health*, 12(1):741.

Ziraba, A., Fotso, J., and Ochako, R. (2009). Overweight and obesity in urban Africa: A problem of the rich or the poor? *BMC Public Health*, 9(1):465.

Zwang, J., Garenne, M., Kahn, K., Collinson, M., and Tollman, S. (2007). Trends in mortality from pulmonary tuberculosis and HIV/AIDS co-infection in rural South Africa (Agincourt). *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 101(9):893–898.

Appendix A: Published Papers

PAPER I

Soc Indic Res
DOI 10.1007/s11205-016-1397-z

Assessing Changes in Household Socioeconomic Status in Rural South Africa, 2001–2013: A Distributional Analysis Using Household Asset Indicators

Chodziwadziwa W. Kabudula[1,2] · Brian Houle[1,3,4] · Mark A. Collinson[1,5,6] · Kathleen Kahn[1,5,6] · Stephen Tollman[1,5,6] · Samuel Clark[1,4,5,7]

Accepted: 18 June 2016

Abstract Understanding the distribution of socioeconomic status (SES) and its temporal dynamics within a population is critical to ensure that policies and interventions adequately and equitably contribute to the well-being and life chances of all individuals. This study assesses the dynamics of SES in a typical rural South African setting over the period 2001–2013 using data on household assets from the Agincourt Health and Demographic Surveillance System. Three SES indices, an absolute index, principal component analysis index and multiple correspondence analysis index, are constructed from the household

✉ Chodziwadziwa W. Kabudula
chodziwadziwa.kabudula@wits.ac.za

Brian Houle
brian.houle@anu.edu.au

Mark A. Collinson
Mark.Collinson@wits.ac.za

Kathleen Kahn
Kathleen.Kahn@wits.ac.za

Stephen Tollman
Stephen.Tollman@wits.ac.za

Samuel Clark
work@samclark.net

[1] MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa

[2] Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK

[3] School of Demography, The Australian National University, Canberra, Australia

[4] Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, USA

[5] INDEPTH Network, Accra, Ghana

[6] Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden

[7] Department of Sociology, University of Washington, Seattle, WA, USA

Published online: 28 June 2016

asset indicators. Relative distribution methods are then applied to the indices to assess changes over time in the distribution of SES with special focus on location and shape shifts. Results show that the proportion of households that own assets associated with greater modern wealth has substantially increased over time. In addition, relative distributions in all three indices show that the median SES index value has shifted up and the distribution has become less polarized and is converging towards the middle. However, the convergence is larger from the upper tail than from the lower tail, which suggests that the improvement in SES has been slower for poorer households. The results also show persistent ethnic differences in SES with households of former Mozambican refugees being at a disadvantage. From a methodological perspective, the study findings demonstrate the comparability of the easy-to-compute absolute index to other SES indices constructed using more advanced statistical techniques in assessing household SES.

Keywords Agincourt · South Africa · Health and Demographic Surveillance System (HDSS) · Socioeconomic status (SES) · Household assets · Absolute index · Principal component analysis · Multiple correspondence analysis · Relative distribution methods

1 Introduction

An individual's or group's position within a hierarchical social structure known as socioeconomic status (SES) influences one's access to and control over desired resources including knowledge, money, power, prestige, and beneficial social connections which shape one's well-being and life chances (Link and Phelan 1995; Mueller and Parcel 1981; Link and Phelan 2005; Link et al. 2008; Phelan et al. 2010). Therefore, it is important to understand the distribution of SES and its temporal dynamics within a population to ensure that policies and interventions adequately and equitably contribute to the well-being and life chances of all individuals.

In low- and middle-income settings, one of the widely used measures of SES is a composite index constructed from a list of household asset items (Ataguba et al. 2011; Barros et al. 2010; Gwatkin et al. 2007; Hong and Mishra 2011; Hosseinpoor et al. 2006; Minujin and Delamonica 2004; Nkonki et al. 2011; Uthman 2009; Van de Poel et al. 2008; Ziraba et al. 2009). The index is often called a "wealth index" or "asset index" (Howe et al. 2012) and the household asset items on which it is derived from include durable goods, housing characteristics, sanitation and access to services. Balen et al. (2010), Howe et al. (2009, 2012), Montgomery et al. (2000) and Sahn and Stifel (2003) have outlined the theoretical basis for the preference of the asset index as a measure of SES in low- and middle-income settings over "direct" measures such as income, expenditure, and financial assets (e.g., savings and pensions). Supporting reasons range from reliability to time and cost effectiveness. For example, the information required to construct the asset index is relatively easy and inexpensive to collect. Additionally, in low- and middle-income settings, household assets provide a better proxy for a household's long-run wealth compared to information on income or expenditures; this is due to seasonal variability in earnings, income from potentially multiple and diverse informal activies, high rates of self-employment, likely recall bias and misreporting.

Booysen et al. (2008), Sahn and Stifel (2003) and Ward (2014) are among others who have demonstrated that data on household asset ownership collected at more than one point in time using a standardized questionnaire can be used to construct an asset index to compare and follow up the changes in the distribution of SES within populations. The Agincourt Health and Demographic Surveillance System (HDSS), which is central to the research programme of the MRC/Wits Rural Public Health and Health Transitions Research Unit has collected data on household asset ownership every 2 years since 2001 using a standardized questionnaire in the Agincourt sub-district in rural northeast South Africa. In this paper, we use these data to construct and compare asset indices and to assess the dynamics of SES in the Agincourt HDSS study population over the period 2001–2013. The focus is on the temporal changes in the ownership of various household asset items and the distribution of SES.

2 Materials and Methods

2.1 Data Sources

The analysis in this paper is based on data on asset indicators collected by the HDSS. The Agincourt system has collected detailed longitudinal data on vital events including births, deaths, in- and out-migrations, as well as complementary data covering health, social and economic indicators in a predominantly rural population in northeast South Africa every year since 1992 (Kahn et al. 2007, 2012). Until 2006, the study included 21 villages. The study area was extended to 26 villages in 2007. Another five villages were added between 2010 and 2012 in response to an expanding trials and evaluation portfolio. The population, of approximately 115,000 people in 2014, is largely Shangaan-speaking and almost a third are former Mozambican refugees who arrived in the area in the early to mid-1980s and their descendants.

Collection of data on household asset indicators that include construction materials of the main dwelling, type of toilet facilities, sources of water and energy, ownership of modern assets and livestock only started in 2001 and has been repeated every 2 years. To assess changes in the asset indicators over the period 2001–2013, we use only the data collected from households in the original 21 villages.

2.2 Statistical Analysis

There are three parts to the analysis. The first part summarizes changes in ownership of various household assets in the Agincourt study population from 2001 to 2013. The second part involves constructing three composite indices that can be used as a measure of SES from the household asset items. The three indices namely absolute index, principal components analysis (PCA) index and multiple correspondence analysis (MCA) index are among the most widely utilized indices in the literature. The three indices are used to assess the robustness of our findings. Similar to the approach adopted by Howe et al. (2008), the three indices are compared with each other using scatter plots and the percentage of households classified into the same and different SES quintiles. The agreement of classification of households into SES quintiles between indices is assessed using Kappa statistics. The Kappa statistic, which takes values between 0 (no agreement better than chance) and 1 (perfect agreement) measures agreement in classification between two methods taking into account the agreement that is expected based on chance alone (Howe et al. 2008). Also similar to the approach adopted by Balen et al. (2010), the Spearman's rank correlation coefficient is utilised for further comparisons of the three indices. The last part of the analysis applies the method of relative distributions developed by Handcock and Morris (1998, 1999) to the asset indices to assess changes in the distribution of SES over time in terms of location and shape. This part of the analysis also takes into account ethnic differences in the distribution of SES as a previous study by Sartorius and colleagues covering the period 2001–2007 showed persistent differentials in SES between the South African and Mozambican populations (Sartorius et al. 2013).

2.2.1 Construction of Asset Indices

The absolute index that we construct has been utilized by a number of other researchers that have analyzed data from the Agincourt HDSS (Houle et al. 2013; Gomez-Olive et al. 2014; Houle et al. 2014; Madhavan et al. 2012). To construct this index, first the items of each asset indicator are assigned a weight so that increasing values correspond to items associated with higher SES. For example, for the asset indicator wall material, *5 = brick; 4 = cement; 3 = other modern material; 2 = mud;* and *1 = other traditional material*. Thereafter, the value assigned to each item of an asset indicator is normalized by dividing it by the value assigned to the item associated with the highest SES. This results in items of a given asset indicator taking values within the range [0, 1]. The asset indicators are then grouped into five broad asset subcategories (modern assets, livestock, power supply, water and sanitation, and dwelling structure). The normalized values of the asset indicators within each subcategory are then summed to yield a subcategory-specific value. Each subcategory-specific value is further normalized so that it too is in the range [0, 1]. Finally, the five subcategory-specific normalized values are summed to produce an overall household asset index that falls in the range [0, 5].

The PCA index was first recommended by Filmer and Pritchett (2001) and is one of the most widely used asset indices (Gwatkin et al. 2007; McKenzie 2005; Minujin and Delamonica 2004). Construction of this index starts by constructing an $n \times p$ matrix, $\mathbf{X}$, representing ownership of p asset items collected from n households. Thereafter, each element of $\mathbf{X}$ is normalized by subtracting from it the column mean and dividing the difference by the column standard deviation to produce another $n \times p$ matrix, $\mathbf{Y}$. Next, a $p \times p$ correlation matrix, $\mathbf{R}$, is computed from the normalized data matrix, $\mathbf{Y}$. This is followed by solving the equation $(\mathbf{R} - \lambda\mathbf{I})\mathbf{V} = 0$ for λ and $\mathbf{V}$, where λ is a vector of eigenvalues, $\mathbf{I}$ is an identity matrix and $\mathbf{V}$ is a matrix of eigenvectors associated with the eigenvalues in λ. Each eigenvector is then scaled so that its sum of squares equals the total variance. The product of the normalized matrix of assets variables, $\mathbf{Y}$, and the matrix of scaled eigenvectors, $\mathbf{V}^*$ produces a set of uncorrelated linear combinations of the asset variables for each household j, known as *principal components*. For each household, the number of principal components equals the number of asset items, and the rank of each component corresponds to the rank of its associated eigenvector. The first component is associated with the most dominant (largest) eigenvalue and explains as much as possible of the variation in the original data. The second component is associated with the second largest eigenvalue and explains as much as possible of the remaining variation in the data, subject to being uncorrelated with the first component. Similarly, each subsequent component explains as much as possible of the remaining variation in the data, while being uncorrelated with the other components. Formally, for household j, the PCA index is computed as

$$A_j = v_{11}^{*}\left(\frac{\mathrm{x}_{j1-\bar{\mathrm{x}}_1}}{s_1}\right) + v_{21}^{*}\left(\frac{\mathrm{x}_{j2-\bar{\mathrm{x}}_2}}{s_2}\right) + \ldots + v_{p1}^{*}\left(\frac{\mathrm{x}_{jp-\bar{\mathrm{x}}_p}}{s_p}\right)$$

where v_{i1}^{*} are the elements of the scaled eigenvector associated with the largest eigenvalue, x_{ji} are the asset ownership values for household j and asset $i, i \in [1, 2 \ldots p]$, and $\bar{\mathrm{x}}_i$ and s_i are respectively, the mean and standard deviation of the asset ownership values across all households for asset item i. In our description of the steps to derive the PCA index we have kept the mathematical details to a minimum. More detailed mathematical descriptions of the steps involved in the PCA technique can be found in Everitt and Hothorn (2011), Rencher (2003).

The procedure used to construct the MCA index is similar to the one used to construct the PCA index but does not assume that the data are continuous and that there is a linear relationship between the observations (Traissac and Martin-Prevel 2012; Booysen et al. 2008; Howe et al. 2012). Because all the asset indicators are discrete or categorical, others have argued that the MCA index is the most appropriate asset-based measure of SES (Booysen et al. 2008; Traissac and Martin-Prevel 2012; Asselin and Anh 2008). In constructing the MCA index we follow the guidelines provided by Booysen et al. (2008) and Asselin and Asselin and Anh (2008). First, the indicators of asset ownership of all households are organized into a matrix **X** of ones and zeros called the "indicator matrix". In the indicator matrix, each categorical asset indicator is decomposed into a set of mutually exclusive and exhaustive binary categories that each take only the value 0 or 1 such that every household has a '1' in exactly one of each asset's set of categories and a '0' in the rest of the asset's categories. Second, a matrix **S** is calculated by taking the χ^2 metric on row/column profiles of **X**. Greenacre (2007) provides the formula for computing **S** as

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}\left(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathrm{T}}\right)\mathbf{D}_c^{-\frac{1}{2}}$$

where **P** is the matrix formed by dividing each element of the matrix **X** by the sum of its elements, **r** is a vector whose elements are the sums of the row elements of the matrix **P**, **c** is a vector whose elements are the sums of the column elements of the matrix **P**, and $\mathbf{D}_r$ and $\mathbf{D}_c$ are diagonal matrices formed from **r** and **c** respectively. Finally, singular value decomposition (SVD) is then performed on the matrix **S** to decompose it into three matrices such that $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathrm{T}}$ (Greenacre 2007). The columns of the matrices **U** and **V** referred to as left and right singular vectors are respectively the eigenvectors of the matrices $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ and $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ and the columns of the diagonal matrix $\mathbf{D}_\alpha$ known as singular values are the square roots of the common positive eigenvalues of $\mathbf{S}\mathbf{S}^{\mathrm{T}}$ and $\mathbf{S}^{\mathrm{T}}\mathbf{S}$. Like in the PCA approach, in constructing a single asset index, the elements in the first column vector of the matrix **V** derived by the SVD are then used as weights of the asset categories. Consequently, as provided by Booysen et al. (2008), the MCA index score for household i is calculated as

$$MCA_i = R_{i1}W_1 + R_{i2}W_2 + \cdots + R_{ij}W_j$$

where R_{ij} is the response of household i to asset category j and W_j is the MCA weight of asset category j.

The PCA and MCA indices are derived from pooled data from all the available years. This approach ensures that indices explain variation over time as well as across households and are not affected by changes in the contribution of particular assets to household welfare (McKenzie 2005). Pooling of the data is not necessary for the absolute index as the procedure used to generate this index assigns the same weight to the same asset item across time.

2.2.2 Assessing Distributional Changes in SES

The method of relative distributions that we apply to the three indices to assess trends in the distribution of SES quantifies differences between the distributions of a set of measurements of an attribute of interest from a population at one time period and another set of measurements of the same attribute from a different population, or from the same population at a later time period. It takes the values of one distribution (the comparison distribution) and expresses them as positions in another distribution (the reference distribution) (Handcock and Morris 1998, 1999). Compared to the standard approach of comparing distributions using summary statistics such as mean, median and variance, which do not consider the entire distributions, the relative distribution analytic approach allows direct comparisons between outcomes across the entire distributions and provides insights that may be missed by the former approach.

Taking 2001 as the baseline year, we obtain the relative distribution for each later time period, t, using the density function of the percentile rank, r, of asset index value,y, in 2001 as

$$g_t(r) = \frac{f_t(y)}{f_0(y)}, \quad 0 < r \leq 1$$

where $f_0(y)$ and $f_t(y)$ are the density functions of the asset index values in 2001 and at a later time period respectively. Basically, the relative distribution, $g_t(r)$, represents the ratio of the population density at asset index value, y, at each later time period, t, to the density in 2001. When there are no differences between the comparison and reference distributions, the relative distribution is uniform or "flat" (taking a value of 1 throughout). When there are differences between the distributions, the relative distribution "rises" or "falls" depending on the direction of the difference. For example, if the proportion of households at a later time period, t, with asset index values equal to the median asset index value in 2001 is less than 50 %, the relative distribution will have a value below 1 at a point on the vertical axis corresponding to 50 % on the horizontal axis.

Following the approach by Handcock and Morris (1998, 1999), the changes in the relative distribution of the asset index values in 2001 and at later time periods are statistically summarized using the entropy statistic and median relative polarization (MRP) index. The entropy statistic used is based on the Kullback–Leibler divergence, which is a measure of the distance between two distributions and is defined by:

$$D(F : F_0) = \int_{-\infty}^{\infty} \log\left(\frac{f(y)}{f_0(y)}\right) dF(y) = \int_{0}^{1} \log(g(r))g(r)dr$$

where $g(r)$ is the probability density function of the relative distribution of asset index values in the reference and comparison distributions and F_0 and F respectively represent the cumulative distribution functions of the reference and comparison distributions of asset index values. We use the entropy statistic to quantify: (1) overall divergence between the comparison and reference distributions; (2) divergence between the location-adjusted reference distribution and the reference distribution; and (3) divergence between the comparison distribution and the location-adjusted reference distribution. The location adjustment used is median adjustment. This is preferred over mean adjustment because of the well-known drawbacks of the mean when distributions are skewed. As for the MRP index, we use it to quantify the extent to which the shape difference between the distributions of asset index values in 2001 and at later time periods takes the form of relative polarization or rising inequality. It is computed as:

$$MRP_t = 4\int_0^1 \left|r - \frac{1}{2}\right| \times g_t(r)dr - 1$$

where $g_t(r)$ is the relative population density at asset index value, y at each time period, t weighted by the absolute difference between the baseline rank of y and the median, $\left|r - \frac{1}{2}\right|$. Its value varies between -1 and 1, with 0 representing no change in the distribution of asset index values at time period t relative to the baseline year, positive values representing more polarization (i.e. increases in the tails of the distribution) and negative values representing less polarization (i.e. convergence towards the center of the distribution). In order to distinguish the contributions from the lower and upper tails of the distribution to the overall polarization, the MRP index is decomposed into lower (LRP) and upper (URP) polarization indices defined respectively as:

$$LRP_t = 8\int_0^{\frac{1}{2}} \left|r - \frac{1}{2}\right| \times g_t(r)dr - 1$$

$$URP_t = 8\int_{\frac{1}{2}}^1 \left|r - \frac{1}{2}\right| \times g_t(r)dr - 1$$

These indices also vary between -1 and 1 and have similar interpretations as the MRP index.

The analysis of ethnic differences in the distribution of SES between the South African and Mozambican populations use the distribution of the asset index values of the Mozambican households as the reference distribution and that of the asset index values of the South African households as the comparison distribution.

2.3 Software

We use STATA version 13.1 (Stata Corp., College Station, USA) to construct the asset indices and to perform the descriptive analyses. We also utilize the R statistical package *reldist* to conduct the relative distribution analysis (Handcock and Aldrich 2002).

2.4 Ethics Statement

The Human Research Ethics Committee (Medical) of the University of the Witwatersrand reviewed and approved the Agincourt HDSS (protocol M960720 and M081145). At the start of surveillance in 1992, community consent was secured from civic and traditional leadership and has continuously been reaffirmed for over two decades through frequent meetings. This is facilitated by the Agincourt Unit's LINC (Learning, Information dissemination and Networking with Community) Office. Three local people working under a coordinator in the LINC office regularly engage with Community Development Forums as well as a Community Advisory Group in the study site. Both are elected committees comprising village members. Community Development Forums, the lowest level of local government, include the Induna who represents the Traditional Council. The LINC office ensures that Forum members understand research objectives and results and are able to raise concerns about the Unit's research in their communities, and provide feedback of research results at community meetings. The Community Advisory Group ensures information flows between the Unit and the community, voices concerns, assesses the potential impact of the Unit's research on the community, and maintains ongoing dialogue and consultation. At the individual and household level, informed verbal consent is obtained from the head of the household or an eligible adult in the household at each annual follow-up surveillance visit. Prior to conducting any interview, a local fieldworker who is well-trained and versed in the Agincourt HDSS methods and the process of verbal informed consent explains in the local language to the respondent the purpose, aims and justification of the HDSS as well as information about confidentiality, privacy and the right to refuse to participate or withdraw from the HDSS. The responsible fieldworker documents the consent process by marking out the respondent on the household roster as well as recording the fieldworker details and date on the spaces provided at the top of the household roster. A verbal consenting process is normal practice for HDSS and the processes followed in the Agincourt HDSS have continued to be accepted by the aforementioned ethics committee. Furthermore, additional ethical clearance was obtained from the same ethics committee for the primary study reported in this paper (protocol M120488).

2.5 Data Availability

Detailed documentation of the Agincourt HDSS data and an anonymized database containing data from 10 % of the surveillance households are freely available on the Agincourt HDSS website (www.agincourt.co.za). The specific customized data used in this study are available on request to interested researchers.

3 Results

3.1 Temporal Changes in Household Asset Ownership

Table 1 shows the percentage of households owning particular asset items in the 21 villages of the Agincourt HDSS over the period 2001–2013. The results indicate substantial increases over time in the proportions of households that own asset items associated with greater modern wealth. One notable change is the increase in the proportion of households with dwellings constructed with either brick or cement walls from 76 % in 2001 to 98 % in 2013. The prevalence of tiles as roof and floor materials of dwellings also respectively increased from 3 and 0.5 % in 2001 to 15 and 14 % in 2013. In addition, the proportion of households using electricity for lighting and cooking respectively increased from 69 and 4 % in 2001 to 96 and 50 % in 2013. Further noticeable changes are the increases in the proportions of households owning stove, fridge, cellphone and car respectively from 41, 40, 37 and 14 % in 2001 to 85, 86, 98 and 20 % in 2013. On the contrary, proportions of households that own asset items associated with traditional wealth such as animal drawn cart and livestock with the exception of chickens have remained persistently low. The prevalence of animal drawn cart remained nearly unchanged from 3 % in 2001 to 1 % in

Table 1 Percentage of households owning particular assets items in villages of Agincourt HDSS, South Africa, over the period 2001–2013 and weights assigned to each asset indicator in absolute, PCA and MCA SES indices

	Percentage of households owning asset							Asset weights		
	2001 (n = 10,974)	2003 (n = 11,501)	2005 (n = 11,341)	2007 (n = 11,253)	2009 (n = 12,760)	2011 (n = 11,549)	2013 (n = 11,363)	Absolute index	PCA index	MCA index
Type of dwelling's wall material										
Brick	1.43	2.06	2.52	3.79	6.05	3.98	4.68	5	0.04	1.66
Cement	75.00	78.99	85.26	88.40	88.11	92.42	93.14	4	0.16	0.34
Wood/other modern	1.22	0.37	0.63	0.73	0.68	0.61	0.55	3	−0.04	−2.41
Mud	21.10	17.80	11.12	6.76	4.69	2.49	1.35	2	−0.19	−3.40
Other traditional	1.25	0.77	0.48	0.32	0.46	0.49	0.27	1	−0.05	−3.62
Type of dwelling's roof material										
Tiles	3.16	4.45	5.60	7.00	9.49	12.37	14.78	3	0.12	2.66
Corrugated iron	90.67	90.71	91.15	90.95	89.38	86.86	84.68	2	−0.05	−0.12
Thatch/other traditional material	6.17	4.84	3.25	2.04	1.13	0.76	0.55	1	−0.11	−4.15
Floor material										
Tiles	0.46	1.01	1.90	2.67	4.58	9.75	13.57	7	0.10	3.28
Cement	90.06	92.17	93.59	94.81	93.62	88.94	85.68	6	0.02	0.01
Carpet	0.25	0.17	0.16	0.20	0.16	0.06	0.12	5	0.00	0.12
Wood/other modern	0.30	0.13	0.10	0.16	0.11	0.10	0.10	4	−0.01	−2.19
Dirt	5.54	4.50	3.45	1.48	0.77	0.88	0.33	3	−0.13	−4.64
Mat	0.20	0.06	0.05	0.06	0.17	0.13	0.06	2	−0.02	−3.09
Other traditional material	3.19	1.97	0.75	0.60	0.60	0.13	0.14	1	−0.09	−4.98
Number of bedrooms in dwelling										
1	36.51	34.14	28.88	25.26	22.36	20.75	19.70	1	−0.20	−1.96
2	32.60	33.72	35.53	35.60	36.80	34.47	32.61	2	−0.01	−0.12

Table 1 continued

	Percentage of households owning asset							Asset weights		
	2001 (n = 10,974)	2003 (n = 11,501)	2005 (n = 11,341)	2007 (n = 11,253)	2009 (n = 12,760)	2011 (n = 11,549)	2013 (n = 11,363)	Absolute index	PCA index	MCA index
3	20.62	21.35	22.64	24.53	25.27	26.79	27.47	3	0.12	1.22
4	8.15	8.64	10.18	11.36	12.25	13.85	15.52	4	0.10	1.81
5 or more	2.11	2.15	2.76	3.25	3.32	4.14	4.71	5	0.06	1.99
Separate kitchen	61.29	72.90	73.08	79.20	76.80	66.20	71.64	1	0.11	0.43
No separate kitchen	38.71	27.10	26.92	20.80	23.20	33.80	28.36	0		−1.10
Separate living room	47.52	49.46	51.44	55.26	52.98	58.36	61.06	1	0.19	1.08
No separate living room	52.48	50.54	48.56	44.74	47.02	41.64	38.94	0		−1.26
Location of toilet facility										
Inside the dwelling	0.22	0.15	0.19	0.52	0.57	0.76	2.09	4	0.03	4.01
In the yard	56.35	59.80	65.99	71.23	73.22	78.47	81.35	3	0.26	0.88
Neighbour's compound	18.08	20.21	17.42	16.15	16.82	14.84	12.76	2	−0.14	−1.57
Bush	25.35	19.84	16.39	12.10	9.39	5.92	3.80	1	−0.21	−2.86
Type of toilet facility										
Modern/flush	0.20	0.15	0.25	0.26	0.28	0.54	2.11	4	0.03	4.33
Ventilated improved pit latrine	0.89	0.48	2.20	4.07	11.14	4.34	9.16	3	0.03	0.97
Traditional latrine/pit	59.29	59.54	63.94	69.19	79.16	75.44	72.44	2	0.23	0.78
No facility	39.61	39.83	33.61	26.48	9.42	19.69	16.29	1	−0.27	−2.33
Source of drinking water										
Tap in the house	0.75	0.50	0.62	0.50	1.41	1.21	0.61	6	0.03	2.31
Tap in the yard	17.99	8.90	16.49	23.43	26.36	30.20	31.75	5	0.14	1.45
Tap on the street	65.35	75.90	80.05	74.19	66.76	65.81	60.77	4	−0.12	−0.42

Table 1 continued

	Percentage of households owning asset							Asset weights		
	2001 (n = 10,974)	2003 (n = 11,501)	2005 (n = 11,341)	2007 (n = 11,253)	2009 (n = 12,760)	2011 (n = 11,549)	2013 (n = 11,363)	Absolute index	PCA index	MCA index
Water truck	0.31	0.06	0.47	0.66	5.26	2.16	6.26	3	0.03	1.07
Well	10.84	14.38	2.19	0.65	0.05	0.20	0.39	2	−0.04	−1.50
Pond. river. dam. rain and other	4.77	0.25	0.19	0.57	0.16	0.42	0.23	1	−0.01	−1.07
Source of power for lighting										
Electricitry	68.66	76.32	89.35	90.38	93.99	95.51	96.44	4	0.22	0.46
Solar	0.06	0.08	0.04	0.98	0.60	0.12	0.18	3	−0.02	−1.92
Battery	0.08	0.10	0.15	0.10	0.10	0.11	0.24	2	0.00	0.07
Other	31.19	23.49	10.47	8.55	5.31	4.26	3.14	1	−0.22	−3.26
Source of power for cooking										
Electricitry	14.39	19.35	21.83	33.69	41.44	43.85	50.00	5	0.16	1.42
Gas	2.38	1.84	2.14	1.48	0.88	0.81	0.41	4	0.03	1.32
Parafin	6.98	5.36	4.41	1.58	0.44	0.24	0.11	3	−0.04	−1.38
Wood	75.72	73.24	71.51	63.11	57.13	55.02	49.38	2	−0.15	−0.69
Other	0.53	0.21	0.11	0.13	0.11	0.08	0.09	1	−0.02	−2.94
Modern assets										
Stove	41.47	42.53	53.44	66.96	75.27	82.96	85.38	1	0.22	1.03
No stove	58.53	57.47	46.56	33.04	24.73	17.04	14.62	0		−1.85
Fridge	40.52	45.63	57.83	68.09	75.02	81.70	86.22	1	0.26	1.13
No fridge	59.48	54.37	42.17	31.91	24.98	18.30	13.78	0		−2.13
TV	53.47	55.21	59.99	64.43	71.64	80.21	84.21	1	0.22	0.96
No TV	46.53	44.79	40.01	35.57	28.36	19.79	15.79	0		−1.96
Video	5.86	7.27	11.81	39.17	56.07	65.88	66.44	1	0.20	1.68
No video	94.14	92.73	88.19	60.83	43.93	34.12	33.56	0		−0.97

Table 1 continued

	Percentage of households owning asset							Asset weights		
	2001 (n = 10,974)	2003 (n = 11,501)	2005 (n = 11,341)	2007 (n = 11,253)	2009 (n = 12,760)	2011 (n = 11,549)	2013 (n = 11,363)	Absolute index	PCA index	MCA index
Satellite dish	0.26	0.27	0.51	1.59	5.40	12.44	17.65	1	0.11	3.09
No satellite dish	99.74	99.73	99.49	98.41	94.60	87.56	82.35	0		−0.18
Landline phone	3.24	2.05	1.63	2.29	1.17	1.73	0.87	1	0.04	1.71
No landline phone	96.76	97.95	98.37	97.71	98.83	98.27	99.13	0		−0.03
Cellphone	36.91	52.01	75.73	85.10	92.25	94.87	97.74	1	0.18	0.64
No cellphone	63.09	47.99	24.27	14.90	7.75	5.13	2.26	0		−2.11
Car	14.39	13.43	14.65	15.43	17.88	19.61	19.50	1	0.13	1.96
No car	85.61	86.57	85.35	84.57	82.12	80.39	80.50	0		−0.39
Motorbike	0.66	0.32	0.36	0.66	0.47	0.82	0.51	1	0.02	2.12
No motorbike	99.34	99.68	99.64	99.34	99.53	99.18	99.49	0		−0.01
Bicycle	13.21	10.32	10.04	8.41	9.98	9.96	5.76	1	0.05	0.94
No bicycle	86.79	89.68	89.96	91.59	90.02	90.04	94.24	0		−0.10
Livestock										
Animal drawn cart	3.28	2.40	1.89	1.92	2.12	1.73	1.28	1	0.03	1.23
No animal drawn cart	96.72	97.60	98.11	98.08	97.88	98.27	98.72	0		−0.03
No cows	84.91	87.19	87.67	88.04	88.39	87.42	87.88	0	−0.10	−0.19
1–3 cows	6.05	5.13	4.03	3.33	3.34	3.55	3.10	1	0.04	0.72
4–10 cows	5.81	5.37	5.60	5.48	5.32	5.75	5.58	2	0.07	1.47
More than 10 cows	2.70	1.71	2.00	2.63	2.31	2.75	2.80	3	0.05	1.91
Unknown number of cows	0.53	0.59	0.70	0.52	0.64	0.53	0.64	4	0.03	1.88
No goats	87.48	89.79	90.20	90.70	91.86	91.66	92.24	0	−0.04	−0.05
1–3 goats	7.64	4.38	3.94	3.94	4.04	4.08	3.60	1	0.02	0.28
4–10 goats	4.28	4.81	4.53	4.06	3.36	3.39	3.38	2	0.03	0.61

Table 1 continued

	Percentage of households owning asset							Asset weights		
	2001 (n = 10,974)	2003 (n = 11,501)	2005 (n = 11,341)	2007 (n = 11,253)	2009 (n = 12,760)	2011 (n = 11,549)	2013 (n = 11,363)	Absolute index	PCA index	MCA index
More than 10 goats	0.48	0.83	1.05	1.13	0.57	0.77	0.56	3	0.01	0.82
Unknown number of goats	0.12	0.18	0.27	0.18	0.16	0.10	0.22	4	0.01	1.13
No chickens	37.58	45.33	60.27	61.01	58.11	52.03	60.62	0	−0.04	−0.19
1–10 chickens	33.34	35.95	29.21	21.35	18.88	20.79	18.63	1	−0.01	−0.21
11–40 chickens	20.24	13.76	7.95	12.40	15.16	21.29	12.85	2	0.05	0.73
More than 40 chickens	1.03	0.37	0.33	1.32	1.14	1.63	0.88	3	0.02	1.08
Unknown number of chickens	7.81	4.60	2.24	3.94	6.70	4.26	7.02	4	0.03	0.66
No pigs	95.74	97.31	97.41	97.73	98.04	97.94	98.29	0	−0.03	−0.02
1–3 pigs	3.73	2.20	2.02	1.55	1.32	1.33	0.96	1	0.02	0.50
4–10 pigs	0.46	0.40	0.43	0.55	0.50	0.59	0.55	2	0.02	1.74
More than 10 pigs	0.05	0.07	0.10	0.12	0.09	0.13	0.15	3	0.01	2.41
Unknown number of pigs	0.02	0.02	0.04	0.06	0.05	0.01	0.04	4	0.01	1.81

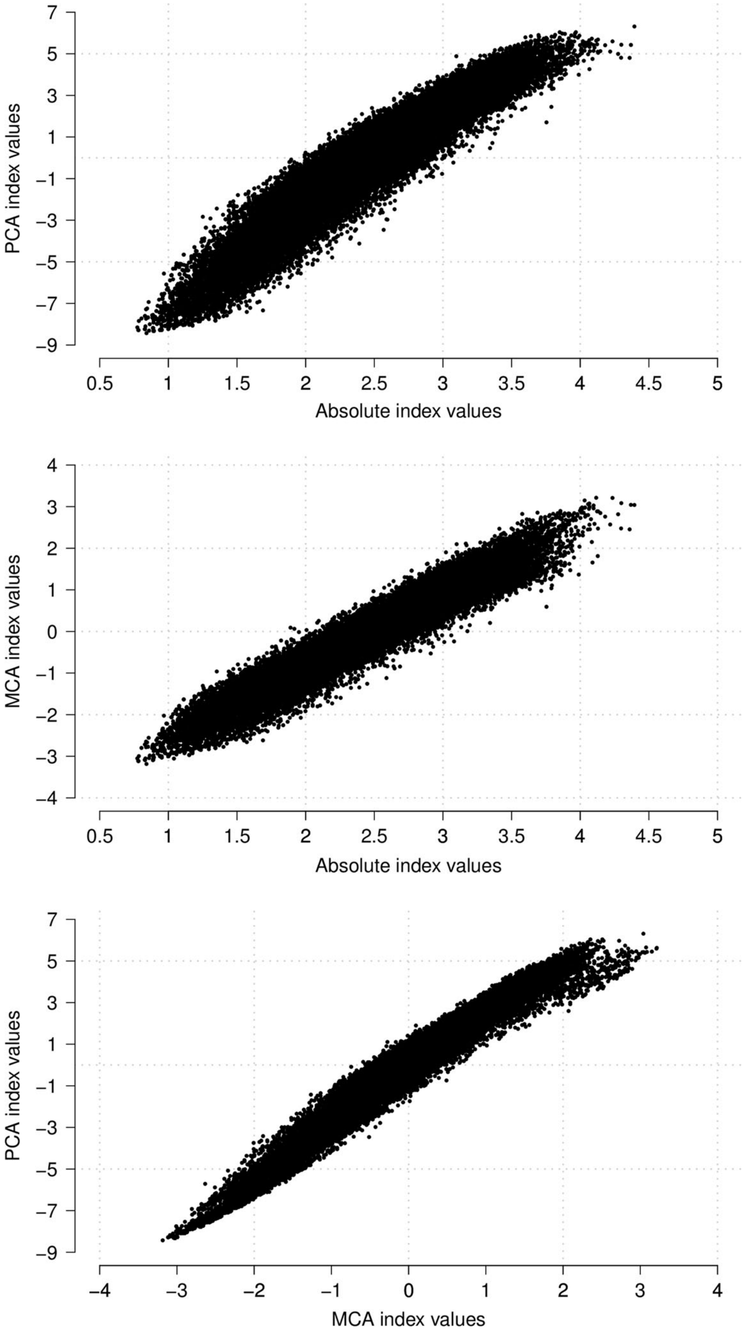


Fig. 1 Pairwise comparisons of asset index values

2013. Similarly, the proportion of households not owning cows or pigs only marginally changed from 85 % in 2001 to 88 % in 2013 for cows and from 96 % in 2001 to 98 % in 2013 for pigs. In addition, not owning goats slightly increased from 87 % in 2001 to 92 % in 2013.

3.2 Comparison of Asset Indices

The last three columns of Table 1 present the weights assigned to each asset item in constructing the three asset indices. For the absolute index, the weights are assigned in such a way that increasing values correspond to items associated with higher SES. For the PCA and MCA indices, positive weights are assigned to items expected to be associated with higher SES (e.g. tiles and cement housing floor materials, bricks and cement housing wall materials and tiles and corrugated iron sheets housing roof materials) and negative weights are assigned to items expected to be associated with lower SES (e.g. mud and other traditional housing floor and wall materials, and thatch and other traditional housing roof materials). However, on average the absolute values of the weights in the MCA index are higher than those in the PCA index. In addition, the ranking of the asset items based on the weights in the MCA and PCA indices show marked differences. From the PCA index, the highest weight is assigned to owning a toilet within the yard followed by owning a fridge and the lowest weight is assigned to not owning any toilet facility followed by sources of power for lighting other than electricity, solar or battery. From the MCA index, the highest weight is assigned to owning a toilet inside the dwelling followed by owning a flush toilet and the lowest weights are assigned to owning a house with the floor made of traditional materials such as dirt.

Despite the differences in the weights assigned to the asset items in the three indices, as shown in Fig. 1 and Table 2, the indices are reasonably comparable. Pairwise comparisons between the values of the indices result in correlation coefficients of at least 0.95. In addition, each pair of indices assigns at least 71 % of households in the same SES quintile with Kappa statistics of at least 0.64. Where a pair of indices places households in different quintiles, movement is generally limited to one quintile, with less than 1 % of households moving between two or more quintiles.

3.3 Distributional Changes in SES

Figure 2 shows the distribution of SES in the villages of the Agincourt HDSS over the period 2001–2013 based on the absolute, PCA and MCA indices. Overall, from one time period to the next, the mean and median values have persistently shifted to the right across all the three indices. Also it is apparent that the level of variability in the values of all the indices, as depicted by the standard deviation values, has progressively declined over time. Clearly, there has been both location and shape shifts in the SES distribution between 2001 and 2013.

Further insights into the key changes that have occurred in the distribution of the SES of households in the villages of Agincourt HDSS over the period 2001–2013 are provided by plots of the relative distribution of the densities of asset index values in selected years (2005, 2009 and 2013) to the density of asset index values in 2001 in Fig. 3, 4 and 5. The plots of overall distribution show the fraction of households in a particular year that fall into each decile of the 2001 SES distribution. The plots of location shift present the pattern of the relative distribution with no shape but only a location (median) shift in the SES

Table 2 Movement of households between quintiles of absolute, PCA and MCA indices

Indices being compared	Correlation coefficient	Percent of households moving between quintiles			Kappa statistic
		Same quintile	One quintile	Two quintiles	
Absolute and PCA	0.9561	71.28	28.11	0.60	0.6410
Absolute and MCA	0.9668	74.73	25.03	0.24	0.6841
PCA and MCA	0.9835	83.05	16.93	0.02	0.7881

distributions. The plots of the shape shift show the pattern of the relative distribution with no median but only a shape shift in the SES distributions.

In all the three indices, the value of the overall relative distributions is higher than one above the 7th decile of the 2001 distribution from 2009. This means that from 2009 there are higher proportions of households with asset index values that are above the asset index value in the 8th decile of the 2001 distribution. The entropy statistics for the overall relative distribution provide further evidence that irrespective of the index used, over time the distribution of SES has become more divergent from that in 2001. Using the absolute index, the entropy statistics moves from 0.127 in 2005 to 0.407 in 2009 and 0.603 in 2013. Using the PCA index, the entropy statistics moves from 0.0818 in 2005 to 0.377 in 2009 and 0.623 in 2013. Finally, if we use the MCA index, the entropy statistics changes from 0.1 in 2005 to 0.467 in 2009 and 0.747 in 2013.

The relative distributions with location shift illustrate that the effect of the median shift is quite large across all the three indices from 2009. In all the indices, changes in the median alone have caused the proportion of households with asset index values corresponding to the highest decile of the 2001 distribution in 2013 to be more than four times that in 2001. In addition, in all the indices the median shift alone has contributed more than 50 % of the overall entropy between the 2001 and 2013 distributions.

The median-adjusted relative distributions, which expose the effect of changes in distributional shape, show that for all the indices, the proportion of households with asset index values corresponding to the middle deciles (4th to 7th deciles) of the 2001 distribution has been increasing over time. Conversely, the proportion of households with asset index values corresponding to the lower and upper deciles of the 2001 distribution has been decreasing over time. This means that the distribution of SES has consistely become less polarized and is converging towards the middle over the years compared to 2001. Further details on the degree of convergence of the SES distribution from the two tails to the middle are provided by the median, lower and upper polarization indices and their corresponding 95 % confidence intervals, as reported in Table 3. The significantly negative values for the median index confirm that the SES distribution has been converging from the two tails to the middle. The significantly negative values for the lower and upper polarization indices confirm further that the convergence has occurred from both tails of the distribution. However, the large negative values for the upper indices compared to the lower indices indicate that the convergence towards the middle deciles from the upper tail of the distribution has been larger than that from the lower tail.

The analysis that takes into account ethnic background of the household head shows that improvement in SES has occurred for both South Africans and Mozambicans (Fig. 6).

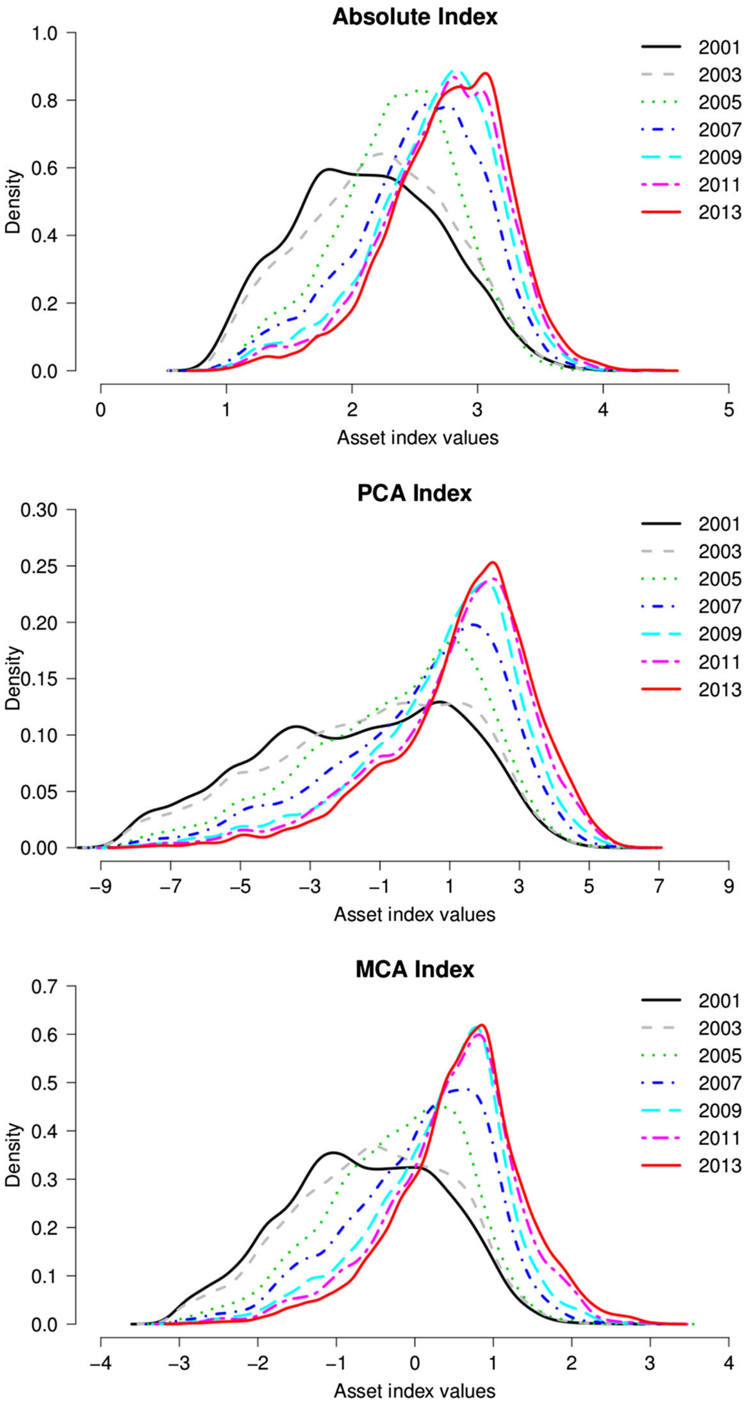


Fig. 2 Kernel density estimates of the distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013

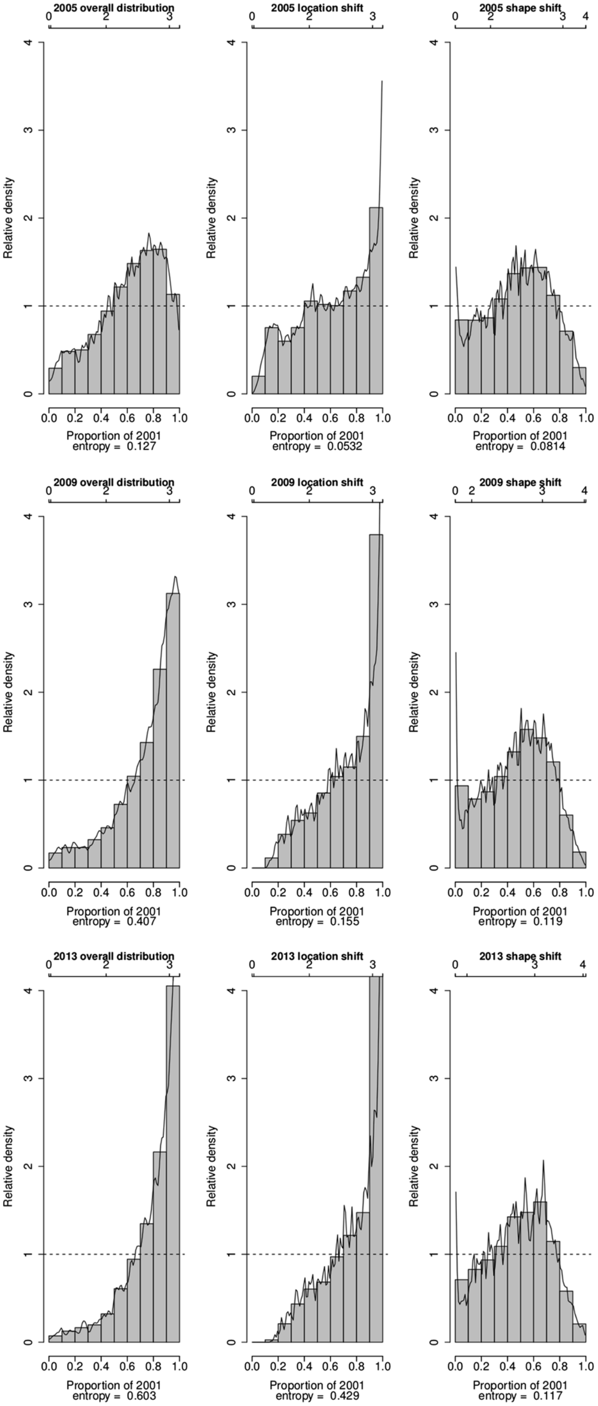


Fig. 3 Changes in the relatative distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013 based on absolute index

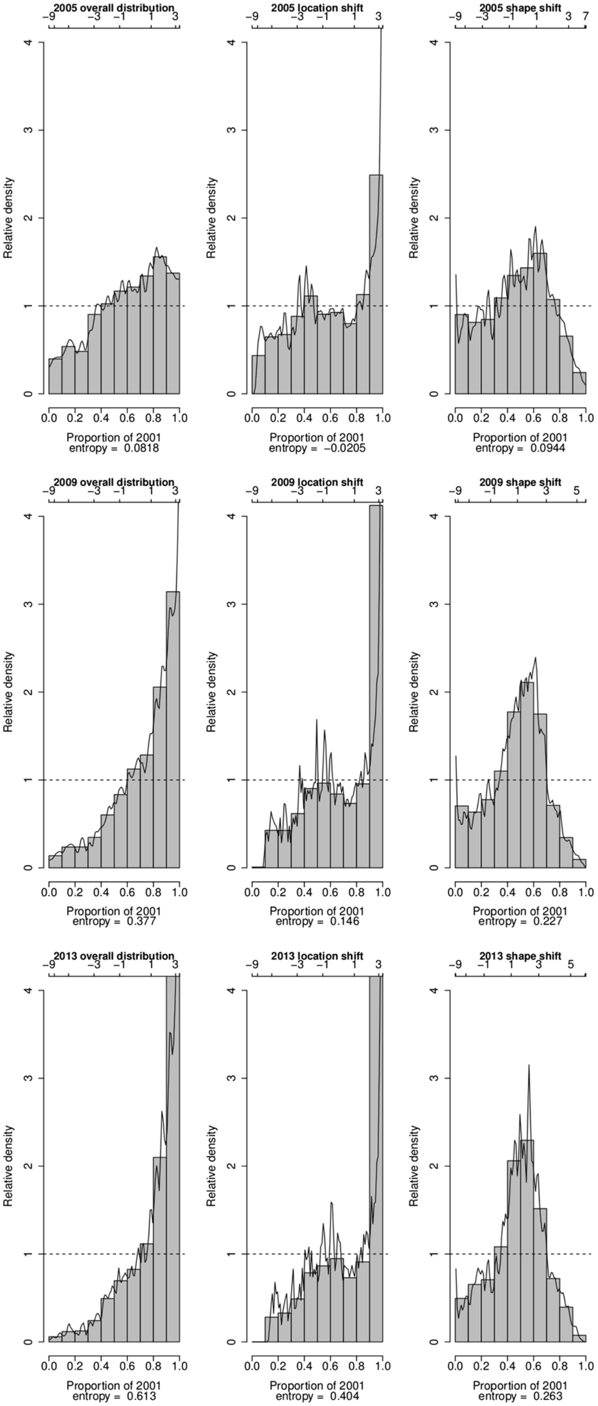


Fig. 4 Changes in the relatative distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013 based on PCA index

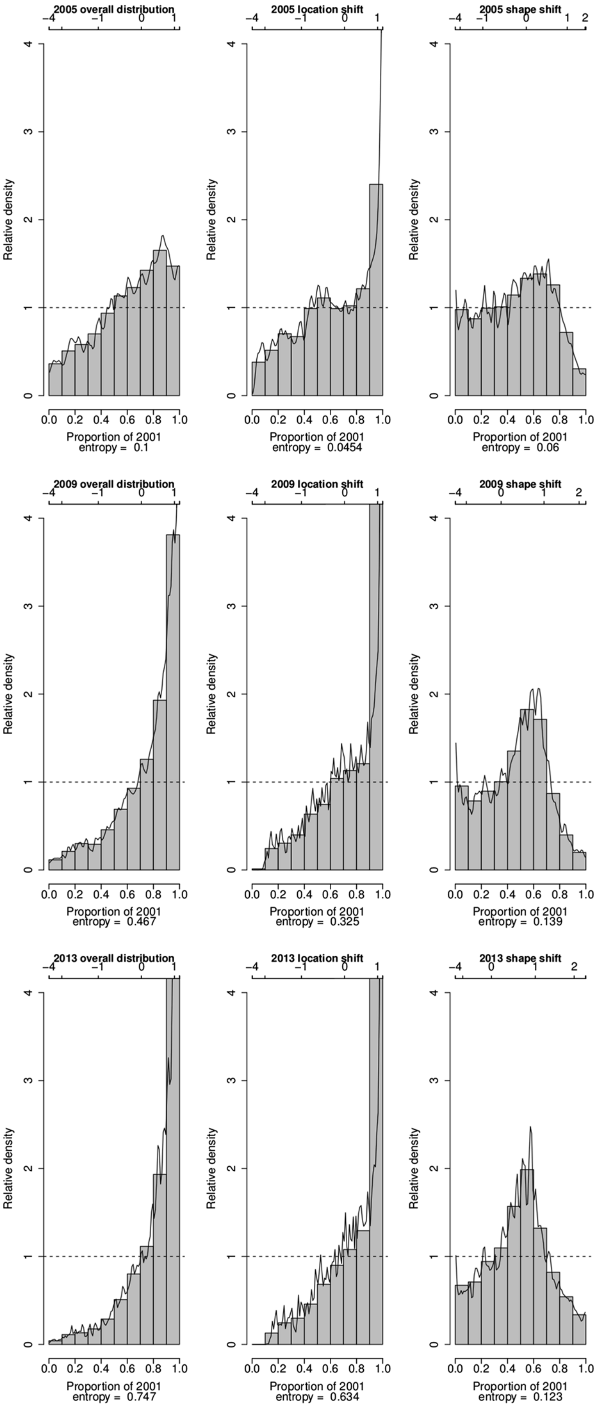


Fig. 5 Changes in the relatative distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013 based on MCA index

Table 3 Median polarization indices

	Lower CI	Estimate	Upper CI	*p* value
Absolute index				
2005 distribution compared to 2001 distribution				
Median	−0.193	−0.178	−0.163	<0.001
Lower	−0.136	−0.106	−0.076	<0.001
Upper	−0.279	−0.250	−0.221	<0.001
2009 distribution compared to 2001distribution				
Median	−0.203	−0.189	−0.174	<0.001
Lower	−0.101	−0.072	−0.043	<0.001
Upper	−0.331	−0.303	−0.274	<0.001
2013 distribution compared to 2001distribution				
Median	−0.224	−0.209	−0.194	<0.001
Lower	−0.159	−0.129	−0.099	<0.001
Upper	−0.318	−0.289	−0.260	<0.001
PCA index				
2005 distribution compared to 2001distribution				
Median	−0.201	−0.186	−0.171	<0.001
Lower	−0.127	−0.097	−0.067	<0.001
Upper	−0.304	−0.275	−0.246	<0.001
2009 distribution compared to 2001distribution				
Median	−0.346	−0.332	−0.318	<0.001
Lower	−0.247	−0.218	−0.189	<0.001
Upper	−0.472	−0.446	−0.420	<0.001
2013 distribution compared to 2001distribution				
Median	−0.389	−0.375	−0.361	<0.001
Lower	−0.326	−0.297	−0.269	<0.001
Upper	−0.479	−0.452	−0.426	<0.001
MCA index				
2005 distribution compared to 2001distribution				
Median	−0.148	−0.133	−0.118	<0.001
Lower	−0.072	−0.042	−0.011	<0.001
Upper	−0.254	−0.225	−0.195	<0.001
2009 distribution compared to 2001distribution				
Median	−0.242	−0.227	−0.213	<0.001
Lower	−0.114	−0.084	−0.055	<0.001
Upper	−0.398	−0.370	−0.343	<0.001
2013 distribution compared to 2001distribution				
Median	−0.271	−0.256	−0.242	<0.001
Lower	−0.207	−0.177	−0.147	<0.001
Upper	−0.364	−0.336	−0.307	<0.001

However, at each single point in time the Mozambicans on average have lower SES compared to the South Africans (Fig. 7). A comparison of the distributions of the SES of the two ethnic groups using relative distribution methods indicate that the differences are

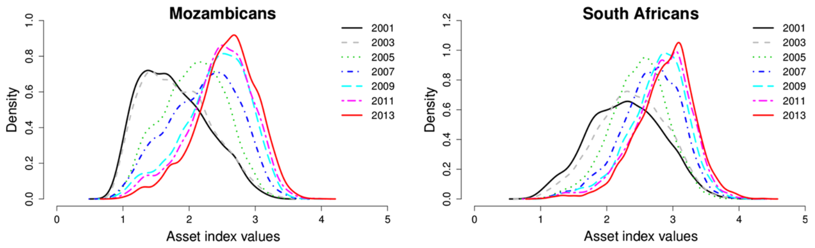


Fig. 6 Kernel density estimates of the distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013 by ethnicity based on absolute index

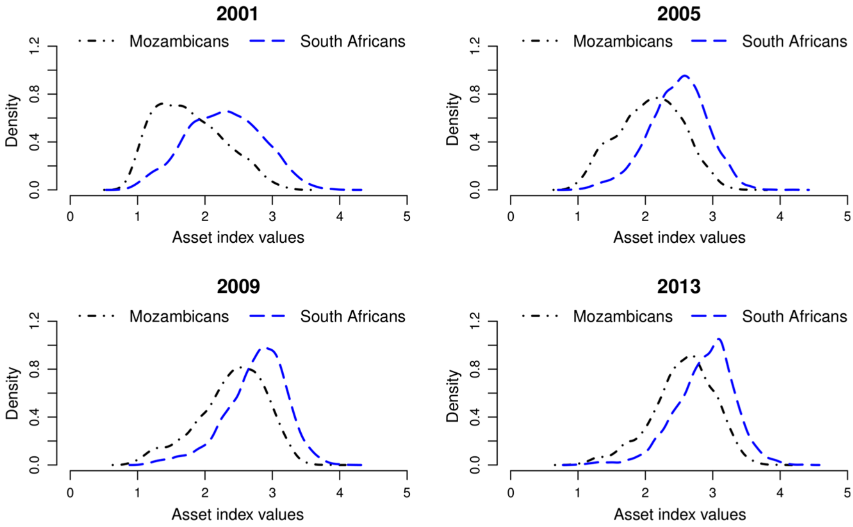


Fig. 7 Ethnic differentials in the distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013 based on absolute index

mainly due to differences in the medians of the distributions (Table 4; Fig. 8). There is little effect of differences in the shape of the distributions.

4 Discussion

Using pooled data on household assets collected every 2 years from 2001 to 2013 from households of the residents of the Agincourt HDSS, we have assessed the dynamics of SES in a typical South African rural setting. We constructed three asset indices: absolute index, PCA index and MCA index from information on ownership of household assets that include construction materials of the main dwelling, type of toilet facilities, sources of water and energy, ownership of modern assets and livestock. Thereafter, we applied the method of relative distributions to the three indices to assess temporal trends in the distribution of SES.

Table 4 Median polarization indices by ethnicity

	Lower CI	Estimate	Upper CI	*p* value
2001				
Median	0.025	0.049	0.073	<0.001
Lower	0.090	0.136	0.182	<0.001
Upper	−0.089	−0.039	0.012	0.066
2005				
Median	−0.134	−0.109	−0.085	<0.001
Lower	−0.148	−0.099	−0.051	<0.001
Upper	−0.168	−0.119	−0.071	<0.001
2009				
Median	−0.125	−0.102	−0.080	<0.001
Lower	−0.125	−0.080	−0.035	<0.001
Upper	−0.169	−0.124	−0.080	<0.001
2013				
Median	−0.072	−0.048	−0.025	<0.001
Lower	−0.073	−0.027	0.020	0.128
Upper	−0.117	−0.070	−0.023	0.002

The analysis is based on the Absolute index

Our findings indicate that the proportion of households that own assets associated with modern wealth such as stove, fridge, cellphone, car, electricity for lighting and cooking and houses constructed with modern floor, wall and roof materials has substantially increased over time. The increase has persisted beyond the time period covered in an earlier study by Sartorius et al. (2013).

On the contrary, ownership of assets associated with traditional wealth such as livestock has persistently been low. This indicates that unlike other rural populations in sub-Saharan Africa, such as a rural population in Senegal studied by Garenne (2015), traditional wealth contributes to the SES of few households in rural South Africa. This is not surprising since South Africa is a middle-income country. From a policy perspective, the general continuous increase in ownership of assets associated with modern wealth is a positive indicator of the impact of the wide-ranging reforms introduced in South Africa by the post-apartheid government that include the provision of free basic services, such as electricity (50 kWh per household per month), water, sanitation and housing to previously disadvantaged populations the majority of whom live in rural areas (Bhorat and van der Westhuizen 2013). Another important factor has been the implementation of non-contributory social grants provided by the state to vulnerable sectors of the population (Collinson 2010; Lund 2002).

Results from the relative distribution analysis in all the three indices show that the median asset index values have shifted to the right and that the distribution of SES has become less polarized and is converging towards the middle. Worth noting however is that the convergence towards the middle is larger from the upper tail than from the lower tail of the SES distribution. This might be an indication that there has been little or no improvement in the SES of the very poor segment of the population. Further analysis of the charactersitics of the individuals whose SES has persistently remained lower can assist in formulating policies that could bring further improvements in SES. The finding that the

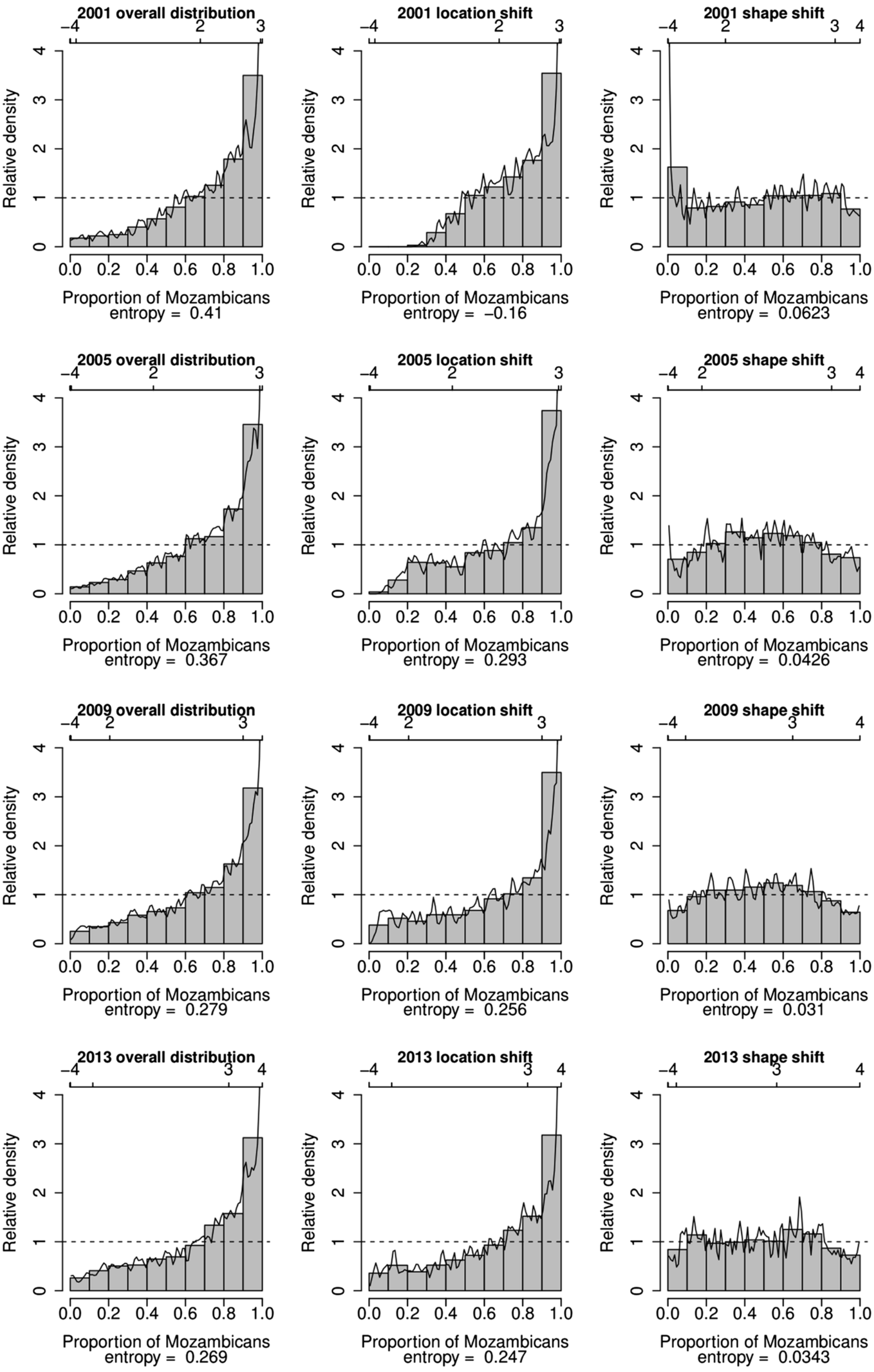


Fig. 8 Relative distributions of ethnic differentials in the distribution of SES in the villages of Agincourt HDSS, South Africa, 2001–2013 based on absolute index

SES of the Mozambican households continues to be lower compared to that of South African households suggests that members of the Mozambican households should be among the target of such policies.

From a methodological perspective, the finding that the conclusion drawn from the analysis using the easy-to-compute absolute index are similar to those from the analysis using indices constructed using more advanced statistical techniques such as PCA and MCA demonstrates the utility of the absolute index in assessing people's SES based on household assets. This finding is consistent with findings by Howe et al. (2008) and Garenne (2015) that SES indices constructed using statistically advanced methods such as PCA offer little advantage over indices constructed using simpler and more intuitive methods such as the absolute index. Since the absolute index has the added property of comparability across time without pooling the data it may be desirable in assessing temporal trends in SES.

Our study uses indices constructed from information on ownership of household assets to assess trends in SES. However, we acknowledge that our approach is by no means the only way to measure SES. Since our indices do not include other factors associated with social exclusion such as gender, education and ethnic background, they may provide only a partial view of the multi-dimensional concept of poverty, inequality and inequity. Nethertheless, our findings provide some interesting insights into the dynamics of SES in rural South Africa in recent years.

5 Conclusion

This study has shown that over the period 2001–2013 the rural population in northeast South Africa has experienced significant improvements in ownership of household assets associated with greater modern wealth and polarization of the distribution of SES has declined. However, the movement towards the middle of the SES distribution has been slower for poorer households. Methodologically, the results demonstrate that the absolute index is comparable to other indices constructed using more advanced statistical techniques in assessing people's SES based on household assets.

Acknowledgments We thank the respondents, field staff and management of the Agincourt Health and Demographic Surveillance System for their respective contributions to the production of the data used in this study.

Funding The Agincourt Health and Demographic Surveillance System has been supported with funding to the MRC/Wits Rural Public Health and Health Transitions Research Unit from the Wellcome Trust, UK (Grants 058893/Z/99/A, 069683/Z/02/Z, 085477/Z/08/Z and 085477/B/08/Z); the South African Medical Research Council, University of the Witwatersrand, and Anglo-American Chairman's Fund, South Africa; the William and Flora Hewlett Foundation (Grant 2008-1840), the Andrew W. Mellon Foundation, and the National Institute on Aging (NIA) of the National Institutes of Health (NIH), USA (Grants 1R24AG032112-01 and 5R24AG032112-03).

Authors Contributions Conceived and designed the experiments: CWK BH MC ST SJC. Analyzed the data: CWK. Contributed reagents/materials/analysis tools: MC KK ST SJC. Wrote the paper: CWK BH MC KK ST SJC.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no competing interests.

References

Asselin, L. M., & Anh, V. T. (2008). Multidimensional poverty and multiple correspondence analysis. In N. Kakwani & J. Silber (Eds.), *Quantitative approaches to multidimensional poverty measurement* (pp. 80–103). London: Palgrave Macmillan.

Ataguba, J. E., Akazili, J., & McIntyre, D. (2011). Socioeconomic-related health inequality in South Africa: Evidence from General Household Surveys. *International Journal for Equity in Health, 10*(1), 48.

Balen, J., McManus, D. P., Li, Y. S., Zhao, Z. Y., Yuan, L. P., Utzinger, J., et al. (2010). Comparison of two approaches for measuring household wealth via an asset-based index in rural and peri-urban settings of Hunan province, China. *Emerging Themes in Epidemiology, 7*, 7. doi:10.1186/1742-7622-7-7.

Barros, F. C., Victora, C. G., Scherpbier, R., & Gwatkin, D. (2010). Socioeconomic inequities in the health and nutrition of children in low/middle income countries. *Revista de Saúde Pública, 44*(1), 1–16.

Bhorat, H., & van der Westhuizen, C. (2013). Non-monetary dimensions of well-being in South Africa, 1993–2004: A post-apartheid dividend? *Development Southern Africa, 30*(3), 295–314.

Booysen, F., Van Der Berg, S., Burger, R., Maltitz, M. V., & Rand, G. D. (2008). Using an asset index to assess trends in poverty in seven Sub-Saharan African countries. *World Development, 36*(6), 1113–1130.

Collinson, M. A. (2010). Striving against adversity: The dynamics of migration, health and poverty in rural South Africa. *Global Health Action, 3*, 5080. doi:10.3402/gha.v3i0.5080.

Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R*. New York: Springer.

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effects without expenditure data or tears: An application to educational enrollments in states of India. *Demography, 38*(1), 115–132.

Garenne, M. (2015). Traditional wealth, modern goods, and demographic behavior in rural Senegal. *World Development, 72*, 267–276.

Gomez-Olive, F. X., Thorogood, M., Bocquier, P., Mee, P., Kahn, K., Berkman, L., et al. (2014). Social conditions and disability related to the mortality of older people in rural South Africa. *International Journal of Epidemiology, 43*(5), 1531–1541.

Greenacre, M. (2007). *Correspondence analysis in practice*. Boca Raton: CRC Press.

Gwatkin, D. R., Rutstein, S., Johnson, K., Suliman, E., Wagstaff, A., & Amouzou, A. (2007). *Socio-economic differences in health, nutrition, and population within developing countries*. Washington, DC: World Bank.

Handcock, M. S., & Aldrich, E. M. (2002). *Applying relative distribution methods in R*. Washington, DC: Center for Statistics and Social Sciences, University of Washington.

Handcock, M. S., & Morris, M. (1998). Relative distribution methods. *Sociological Methodology, 28*(1), 53–97.

Handcock, M. S., & Morris, M. (1999). *Relative distribution methods in the social sciences*. New York: Springer.

Hong, R., & Mishra, V. (2011). Effect of wealth inequality on chronic under-nutrition in Cambodian children. *Journal of Health, Population and Nutrition, 24*(1), 89–99.

Hosseinpoor, A. R., Van Doorslaer, E., Speybroeck, N., Naghavi, M., Mohammad, K., Majdzadeh, R., et al. (2006). Decomposing socioeconomic inequality in infant mortality in Iran. *International Journal of Epidemiology, 35*(5), 1211–1219.

Houle, B., Clark, S. J., Gómez-Olivé, F. X., Kahn, K., & Tollman, S. M. (2014). The unfolding counter-transition in rural South Africa: Mortality and cause of death, 1994–2009. *PLoS ONE, 9*(6), e100420.

Houle, B., Stein, A., Kahn, K., Madhavan, S., Collinson, M., Tollman, S. M., et al. (2013). Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status. *International Journal of Epidemiology, 42*(5), 1444–1454.

Howe, L. D., Galobardes, B., Matijasevich, A., Gordon, D., Johnston, D., Onwujekwe, O., et al. (2012). Measuring socio-economic position for epidemiological studies in low-and middle-income countries:

A methods of measurement in epidemiology paper. *International Journal of Epidemiology*, *41*, 871–886. doi:10.1093/ije/dys037.

Howe, L. D., Hargreaves, J. R., Gabrysch, S., & Huttly, S. R. (2009). Is the wealth index a proxy for consumption expenditure? A systematic review. *Journal of Epidemiology and Community Health, 63*(11), 871–877.

Howe, L. D., Hargreaves, J. R., & Huttly, S. R. (2008). Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. *Emerging Themes in Epidemiology, 5*, 3.

Kahn, K., Collinson, M. A., Gómez-Olivé, F. X., Mokoena, O., Twine, R., Mee, P., et al. (2012). Profile: Agincourt Health and Socio-demographic Surveillance System. *International Journal of Epidemiology, 41*(4), 988–1001.

Kahn, K., Tollman, S. M., Collinson, M. A., Clark, S. J., Twine, R., Clark, B. D., et al. (2007). Research into health, population and social transitions in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System. *Scandinavian Journal of Public Health, 35*(69 suppl), 8–20.

Link, B. G., & Phelan, J. (1995). Social conditions as fundamental causes of disease. *Journal of Health & Social Policy*, *35*(Extra Issue), 80–94.

Link, B. G., & Phelan, J. C. (2005). Fundamental sources of health inequalities. In D. Mechanic, L. B. Rogut, D. C. Colby & J. R. Knickman (Eds.), *Policy Challenges in Modern Health Care* (pp. 71–84). New Jersey: Rutgers University Press.

Link, B. G., Phelan, J. C., Miech, R., & Westin, E. L. (2008). The resources that matter: Fundamental social causes of health disparities and the challenge of intelligence. *Journal of Health and Social Behavior, 49*(1), 72–91.

Lund, F. (2002). Social security and the changing labour market: Access for non-standard and informal workers in South Africa. *Social Dynamics, 28*(2), 177–206.

Madhavan, S., Schatz, E., Clark, S., & Collinson, M. (2012). Child mobility, maternal status, and household composition in rural South Africa. *Demography, 49*(2), 699–718.

McKenzie, D. J. (2005). Measuring inequality with asset indicators. *Journal of Population Economics, 18*(2), 229–260.

Minujin, A., & Delamonica, E. (2004). Socio-economic inequalities in mortality and health in the developing world. *Demographic Research*. doi:10.4054/DemRes.2004.S2.13.

Montgomery, M. R., Gragnolati, M., Burke, K. A., & Paredes, E. (2000). Measuring living standards with proxy variables. *Demography, 37*(2), 155–174.

Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child Development*, *52*(1), 13–30.

Nkonki, L. L., Chopra, M., Doherty, T. M., Jackson, D., Robberstad, B., et al. (2011). Explaining household socio-economic related child health inequalities using multiple methods in three diverse settings in South Africa. *International Journal for Equity in Health, 10*(1), 13.

Phelan, J. C., Link, B. G., & Tehranifar, P. (2010). Social conditions as fundamental causes of health inequalities theory, evidence, and policy implications. *Journal of Health and Social Behavior, 51*(1 suppl), S28–S40.

Rencher, A. C. (2003). *Methods of multivariate analysis*. New Jersey: Wiley.

Sahn, D. E., & Stifel, D. (2003). Exploring alternative measures of welfare in the absence of expenditure data. *Review of Income and Wealth, 49*(4), 463–489.

Sartorius, K., Sartorius, B., Tollman, S., Schatz, E., Kirsten, J., & Collinson, M. (2013). Rural poverty dynamics and refugee communities in South Africa: A spatial-temporal model. *Population, Space and Place, 19*(1), 103–123.

Traissac, P., & Martin-Prevel, Y. (2012). Alternatives to principal components analysis to derive asset-based indices to measure socio-economic position in low-and middle-income countries: The case for multiple correspondence analysis. *International Journal of Epidemiology, 41*(4), 1207–1208.

Uthman, O. A. (2009). Decomposing socio-economic inequality in childhood malnutrition in Nigeria. *Maternal & Child Nutrition, 5*(4), 358–367.

Van de Poel, E., Hosseinpoor, A. R., Speybroeck, N., Van Ourti, T., & Vega, J. (2008). Socioeconomic inequality in malnutrition in developing countries. *Bulletin of the World Health Organization, 86*(4), 282–291.

Ward, P. (2014). Measuring the level and inequality of wealth: An application to China. *Review of Income and Wealth, 60*(4), 613–635.

Ziraba, A., Fotso, J., & Ochako, R. (2009). Overweight and obesity in urban Africa: A problem of the rich or the poor? *BMC Public Health, 9*(1), 465.

PAPER II

Kabudula et al. BMC Public Health (2017) 17:424
DOI 10.1186/s12889-017-4312-x

BMC Public Health

RESEARCH ARTICLE Open Access

Progression of the epidemiological transition in a rural South African setting: findings from population surveillance in Agincourt, 1993–2013

Chodziwadziwa W. Kabudula[1,2*], Brian Houle[1,3,4], Mark A. Collinson[1,5,6], Kathleen Kahn[1,5,6], Francesc Xavier Gómez-Olivé[1,5], Samuel J. Clark[1,4,5,7] and Stephen Tollman[1,5,6]

Abstract

Background: Virtually all low- and middle-income countries are undergoing an epidemiological transition whose progression is more varied than experienced in high-income countries. Observed changes in mortality and disease patterns reveal that the transition in most low- and middle-income countries is characterized by reversals, partial changes and the simultaneous occurrence of different types of diseases of varying magnitude. Localized characterization of this shifting burden, frequently lacking, is essential to guide decentralised health and social systems on the effective targeting of limited resources. Based on a rigorous compilation of mortality data over two decades, this paper provides a comprehensive assessment of the epidemiological transition in a rural South African population.

Methods: We estimate overall and cause-specific hazards of death as functions of sex, age and time period from mortality data from the Agincourt Health and socio-Demographic Surveillance System and conduct statistical tests of changes and differentials to assess the progression of the epidemiological transition over the period 1993–2013.

Results: From the early 1990s until 2007 the population experienced a reversal in its epidemiological transition, driven mostly by increased HIV/AIDS and TB related mortality. In recent years, the transition is following a positive trajectory as a result of declining HIV/AIDS and TB related mortality. However, in most age groups the cause of death distribution is yet to reach the levels it occupied in the early 1990s. The transition is also characterized by persistent gender differences with more rapid positive progression in females than males.

Conclusions: This typical rural South African population is experiencing a protracted epidemiological transition. The intersection and interaction of HIV/AIDS and antiretroviral treatment, non-communicable disease risk factors and complex social and behavioral changes will impact on continued progress in reducing preventable mortality and improving health across the life course. Integrated healthcare planning and program delivery is required to improve access and adherence for HIV and non-communicable disease treatment. These findings from a local, rural setting over an extended period contribute to the evidence needed to inform further refinement and advancement of epidemiological transition theory.

Keywords: South Africa, Agincourt, Rural, Mortality, HIV/Aids, Cause composition, Verbal autopsy, InterVA, Non-communicable diseases, epidemiological transition

* Correspondence: chodziwadziwa.kabudula@wits.ac.za
[1] MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[2] Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK
Full list of author information is available at the end of the article

Background

Over time, mortality and disease patterns in human populations transition from very high and fluctuating mortality concentrated at younger ages and largely caused by infectious diseases and nutritional deficiencies to relatively stable low mortality concentrated at older ages and largely caused by non-communicable diseases and injuries – the 'epidemiological transition' [1]. High-income countries experienced this transition in an orderly way along a unidirectional path during the first half of the twentieth century [1]. The first phase of the transition was characterized by high, fluctuating mortality dominated by epidemics of infectious diseases, famines and wars. Thereafter, mortality rates declined progressively and degenerative diseases started to replace infectious diseases as the major causes of morbidity and death. Finally, in later stages of the transition, non-communicable diseases such as cardiovascular diseases, diabetes and cancers, and accidents became the main causes of death, and mortality rates eventually stabilized at relatively low levels [1–3]. In low- and middle-income countries the epidemiological transition is still underway and its progress is more varied compared to the experience of high-income countries. Observed changes in mortality and disease patterns in most low- and middle-income countries including those in sub-Saharan Africa reveal transitions that are characterized by reversals, partial changes and simultaneous occurrence of different types of diseases [4–14].

For a long time in South Africa there was a steady decrease in the level of overall mortality. This trend was reversed by the HIV/AIDS epidemic that dramatically increased overall mortality from the mid-1990s to the mid-2000s [6, 15–20]. In recent years, the availability and use of antiretroviral treatment is reducing HIV/AIDS-related mortality and life expectancy is rising [18, 21, 22]. At the same time, modernization, economic and social development over the past two decades have resulted in the adoption of lifestyle practices that expose South Africans to a variety of risk factors for non-communicable diseases and injuries. Hence, the cause of death profile of South Africans increasingly includes non-communicable diseases, violence and injuries [18, 21–30].

As the epidemiological transition continues to unfold in South Africa, influenced by broader demographic, socioeconomic, technological, political, and cultural changes, there is ongoing need to quantify and characterize it and its implications in different sub-populations. This will reveal the history of the burden of disease affecting different ethnic and social groups and help identify and prioritize the interventions with potential for the greatest effect now and in the near future. This need was highlighted by the Global Burden of Disease study [31], which characterized the extent of regional heterogeneity in the trajectories of the epidemiological transition and called for greater availability and understanding of local, national, and regional data. Characterizing the shifting burden of mortality over time is critical in areas without reliable data – particularly rural settings where a greater evidence base can inform the targeting of limited resources and identify rural-urban differences and disparities.

Using mortality and cause of death data from the Agincourt Health and Socio-Demographic Surveillance System (HDSS), this article provides a comprehensive assessment of the epidemiological transition in a rural population in northeast South Africa over the period 1993–2013. This period spans major socio-political changes, the start of the HIV/AIDS epidemic and availability of antiretroviral treatment. In the article we significantly improve, update and extend measures of the trends in mortality and cause of death profiles for the Agincourt study population that have appeared earlier [6, 18, 28]. Importantly, unlike the previous work we operationalize the epidemiological transition using a statistical framework that allows us to characterize its progress relating overall mortality levels to changes in the cause composition and conduct statistical tests of changes and differentials. The longitudinal empirical evidence from this study adds a further rural South African dimension to the sparse literature on the current experience of the epidemiological transition across diverse places and contexts in low- and middle-income settings.

Methods

Data

We use mortality and cause of death data collected from 1993 to 2013 as part of annual updates of vital events conducted using the Agincourt HDSS in a population occupying 27 villages in rural northeast South Africa [32, 33]. The population is largely Shangaan (Tsonga)-speaking. Former Mozambican refugees, who arrived in the study area in the early to mid-1980s in the course and aftermath of civil war, and their descendants, make up about 30% of the population. The population has been under epidemiological and demographic surveillance since 1992 and vital events were updated at approximately 15- to 18-month intervals between 1993 and 1999, and annually since 1999.

Although the population has limited access to infrastructure and public sector services, it has experienced substantial socioeconomic changes over the years. As documented in our earlier study [34], the proportion of households that own assets associated with greater modern wealth has increased substantially over time. For example, the proportion of households with dwellings constructed with either brick or cement walls increased from 76% in 2001 to 98% in 2013; and the prevalence of tiles as roof and floor materials of dwellings increased respectively from 3% and 0.5% in 2001 to 15% and 14% in 2013. In addition, the use of electricity for lighting and cooking respectively increased from 69% and 4% of households in 2001 to 96% and 50% of households in 2013. Other notable increases are in the proportion of households owning stove, fridge, cellphone and car respectively from 41%, 40%, 37% and 14% in 2001 to 85%, 86%, 98% and 20% in 2013.

For individuals identified as having died between the annual surveillance update rounds, verbal autopsy (VA) interviews were conducted with their caregivers to elicit signs and symptoms of the illness or injury prior to their death. The interviews were conducted one to 11 months after death using a locally validated, local-language VA instrument [33, 35].

Given the rigorous processes involved in the collection, quality assurance and processing of HDSS data [14, 36], the data from the Agincourt HDSS population is one of the rare high-quality and methodologically consistent longitudinal health and demographic dataset for populations in resource-poor low- and middle-income settings. The available mortality and cause of death information by age and sex over an extended period provides a unique opportunity for assessing how populations in low- and middle-income settings, including those in rural sub-Saharan Africa are currently experiencing the epidemiological transition.

Assigning causes of death

We use the InterVA-4 probabilistic model (version 4.03) to assign probable causes of death to every death with a complete VA interview. For each death, the InterVA-4 model assigns up to three likely causes of death with associated likelihoods [37]. An indeterminate cause of death is assigned when the VA information is inadequate for the model to arrive at any cause of death. We opted for InterVA-4 as opposed to physician-coded causes of death because the InterVA-4 model assigns causes of death in a standardized, automated manner that is much quicker and more consistent than the former (particularly for assessing changes over time and across settings). Additionally, causes of death derived from InterVA-4 have been found to not substantially differ from those generated by physician coding [38].

Statistical analysis

Trends in mortality and causes of death

Similar to some earlier studies [28, 39], we use discrete-time event history analysis (DTEH) [40] to estimate overall and cause-specific annual hazards of death as functions of sex, age and time period. The *annual hazard* of dying is the probability of dying during a one-year interval starting on a particular date experienced by living individuals, conditional on their state at the beginning of the interval. An individual's continuously evolving *state* is described by the combination of values taken by both constant and time-varying variables, for this study, sex, age and time period.

One of the basic requirements of DTEH is the splitting of each individual's survival history into a set of discrete person years [40]. We create a person-year file that contains one record for each full year lived by each individual in the study population. For example, individuals who died after one year of surveillance contribute one person-year each while those who died after five years of surveillance contribute five person-years. Only *completely observed* person-years are included in the data set except when an individual dies before completing a person-year time unit. Survival histories are truncated for individuals who were alive at the beginning or end of the study and for those who migrated in/out during the study.

After constructing the person-year file we estimate the annual hazards of dying using logistic regression models [40–44]. Binary logistic regression models are used for estimates of the risk of dying from all possible causes, and multinomial logistic regression models are used to obtain estimates of the risk of dying from causes in broad cause of death categories. Using the estimated annual hazards of death, we construct standard life tables to derive life expectancies at birth and adult mortality rates (the probability of dying between ages 15 and 60 for those who survive to age 15 if subjected to age-specific mortality rates between those ages for the specified calendar year).

In order to contextualize the dynamics of the HIV epidemic and the availability of antiretroviral treatment over time, the years of the study are divided into the following time periods: 1993–1997, 1998–2000, 2001–2003, 2004–2007, 2008–2010 and 2011–2013. We also categorize age into the following commonly used age groups: 0–4, 5–14, 15–49, 50–64 and 65+. For the cause-specific analyses, the most likely causes of death generated by the InterVA-4 model except indeterminate are categorized into four broad groups: (1) HIV/AIDS and TB; (2) other communicable, maternal, perinatal, and nutritional diseases (excluding HIV/AIDS and TB); (3) non-communicable diseases; and (4) injuries, consistent with the burden of disease classification system in South Africa [23].

Changes in mortality and cause of death patterns

Following a common, standard approach to analyzing changes in mortality and cause of death patterns, we divide the most likely causes of death generated by the InterVA-4 model into three broad cause groups that can be compared with existing publications: Group I (communicable diseases, maternal, and perinatal conditions and nutritional deficiencies), Group II (non-communicable diseases), and Group III (accidents and injuries) [45, 46]. The proportion of deaths attributed to each cause group ranges from 0 to 1 and the set of proportions for all of the cause groups sums to 1 after excluding indeterminate causes. We follow Salomon and Murray [46] to relate the distribution of deaths among cause groups to the overall level of mortality. We fit estimates of age and cause-specific mortality fractions to a set of regression equations of the following form

$$Y_{i1} = \beta_0 + \beta_1 \ln(M_i) + \varepsilon_{i1}, \text{and} \tag{1}$$

$$Y_{i2} = \gamma_0 + \gamma_1 \ln(M_i) + \varepsilon_{i2}. \tag{2}$$

where i indexes age; Y_{i1} and Y_{i2} are the log ratios of the cause-specific fractions for Group II causes (P_2) and Group III causes (P_3) relative to the cause-specific fraction for Group I causes (P_1): $Y_{i1} = \ln\left(\frac{P_2}{P_1}\right)$ and $Y_{i2} = \ln\left(\frac{P_3}{P_1}\right)$; M_i is the all-cause mortality rate; β_0 and γ_0 are constant terms and ε_{i1} and ε_{i2} are error terms. The coefficients are estimated using *seemingly unrelated regression* models, separately for each sex and age group. These models provide efficient means of jointly obtaining estimates from a set of equations each with its own error term that may be correlated with the error terms of other equations. As in Salomon and Murray [46] we compute predicted values for Y_1 and Y_2 for each observation in the dataset. Those predicted values are transformed into predicted proportions for each cause group using the multivariate logistic transformation:

$$P_j = \frac{\exp(Y_j)}{1+\sum_{j=1}^{J-1} \exp(Y_j)}$$

where $J = 3$ and P_3 is $1 - P_1 - P_2$.

Software

All analyses have been conducted using Stata version 14.1 (Stata Corp., College Station, USA).

Results

Over the period 1993–2013 the Agincourt HDSS recorded a total of 13,472 deaths in 1,604,085 person-years of follow-up. Table 1 presents the person-years and number of deaths grouped by time period and cause of death categories. VA interviews were available for 92% of the deaths. VA interviews were not conducted for the other 8% of the deaths mainly due to failure to contact suitable respondents. The InterVA-4 model assigned undetermined cause of death to 6.2% of the deaths with VA interviews.

Trends in mortality and cause of death

Table 2 presents summed annual probabilities of dying from all causes for all age groups (per 1000), adult (ages 15–64) mortality rates (per 1000) and life expectancies at

Table 1 Person years and number of deaths by time period and cause of death categories, Agincourt, South Africa, 1993–2013

Sex	Indicator	1993–1997	1998–2000	2001–2003	2004–2007	2008–2010	2011–2013
Female	Person years	174,518	108,599	110,608	155,062	138,883	145,799
	Number of Deaths						
	Total	773	677	1019	1651	1229	1096
	HIV/AIDS & TB	200	226	505	825	476	286
	Other Communicable	139	109	126	219	220	237
	Non Communicable	248	203	238	391	424	452
	Injuries	47	30	38	46	29	31
	Indeterminate	52	46	44	93	48	41
	VA interview not done	87	63	68	77	32	49
Male	Person years	161,119	101,311	102,972	143,188	127,695	134,331
	Number of Deaths						
	Total	900	708	1115	1833	1363	1108
	HIV/AIDS & TB	243	256	480	755	528	300
	Other Communicable	120	97	126	236	271	221
	Non Communicable	234	158	230	420	337	352
	Injuries	130	79	119	160	102	127
	Indeterminate	44	20	40	74	49	46
	VA interview not done	129	98	120	188	76	62

Table 2 Trends in selected mortality indicators, Agincourt, South Africa, 1993–2013

Sex	Year	Annual probability of dying (95% CI)	Adult mortality rate (95% CI)	Life Expectancy at birth (95% CI)
Female	1993	4.5 (3.8,5.2)	200.3 (159.7249.6)	73.7 (70.7,76.7)
	1994	4.5 (3.8,5.2)	171.8 (134.2218.6)	73.2 (70.7,75.6)
	1995	3.8 (3.2,4.5)	127.6 (93.6172.8)	74.8 (72.3,77.2)
	1996	4.3 (3.6,5.0)	166.7 (127.6216.4)	74 (71.5,76.5)
	1997	4.5 (3.9,5.2)	222.7 (180.1273.6)	74.6 (71.7,77.4)
	1998	5.5 (4.8,6.3)	242 (202.2288.2)	70.6 (67.9,73.3)
	1999	6.1 (5.3,6.9)	296.5 (254.2344.1)	69.3 (66.6,71.9)
	2000	6.4 (5.6,7.2)	322.5 (275.3375.4)	67.7 (65.1,70.2)
	2001	8.1 (7.3,9.1)	392 (347,440.6)	62.2 (59.8,64.5)
	2002	8.7 (7.9,9.7)	469.2 (421.8519.3)	60.8 (58.3,63.2)
	2003	9.6 (8.6,10.6)	463.5 (420.2509)	59.6 (57.2,62.1)
	2004	9.4 (8.5,10.4)	503.3 (457.8550.8)	59.6 (57.2,62)
	2005	11.3 (10.3,12.4)	536.4 (493.6580.5)	56.5 (54.3,58.7)
	2006	9.9 (9.0,11.0)	476.2 (432,522.5)	59.7 (57.3,62)
	2007	10.2 (9.3,11.2)	517.9 (474.6562.7)	58.6 (56.3,60.8)
	2008	9.7 (8.8,10.6)	464.3 (422.8507.9)	60.4 (58.3,62.4)
	2009	8.0 (7.3,8.9)	378.6 (338.3421.9)	65.3 (63.2,67.4)
	2010	8.0 (7.2,8.8)	377.1 (337,420.3)	65.4 (63.3,67.5)
	2011	7.9 (7.2,8.8)	359 (321.8399.2)	65.2 (63.1,67.3)
	2012	7.2 (6.5,8)	317.2 (280.9357)	68.5 (66.4,70.5)
	2013	6.7 (6,7.4)	294.9 (258.2335.5)	70.2 (68.2,72.3)
Male	1993	5.4 (4.7,6.3)	281.8 (229.5343.3)	67.8 (64.7,70.9)
	1994	6 (5.2,6.9)	324.4 (267.8389.4)	63.6 (61.3,65.8)
	1995	5 (4.3,5.8)	343.5 (284.6410.6)	67.2 (64.7,69.8)
	1996	5.3 (4.6,6.2)	337.8 (282.7400.3)	66.9 (64,69.7)
	1997	5.4 (4.7,6.3)	331 (277.1392.2)	65.7 (63.1,68.4)
	1998	6.2 (5.4,7.1)	373 (317.6434.7)	64.4 (61.7,67)
	1999	7.1 (6.3,8.1)	358.5 (306.6416.1)	62.3 (59.8,64.8)
	2000	7 (6.2,8)	382 (328.9440.4)	62.3 (59.8,64.7)
	2001	9.1 (8.1,10.1)	499.3 (446.5554.7)	57.2 (55,59.4)
	2002	10.5 (9.5,11.7)	550.4 (499.3603)	53.6 (51.5,55.7)
	2003	11.8 (10.7,12.9)	601.3 (552.8650.3)	51.7 (49.6,53.8)
	2004	12 (10.9,13.2)	662.6 (615.9708.8)	51.3 (49.3,53.3)
	2005	12.1 (11,13.3)	588.5 (541.2636.4)	52.3 (50.4,54.1)
	2006	11.9 (10.9,13.1)	637.1 (591.5682.6)	52.1 (50.1,54)
	2007	13.2 (12.1,14.4)	650 (607.3692.4)	49.9 (48.1,51.7)
	2008	12.1 (11.1,13.2)	581.7 (539.4624.6)	52.1 (50.3,53.8)
	2009	9.8 (9,10.8)	517.4 (473.2563.2)	56.6 (54.7,58.5)
	2010	9.3 (8.5,10.3)	511.9 (466.4559.1)	57.6 (55.7,59.5)
	2011	8.6 (7.8,9.5)	445.4 (398.7495.1)	59.8 (58,61.6)
	2012	7.7 (6.9,8.5)	426.1 (380.2475)	61.5 (59.6,63.5)
	2013	7.9 (7.2,8.8)	399.7 (355.4447.3)	61.3 (59.4,63.2)

Adult mortality rate is the probability of dying between ages 15 and 59 birth. These estimates describe trends in all-cause mortality and are also shown in Fig. 1 panels (a), (b) and (c).

The annual probability of dying from all causes for all ages was about 5.4 and 4.5 per 1000 person years in 1993 for males and females, respectively. Those started to increase rapidly around 1997 for both males and females and reached a peak level around 2007 of 13.2 per 1000 person years for males and 11.3 per 1000 person years around 2005 for females, before starting to decline in more recent years. By 2013, the annual probability of dying from all causes had reduced from peak levels to 7.9 per 1000 person years for males and 6.7 per 1000 person years for females. At the peak overall mortality for both sexes had more than doubled to about 2.5 times its starting value.

Adult mortality rates exhibit a similar pattern. From a base of 281.8 and 200.3 per 1000 person years in 1993 for males and females, adult mortality rose to 650 per 1000 person years around 2007 for males and 536 per 1000 person years around 2005 for females, before starting to decline. By 2013 adult mortality rates had reduced from the peak levels to 399.7 per 1000 person years for males and 294.9 per 1000 person years for females.

Trends in life expectancy at birth reflect trends in all-age and adult mortality for both sexes. For females, life expectancy at birth dropped from about 74 years in 1993 to about 57 years in 2005 (a loss of 17 years) and returned to around 70 years by 2013. For males, life expectancy at birth dropped from about 68 years in 1993 to about 50 years in 2007 (a loss of 18 years) and increased to around 61 years by 2013.

Figure 1 panel (d) shows the predicted summed annual probability of dying per 1000 person years by year and cause of death, estimated using a multinomial logistic regression. The estimated probability of dying from each cause is added to that of the other causes. Hence, the cumulative area under each curve represents the probability of dying from the successively combined causes in a given year. The annual probability of dying from HIV/AIDS and TB increased dramatically after 2000, reached a peak around 2004–2005 and has been decreasing since 2007. However, the level of HIV/AIDS

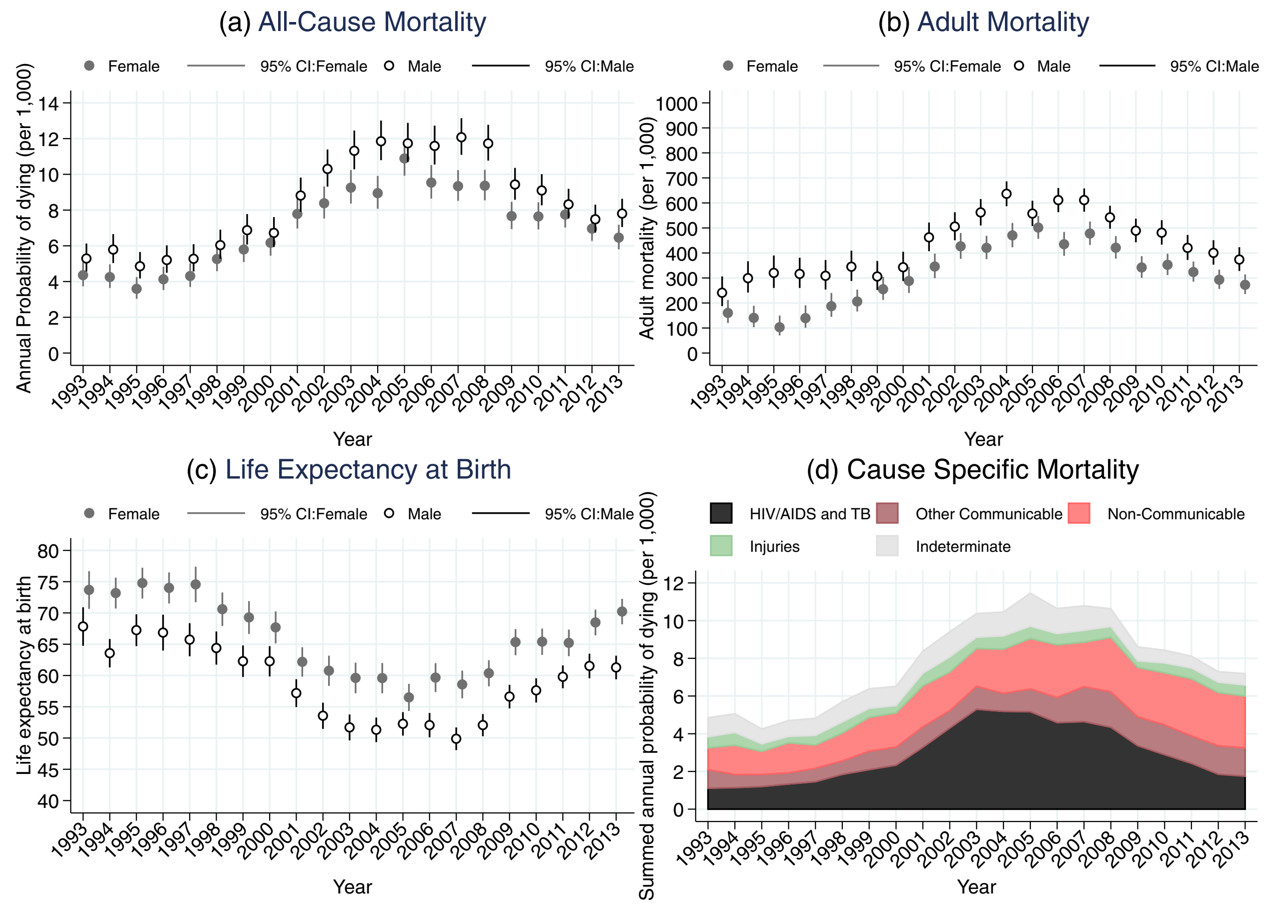


Fig. 1 Trends in selected mortality and cause of death indicators, Agincourt, South Africa, 1993–2013. **a** All-Cause Mortality. **b** Adult Mortality. **c** Life Expectancy at Birth. **d** Cause Specific Mortality

and TB related mortality in the more recent time is still higher than the level in 1993. Non-communicable diseases have consistently been the next largest cause of death and the probability of dying from them has increased steadily over time. The probability of dying from accidents and injuries has remained steady and low although there is a major difference between males and females. Figure 2 shows summaries of the same trends in the probabilities of dying from the different cause of death categories with the years of follow-up divided into six time periods.

Estimates of the risk of dying from different causes as a function of sex, age and time period are presented in Table S1 (see Additional file 1). Trends in the probabilities of dying from each of the cause of death categories by sex, age and time period, and the age-specific marginal linear predictions of dying from selected causes in subsequent time periods relative to 1993–1997 obtained from these estimates, are displayed in Figs. 3, 4 and 5.

The vertical scales for the probabilities of dying in Fig. 3 are appreciably different between the different age categories. Throughout time and for all ages, males have higher probability of dying from all causes compared to females. In all time periods, those aged 65+ have the highest probabilities of dying from all causes followed respectively in descending order by those aged 50–64, 0–4, 15–49 and 5–14.

In the 65+ age category non-communicable diseases have been the leading cause of death for both males and females and mortality associated with them has been rising steadily. Significant increases in non-communicable disease mortality started to emerge in this age category from 2004 to 2007 (RRR (95% CI): 1.51 (1.23, 1.84), p < 0.001 for males and RRR (95% CI): 1.24 (1.02, 1.50), p = 0.0271 for females in 2004–2007 relative to 1993–1997). By 2011–2013, the risk of dying from non-communicable diseases in this age category reached 1.69 (95% CI: 1.38, 2.06, p < 0.001) times as large as 1993–1997 for males and 1.65 (95% CI: 1.38, 1.99, p < 0.001) times as large as 1993–1997 for females.

In the 50–64-age category non-communicable diseases have also been an important cause of death although the

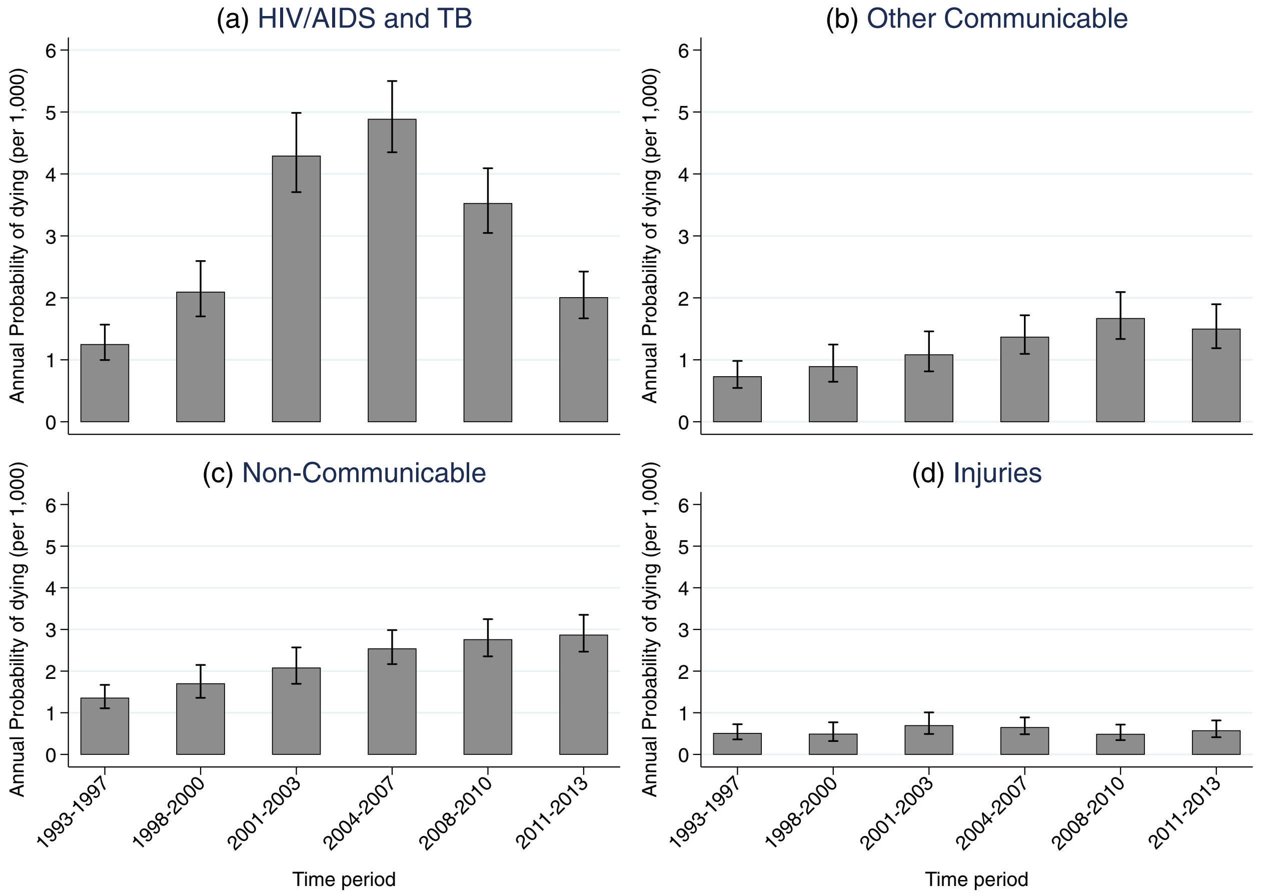


Fig. 2 Trends in annual probability of dying by cause of death, Agincourt, South Africa, 1993–2013. **a** HIV/AIDS and TB. **b** Other Communicable. **c** Non-Communicable. **d** Injuries

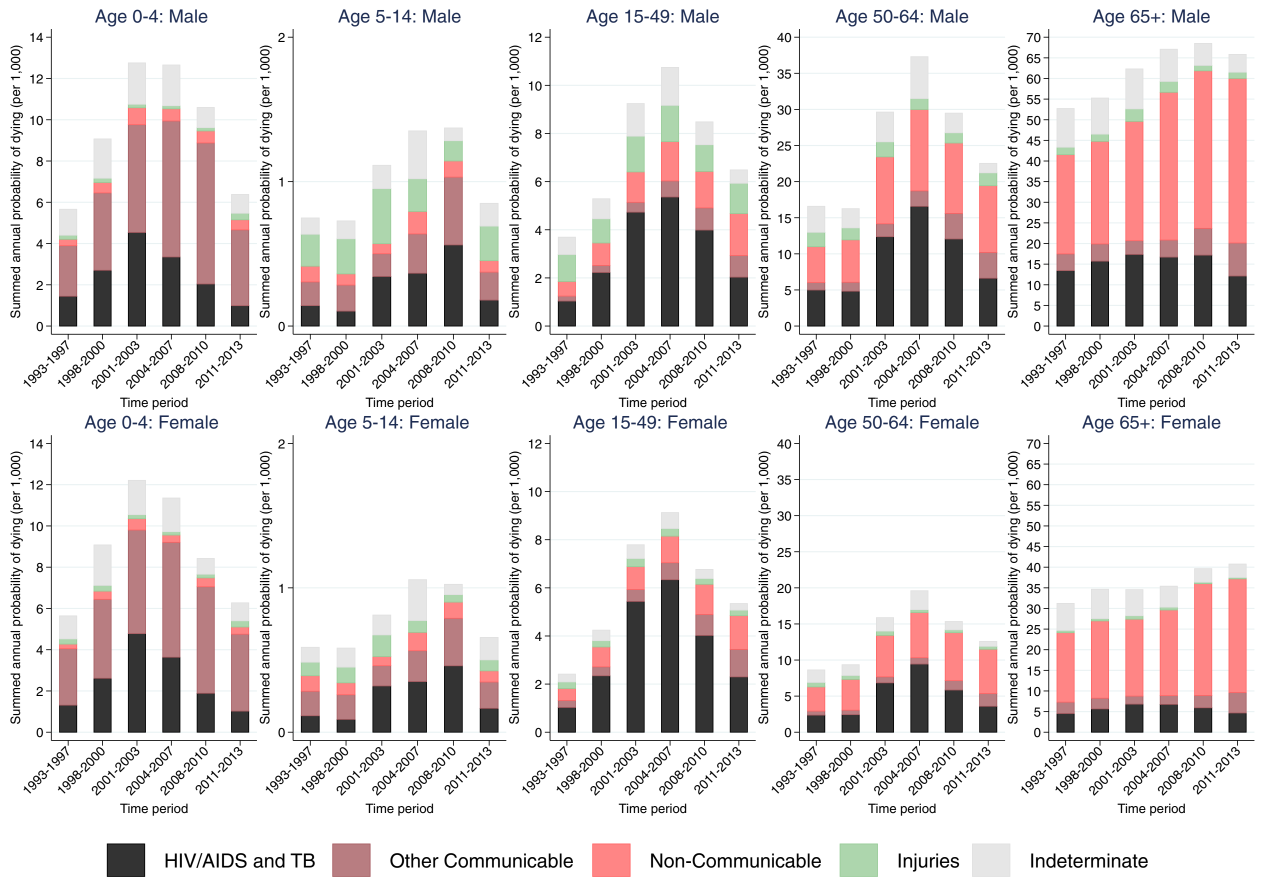


Fig. 3 Trends in annual probability of dying by age, sex and cause of death, Agincourt, South Africa, 1993–2013

risk of dying from them over time relative to 1993–1997 has remained almost constant for both males and females. The risk of dying from non-communicable diseases relative to 1993–1997 increased to statistically significant levels in 2001–2003 (RRR (95% CI): 1.87 (1.37,2.56), p < 0.001) and remained constant thereafter in males and was only statistically significant in 2001–2003 (RRR (95% CI): 1.58 (1.14,2.21), p = 0.007) in females.

HIV/AIDS and TB have also contributed significantly to mortality in the 50–64 age category for both males and females. The probability of dying from HIV/AIDS and TB in this age category steadily increased from the late 1990s to the early 2000s and reached a peak in 2004–2007 (RRR (95% CI): 3.36 (2.55, 4.43), p < 0.001 in 2004–2007 relative to 1993–1997 for males; RRR (95% CI): 4.02 (3.00, 5.37), p < 0.001 in 2004–2007 relative to 1993–1997 for females). Since then, HIV/AIDS and TB mortality has steadily declined and has reached the baseline level for males (RRR (95% CI): 1.33 (0.96,1.83), p = 0.083 in 2011–2013 relative to 1993–1997) and is closer to baseline level for females (RRR (95% CI): 1.52 (1.08–2.13), p = 0.016 in 2011–2013 relative to 1993–1997).

In the 15–49 age category for both males and females HIV/AIDS and TB have been the leading causes of death. The probability of dying from HIV/AIDS and TB in this age category increased dramatically from the late 1990s to the early 2000s, reached a peak in 2004–2007 (RRR (95% CI): **5.19** (4.25, 6.33), p < 0.001 in 2004–2007 relative to 1993–1997 for males; RRR (95% CI): **6.20** (5.12, 7.50), p < 0.001 in 2004–2007 relative to 1993–1997 for females) and has steadily decreased since then. Notwithstanding, HIV/AIDS and TB mortality in the most recent time periods is still above what it was during the early 1990s for both males (RRR (95% CI): 1.96 (1.56,2.46), p < 0.001 in 2011–2013 relative to 1993–1997) and females (RRR (95% CI): 2.24 (1.80–2.78), p < 0.001 in 2011–2013 relative to 1993–1997). Both males and females in the 15–49 age category have also experienced steady increases in the risk of dying from non-communicable diseases over the years. By 2011–2013, the risk of dying from non-communicable diseases in this age category reached 2.89 (95% CI: 2.22, 3.76, p < 0.001) times as large as 1993–1997 for males and 2.84 (95% CI: 2.17, 3.71, p < 0.001) times as large as 1993–1997 for females.

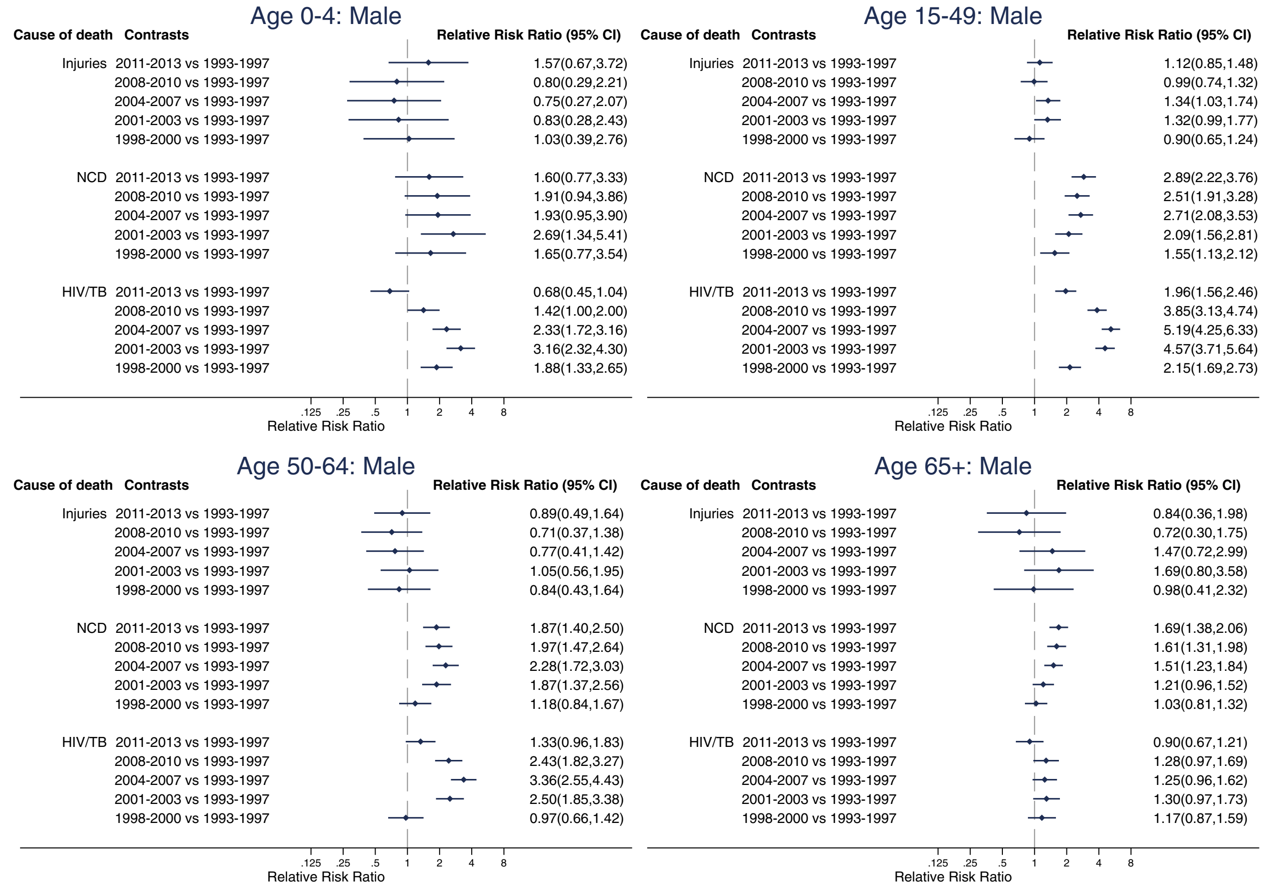


Fig. 4 Age-specific marginal linear predictions of dying from selected causes of death in subsequent time periods relative to 1993–1997 for males, Agincourt, South Africa, 1993–2013

In the 5–14 age category overall mortality increased steadily from the late 1990s, reached a peak in 2004–2007 and remained high until 2008–2010. Overall mortality in this age group only started to decline in the most recent time period following reductions in mortality from HIV/AIDS and TB. The risk of dying from HIV/AIDS and TB in this age category reached peak levels of 3.97 (95% CI: 2.07, 7.62, $p < 0.001$) times as large as 1993–1997 for males and 4.04 (95% CI: 2.09, 7.79, $p < 0.001$) times as large as 1993–1997 for females in 2008–2010 but dropped to the same level as 1993–1997 in 2011–2013 (RRR (95% CI): 1.27 (0.56, 2.86), $p = 0.565$ in 2011–2013 relative to 1993–1997 for males; RRR (95% CI): 1.45 (0.64, 3.28), $p = 0.371$ in 2011–2013 relative to 1993–1997 for females).

In the 0–4 age group mortality has almost exclusively been from HIV/AIDS and TB and other communicable diseases. Trends in the pattern of overall mortality in this age category have mirrored those of the HIV/AIDS mortality pattern. For both males and females, HIV/AIDS mortality in this age group increased steadily from the mid 1990s, reached a peak in 2001–2003 (RRR (95% CI): 3.16 (2.32, 4.30), $p < 0.001$ in 2001–2003 relative to 1993–1997 for males; RRR (95% CI): 3.66 (2.69, 5.00), $p < 0.001$ in 2001–2003 relative to 1993–1997 for females) and has steadily decreased since then. In the most recent time period the level of HIV/AIDS mortality in this age category is equivalent to the level it was during the 1993–1997 time period (RRR (95% CI): 0.68 (0.45, 1.04), $p = 0.075$ in 2011–2013 relative to 1993–1997 for males; RRR (95% CI): 0.78 (0.51, 1.19), $p = 0.249$ in 2011–2013 relative to 1993–1997 for females).

Shifts in mortality and cause of death patterns

Table 3 and Figs. 6 and 7 present results from the analysis of the progression of the epidemiological transition in the Agincourt study population for each sex, age category. Table 3 shows the estimated coefficients from the seemingly unrelated regression models. For most age groups, the coefficients for all-cause mortality are statistically significant. Figures 6 and 7 show, for males and

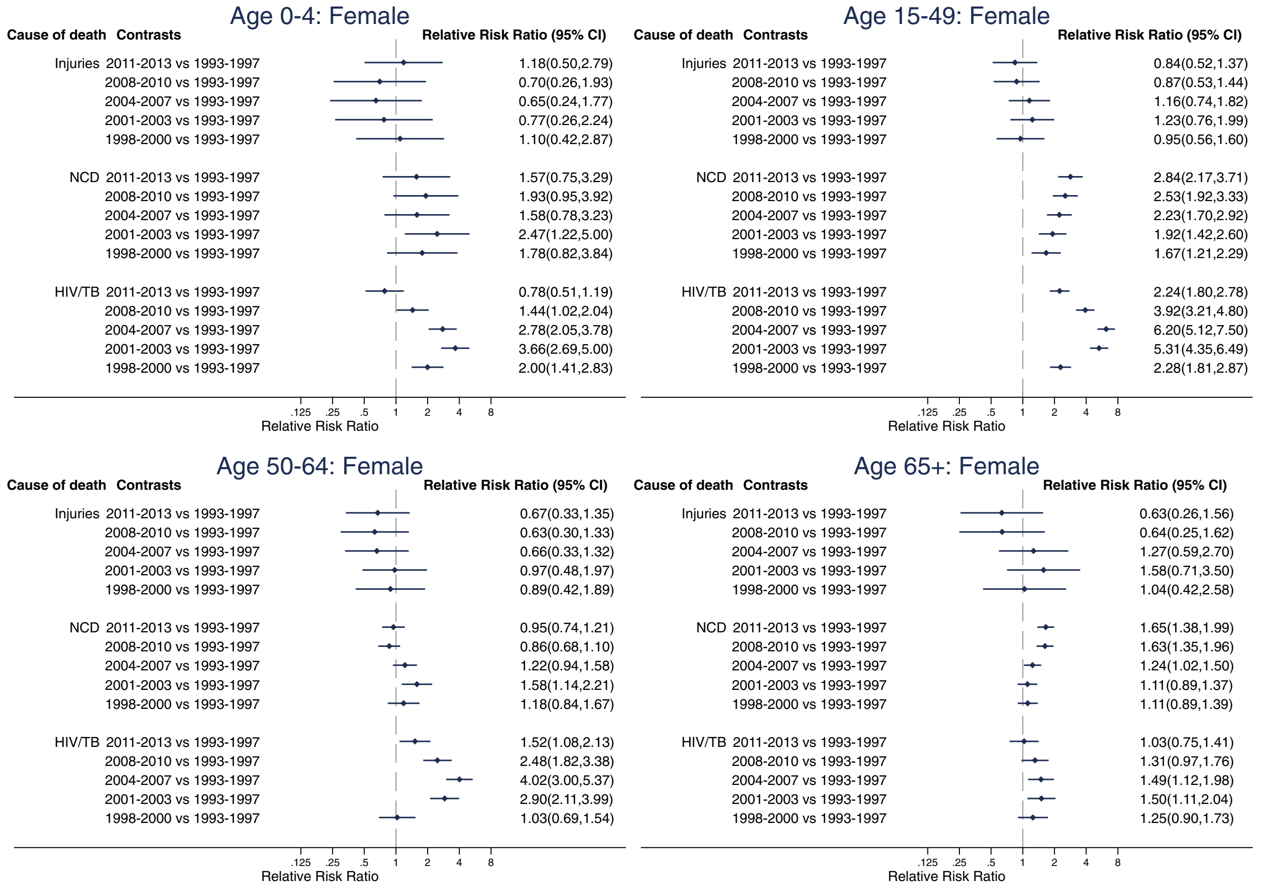


Fig. 5 Age-specific marginal linear predictions of dying from selected causes of death in subsequent time periods relative to 1993–1997 for females, Agincourt, South Africa, 1993–2013

Table 3 Seemingly unrelated regression estimates of log ratios of cause fractions on all-cause mortality

Sex	Coefficients for ln(P_2/P_1)					Coefficients for ln(P_3/P_1)			
	Age	ln(M_i) (95% CI)	*p*-value	Intercept (95% CI)	*p*-value	ln(M_i) (95% CI)	*p*-value	Intercept (95% CI)	*p*-value
Female	0–4	−0.60 (−0.91,–0.28)	< 0.001	−5.68 (−7.16,–4.21)	< 0.001	−0.51 (−1.54,0.52)	0.329	−5.77 (−10.61,–0.93)	0.019
	5–14	−0.20 (−1.97,1.58)	0.830	−3.05 (−15.71,9.62)	0.637	−1.89 (−3.78,0.01)	0.051	−15.09 (−28.62,–1.57)	0.029
	15–49	−0.49 (−1.01,0.03)	0.063	−3.92 (−6.60,–1.24)	0.004	-1.24 (−1.66,–0.82)	< 0.001	−9 (−11.17,–6.84)	< 0.001
	50–64	−0.61 (−0.97,–0.26)	0.001	−2.69 (−4.21,–1.16)	0.001	−1.15 (−1.84,–0.46)	0.001	−7.37 (−10.35,–4.39)	< 0.001
	65+	1.70 (0.50,2.91)	0.006	6.54 (2.56,10.52)	0.001	−4.06 (−6.03,–2.09)	< 0.001	−16.16 (−22.67,–9.66)	< 0.001
Male	0–4	0.01 (−0.71,0.72)	0.983	−2.56 (−5.87,0.75)	0.129	−2.9 (−3.96,–1.83)	< 0.001	−17.25 (−22.18,–12.32)	< 0.001
	5–14	−0.85 (−2.28,0.59)	0.249	−7.34 (−17.19,2.51)	0.144	−1.8 (−3.12,–0.47)	0.008	−12.98 (−22.06,–3.9)	0.005
	15–49	−0.74 (−1.08,–0.39)	< 0.001	−4.63 (−6.32,–2.94)	< 0.001	−1.13 (−1.42,–0.84)	< 0.001	−6.57 (−8.02,–5.12)	< 0.001
	50–64	−0.59 (−1.05,–0.13)	0.011	−2.50 (−4.20,–0.79)	0.004	−1.66 (−2.29,–1.03)	< 0.001	−8.01 (−10.35,–5.67)	< 0.001
	65+	0.46 (−0.36,1.28)	0.270	1.73 (−0.54,4.00)	0.136	−1.08 (−3.89,1.73)	0.451	−5.41 (−13.18,2.36)	0.172

M_i is the all-cause mortality rate for age category *i*; P_1 is cause-specific fraction for Group I causes; P_2 is cause-specific fraction for Group II causes and P_3 is cause-specific fraction for Group III causes.

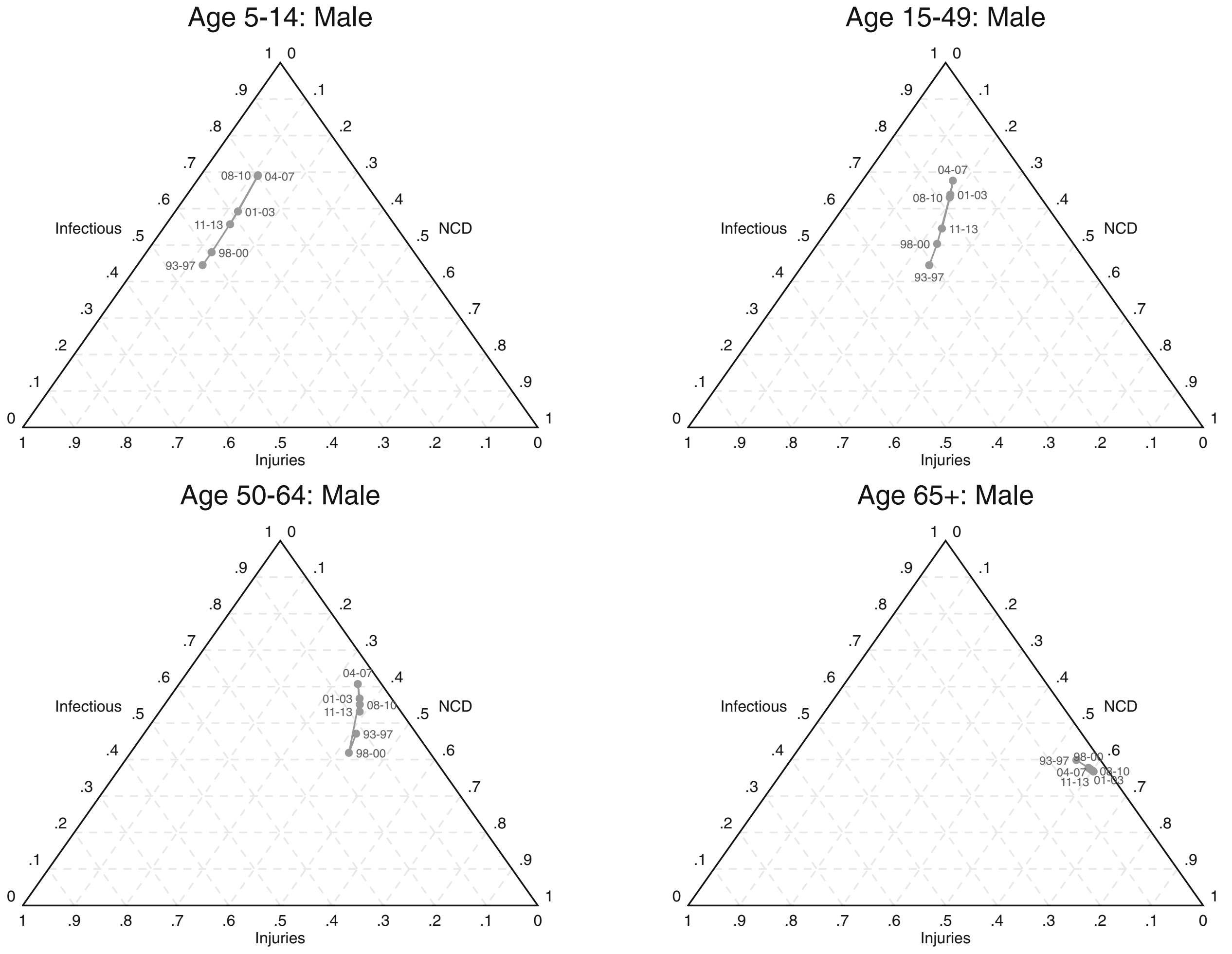


Fig. 6 Shifts in mortality and cause of death patterns for males, Agincourt, South Africa, 1993–2013

females respectively, the actual predicted proportions of deaths attributed to each cause group using ternary diagrams. A ternary diagram is a graphical representation of three variables which sum to a constant and depicts ratios of three variables as positions in an equilateral triangle. The diagrams reveal more detail on the progression of the transition for each sex and age category. The points within each diagram are labeled with an abbreviation of the time period that omits the century, for example 93–97 for 1993–1997. Each point simultaneously represents the fraction of deaths attributed to the standard Group I (communicable diseases, maternal, and perinatal conditions and nutritional deficiencies), Group II (non- communicable diseases), and Group III (accidents and injuries) cause categories. The fraction of deaths attributed to Group I causes is represented as the perpendicular distance from the bottom of the triangle to the top vertex; the fraction to Group II causes is the perpendicular distance from the left side of the triangle to the right vertex; and the fraction to Group III is the perpendicular distance from the right side of the triangle to the left vertex. For example, the point labeled 93–97 for males aged 15–49 represents 47% of deaths from Group I, 22% deaths from Group II and 31% deaths from Group III. Diagrams for the 0–4 age group are not included because there is little movement in this age group. Thus the predictions show the direction and magnitude of changes in cause of death patterns as mortality changes.

From the perspective of classical understanding of epidemiological transition, the diagrams show reversals in the progression of the transition in most age groups during the periods of rising HIV/AIDS mortality. For most age groups, the reversals peak during the 2004–2007 time period. The classic epidemiological transition is on track (conforms to standard understanding) in recent years following the widespread availability and uptake of antiretroviral treatment. However, the progression of the

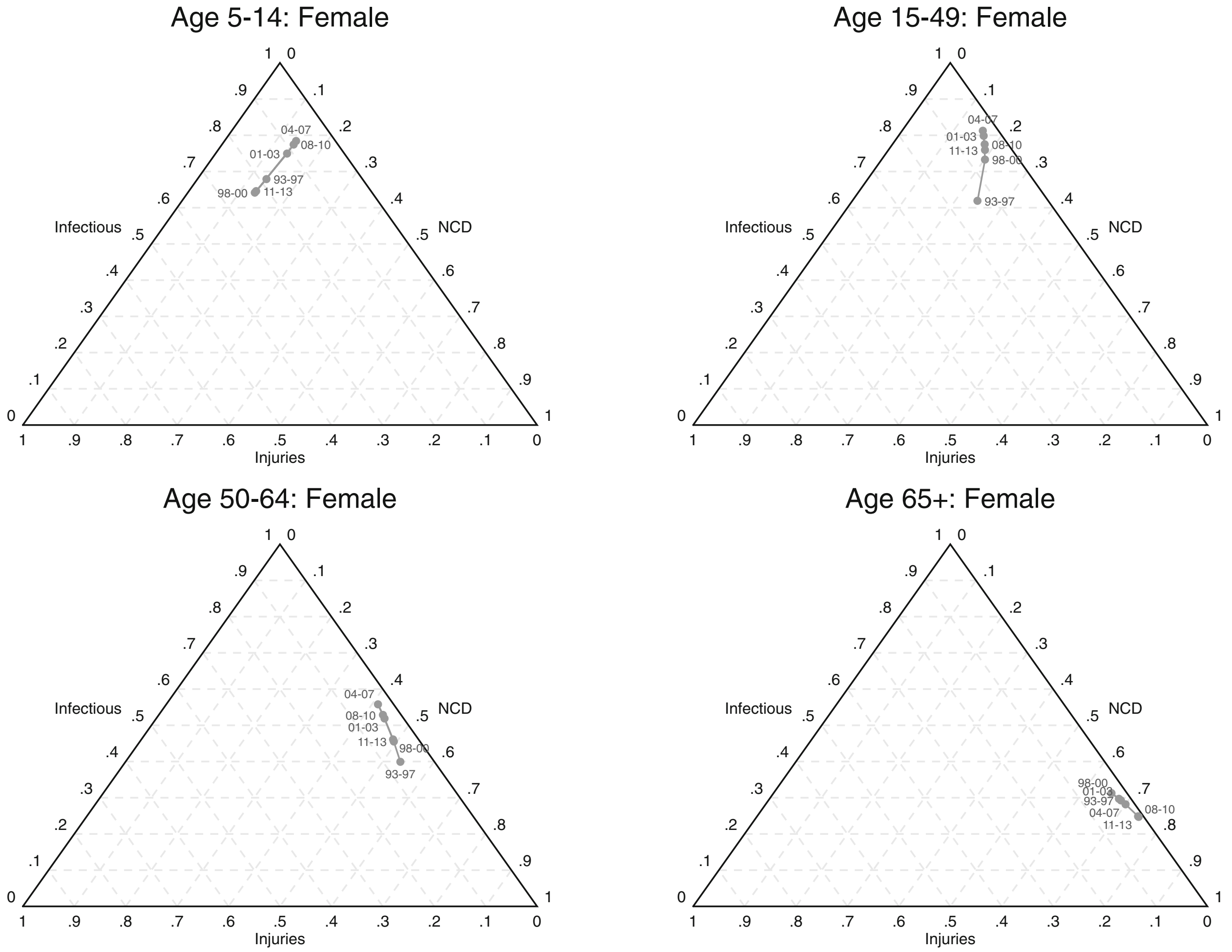


Fig. 7 Shifts in mortality and cause of death patterns for females, Agincourt, South Africa, 1993–2013

transition for most age categories has not yet quite recovered back to the stage it was in the early 1990s.

The diagrams also reveal the simultaneous impacts and interactions of different cause groups as the epidemiological transition progresses in the positive direction, i.e. overall mortality falling. The time trajectory of paths charting the progression of the transition varies by sex. For the older age groups, the relative importance of mortality from non-communicable diseases increases first for females and then for males, although non-communicable diseases become the dominant cause of death for both sexes as all-cause mortality falls.

For young and middle-aged adults (15–49 years), the relative contribution of mortality from injuries increases more for males compared to females as the transition progresses. For young and middle-aged adult females, the relative contribution of mortality from communicable causes is higher than for their male counterparts.

Discussion

This paper has assessed the progress of the epidemiological transition in a rural population in South Africa undergoing profound health and social changes, using mortality and cause of death data collected over two decades through a robust health and socio-demographic surveillance system. The findings improve, update and extend published trends in mortality and cause of death profiles [6, 18, 28] by including data from more recent years that cover the widespread availability and uptake of ART. Further, the analytical approach allows for the progress of the epidemiological transition to be empirically assessed by relating overall mortality levels to changes in the cause composition over time.

The results clearly exhibit elements of the "counter" and "protracted" epidemiological transitions proposed by Frenk [5] based on experiences in Mexico. The epidemiological transition in the Agincourt population began a reversal in the early 1990s [6, 18] that continued until around 2004–2007. This reversal was driven mostly by increases in mortality attributable to HIV/AIDS and TB. Only in recent years has the transition reversed again and started to move in the positive direction, with falling overall mortality and standard (as predicted by the classic theory of the epidemiological transition) changes to the cause of death distribution. This results from the widespread availability and uptake of antiretroviral treatment (ART) that has successfully reduced the number of deaths attributable to HIV/AIDS and TB. Provision of ART started in three district hospitals surrounding the study area between 2004 and 2005 [28, 47]. In 2007 a private community health centre specializing in HIV care and treatment services and run in partnership with the Department of Health, the Bhubezi Community Health Centre, started operating in the study area [47]. Provision of ART thereafter extended to public primary care clinics in the study area between 2008 and 2009 and has become widespread since 2010. However, despite improvements in recent years and overall mortality in children under the age of five years reaching the levels of 20 years before, largely due to the success of prevention of mother-to-child transmission (PMTCT) programmes [48], in most age groups indicators of the epidemiological transition have yet to reach the levels they occupied in the early 1990s. Thus the epidemiological transition is still evolving, having been significantly delayed by the HIV/AIDS epidemic. The progress of the transition has also been characterized by persistent gender differences with faster positive progression in females than males. Similar to other southern African settings, this may be because rates of HIV testing and linkage to and retention in care are higher in females than in males [49–52].

We acknowledge several limitations to this study. First, since updates of vital events in the Agincourt HDSS occur once a year there is a possibility that some still births, neonatal and infant deaths may not be recorded particularly when births and deaths occur between consecutive household visits [32]. However, this bias is minimal in recent years because since 2000 names of the most recent child born to each woman appear on the pre-populated household roster and since 2006 there is careful probing for pregnancies and births since the last recorded child by asking about pregnancy status of every woman of childbearing [32]. Second, we used data from one defined geographic region in rural South Africa. As such, the applicability of our findings elsewhere may not be easy to establish. However, similar to another earlier study [14], this study provides clear evidence of the major interruption to the classical epidemiological transition brought about by the HIV/AIDS epidemic. Second, while this study's goal was to characterize mortality patterns over time and empirically assess the changing relations between overall mortality levels and cause compositions, focusing on population-level patterns may mask heterogeneity in these patterns by social, economic and other indicators. Future analyses exploring heterogeneity in transition trajectories by social groups may identify important differentials and disparities as well as potential explanations of underlying patterns and drivers of epidemiological change.

Evidence suggests that the Agincourt population is undergoing dynamic socioeconomic change [34] while concurrently experiencing high prevalence of HIV [53] and risk factors for cardiometabolic diseases, particularly hypertension [54]. Our findings imply that the epidemiological transition will continue to be protracted in the near future, especially in the middle adult age categories. As more people living with HIV/AIDS access antiretroviral treatment, concentration of mortality will shift towards older age categories and the contribution of cardiovascular and other chronic non-communicable diseases will become more apparent. Further, while baseline data suggests little interaction of ART and cardiometabolic disease risk [54], greater ART uptake and resulting prolonged survival highlights the need for further studies on the interaction of HIV, cardiometabolic disease and ageing. Hence, our results suggest a need to realign the health care system to cater concurrently for multiple disease conditions.

Conclusion

This study has provided a detailed examination of the changing epidemiological profile of a rural South African population prior to and throughout the emergence of the HIV/AIDS epidemic in the absence of treatment, and the resulting changes in the context of PMTCT and ART rollout. Grounded in a robust statistical framework permitting detailed empirical assessment relating mortality levels to cause of death composition, our findings suggest that the Agincourt population is experiencing a protracted transition, with multiple stages overlapping and changes incomplete. This calls for continuous monitoring of the trajectory of the transition in order to advise policy makers around health planning and resource allocation and highlights the value of HDSS. Increasingly, the intersection and interaction of HIV and ART, non-communicable disease risk factors such as rising hypertension, obesity and type-2 diabetes and complex social, economic and behavioral changes occurring in the population (for example, rising labour migration in young women [55]) will impact continued progress in reducing premature mortality and improving health. This study highlights the need for integrated healthcare planning and program delivery to improve access and adherence to treatment for

HIV and non-communicable diseases. Finally, our findings from a local, rural setting over an extended period contribute to the evidence base to inform further refinement and advancement of health and epidemiological transition theory.

Additional file

Additional file 1: Table S1. Multinomial logistic regression of death by cause, sex, age, and time period. (DOCX 51 kb)

Acknowledgments
The respondents, field staff and management of the Agincourt Health and socio-Demographic Surveillance System are sincerely acknowledged for their respective contributions to the production of the data used in this study.

Funding
The MRC/Wits Rural Public Health and Health Transitions Research Unit which runs the Agincourt Health and socio-Demographic Surveillance System has been supported by the Wellcome Trust, UK (grants 058893/Z/99/A, 069683/Z/02/Z, 085477/Z/08/Z, and 085477/B/08/Z); the South African Medical Research Council; and University of the Witwatersrand, South Africa. PRICELESS-SA (Priority Cost-effective Lessons Systems Strengthening) provided support through a grant from the South African Medical Research Council (D1305910–01).

Availability of data and materials
Detailed documentation of the Agincourt HDSS data and an anonymized database containing data from 10% of the surveillance households are available for public access on the Agincourt HDSS website (http://www.agincourt.co.za). The Agincourt HDSS core demographic data are also routinely deposited for public access in the INDEPTH Network Data Repository (http://www.indepth-ishare.org/). The specific customized data used in this study can be made available on request to interested researchers.

Authors' contributions
CWK, MC, ST and SJC conceived the study. CWK and BH analyzed the data. KK led determination of cause of death. MC, KK, ST and SJC provided overall guidance to the conduct of the study. CWK wrote the first draft of the manuscript. All authors reviewed the manuscript and approved it for submission.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
The Agincourt HDSS, including the conduct of verbal autopsies, was reviewed and approved by the Human Research Ethics Committee (Medical) of the University of the Witwatersrand (protocols M960720 and M110138). Additional ethical clearance was also obtained from the same ethics committee (protocol M120488) for the primary analyses reported in this article. As described elsewhere [34], at the individual and household level, informed verbal consent is obtained at each surveillance visit from the head of the household (or an eligible adult in the household). The verbal informed consent process is explained by a local fieldworker who is well-trained and versed in the Agincourt HDSS methods. The fieldworker explains in the local language to the respondent the purpose, aims and justification of the HDSS as well as information about confidentiality, privacy and the right to refuse to participate or withdraw from the HDSS before conducting any interview. Interviews only occur after verbal agreement to participate from the respondent. Documentation of the consent process includes marking out the respondent on the household roster and recording the fieldworker details and date. This verbal consent process is standard across all INDEPTH (www.indepth-network.org) HDSSs, given the impossibility of contacting every person in the HDSS (as some spend part of the year away from home). The Agincourt HDSS verbal consenting processes have continued to be accepted by the aforementioned ethics committee.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [2]Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK. [3]School of Demography, The Australian National University, Canberra, Australia. [4]CU Population Center, Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, USA. [5]INDEPTH Network, Accra, Ghana. [6]Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. [7]Department of Sociology, The Ohio State University, Columbus, OH, USA.

Received: 2 November 2016 Accepted: 26 April 2017
Published online: 10 May 2017

References

1. Omran AR. The epidemiologic transition: a theory of the epidemiology of population change. The Milbank Memorial Fund Quarterly. 1971;49(4):509–38.
2. Rogers RG, Hackenberg R. Extending epidemiologic transition theory: a new stage. Soc Biol. 1987;34(3–4):234.
3. Olshansky SJ, Ault AB. The fourth stage of the epidemiologic transition: the age of delayed degenerative diseases. The Milbank Memorial Fund Quarterly. 1986;64:355–91.
4. Caselli G, Mesle F, Vallin J. Epidemiologic transition theory exceptions. Genus. 2002;58(1):9–51.
5. Frenk J, Bobadilla JL, Sepúlveda J, Cervantes ML. Health transition in middle-income countries: new challenges for health care. Health Policy Plan. 1989;4(1):29.
6. Kahn K, Garenne ML, Collinson MA, Tollman SM. Mortality trends in a new South Africa: hard to make a fresh start. Scand J Public Health. 2007;35(69 suppl):26.
7. Moser K, Shkolnikov V, Leon DA. World mortality 1950-2000: divergence replaces convergence from the late 1980s. Bull World Health Organ. 2005;83(3):202–9.
8. Omran A, R. The epidemiologic transition theory revisited thirty years later. World Health Statistics Quarterly. 1998;51(2/3/4):99–119.
9. Masquelier B, Waltisperger D, Ralijaona O, Pison G, Ravélo A. The epidemiological transition in Antananarivo, Madagascar: an assessment based on death registers (1900-2012). Glob Health Action. 2014;7
10. Agyei-Mensah S, Aikins AG. Epidemiological transition and the double burden of disease in Accra. Ghana J Urban Health. 2010;87(5):879–97.
11. Bawah A, Houle B, Alam N, Razzaque A, Streatfield PK, Debpuur C, et al. The evolving demographic and health transition in four low-and middle-income countries: evidence from four sites in the INDEPTH network of longitudinal health and demographic surveillance systems. PLoS One. 2016;11(6):e0157281.
12. Engelaer FM, Koopman JJ, van Bodegom D, Eriksson UK, Westendorp RG. Determinants of epidemiologic transition in rural Africa: the role of socioeconomic status and drinking water source. Trans R Soc Trop Med Hyg. 2014;108(6):372–9.
13. Santosa A, Wall S, Fottrell E, Högberg U, Byass P. The development and experience of epidemiological transition theory over four decades: a systematic review. Glob Health Action. 2014;7
14. Santosa A, Byass P. Diverse empirical evidence on epidemiological transition in low-and middle-income countries: population-based findings from INDEPTH network data. PLoS One. 2016;11(5):e0155753.
15. Karim SSA, Churchyard GJ, Karim QA, Lawn SD. HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. Lancet. 2009;374(9693):921–33.
16. United Nations DoE, Social Affairs PD. World Population Prospects: The 2010 Revision, Volume I: Comprehensive Tables. ST/ESA/SER.A/313.

17. Bradshaw D, Laubscher R, Dorrington R, Bourne DE, Timaeus IM. Unabated rise in number of adult deaths in South Africa. SAMJ. 2004;94(4):278–9.
18. Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. Glob Health Action. 2014;7
19. Zwang J, Garenne M, Kahn K, Collinson M, Tollman SM. Trends in mortality from pulmonary tuberculosis and HIV/AIDS co-infection in rural South Africa (Agincourt). Trans R Soc Trop Med Hyg. 2007;101(9):893–8.
20. Tollman SM, Kahna K, Garenne M, Gear JS. Reversal in mortality trends: evidence from the Agincourt field site, South Africa, 1992-1995. AIDS. 1999;13(9):1091–7.
21. Herbst AJ, Cooke GS, Bärnighausen T, KanyKany A, Tanser F, Newell ML. Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal. South Africa Bull World Health Organ. 2009;87(10):754–62.
22. Herbst AJ, Mafojane T, Newell ML, others. Verbal autopsy-based cause-specific mortality trends in rural KwaZulu-Natal, South Africa, 2000–2009. Popul Health Metr. 2011;9(1):47.
23. Bradshaw D, Groenewald P, Laubscher R, Nannan N, Nojilana B, Norman R, et al. Initial burden of disease estimates for South Africa, 2000. SAMJ. 2003;93(9):682–8.
24. Bradshaw D, Nannan N, Groenewald P, Joubert J, Laubscher R, Nijilana B, et al. Provincial mortality in South Africa, 2000-priority-setting for now and benchmark for the future. SAMJ. 2005;95(7):496–503.
25. Groenewald P, Bradshaw D, Daniels J, Zinyakatira N, Matzopoulos R, Bourne D, et al. Local-level mortality surveillance in resource-limited settings: a case study of cape town highlights disparities in health. Bull World Health Organ. 2010;88(6):444–51.
26. Hosegood V, Vanneste AM, Timæus IM. Levels and causes of adult mortality in rural South Africa: the impact of AIDS. AIDS. 2004;18(4):663.
27. Tollman SM, Kahn K, Sartorius B, Collinson MA, Clark SJ, Garenne ML. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. Lancet. 2008;372(9642):893–901.
28. Houle B, Clark SJ, Gómez-Olivé FX, Kahn K, Tollman SM. The unfolding counter-transition in rural South Africa: mortality and cause of death, 1994–2009. PLoS One 2014;9(6):e100420.
29. Bradshaw D, Schneider M, Dorrington R, Bourne DE, Laubscher R. South African cause-of-death profile in transition-1996 and future trends. SAMJ. 2002;92:618–23.
30. Kahn K, Tollman SM, Garenne M, Gear JSS. Who dies from what? Determining cause of death in South Africa's rural north-east. Tropical Med Int Health. 1999;4(6):433–41.
31. Naghavi M, Wang H, Lozano R, Davis A, Liang X, Zhou M, et al. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global burden of disease study 2013. Lancet. 2015;385(9963):117–71.
32. Kahn K, Collinson MA, Gómez-Olivé FX, Mokoena O, Twine R, Mee P, et al. Profile: Agincourt health and socio-demographic surveillance system. Int J Epidemiol. 2012;41(4):988–1001.
33. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, et al. Research into health, population and social transitions in rural South Africa: data and methods of the Agincourt health and demographic surveillance system. Scand J Public Health. 2007;35(69 suppl):8–20.
34. Kabudula CW, Houle B, Collinson MA, Kahn K, Tollman S, Clark S. Assessing Changes in Household Socioeconomic Status in Rural South Africa, 2001–2013: A distributional analysis using household asset indicators. Soc Indic Res1–27.
35. Kahn K, Tollman SM, Garenne M, Gear JSS. Validation and application of verbal autopsies in a rural area of South Africa. Tropical Med Int Health. 2000;5(11):824–31.
36. Kahn K. Dying to make a fresh start: mortality and health transition in a new South Africa: Umea University; 2006.
37. Byass P, Chandramohan D, Clark SJ, D'Ambruoso L, Fottrell E, Graham WJ, et al. Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. Glob Health Action. 2012;5
38. Byass P, Kahn K, Fottrell E, Mee P, Collinson MA, Tollman SM. Using verbal autopsy to track epidemic dynamics: the case of HIV-related mortality in South Africa. Popul Health Metr. 2011;9:46.
39. Clark SJ, Collinson MA, Kahn K, Drullinger K, Tollman SM. Returning home to die: circular labour migration and mortality in South Africa 1. Scand J Public Health. 2007;35(Suppl 69):35–44.
40. Allison PD. Event history analysis: regression for longitudinal event data. Thousand Oaks, CA: Sage; 1984.
41. Allison PD. Discrete-time methods for the analysis of event histories. Sociol Methodol. 1982;13(1):61–98.
42. Allison PD. Survival analysis using SAS: a practical guide: SAS Institute; 2010.
43. Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. J Amer Stat Assoc. 1988;83(402):414–25.
44. Van Hook J, Altman CE. Using discrete-time event history fertility models to simulate total fertility rates and other fertility measures. Popul Res Policy Rev. 2013;32(4):585–610.
45. Lopez AD, Mathers CD, Ezzati M, Jamison DT, Murray CJL. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. Lancet. 2006;367(9524):1747–57.
46. Salomon JA, Murray CJL. The epidemiologic transition revisited: compositional models for causes of death by age and sex. Popul Develop Rev. 2002;28(2):205–28.
47. Mee P, Collinson MA, Madhavan S, Root ED, Tollman SM, Byass P, et al. Evidence for localised HIV related micro-epidemics associated with the decentralised provision of antiretroviral treatment in rural South Africa: a spatio-temporal analysis of changing mortality patterns (2007-2010). J Global Health. 2014;4(1):010403.
48. Byass P, Kabudula CW, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. A successful failure: missing the MDG4 target for under-five mortality in South Africa. PLoS Med. 2015;12(12):e1001926.
49. Bärnighausen T, Tanser F, Herbst K, Mutevedzi T, Mossong J, Newell M-L. Structural barriers to antiretroviral treatment: a study using population-based CD4 cell count and linked antiretroviral treatment programme data. Lancet. 2013;382:S5.
50. Bassett IV, Regan S, Luthuli P, Mbonambi H, Bearnot B, Pendleton A, et al. Linkage to care following community-based mobile HIV testing compared with clinic-based testing in Umlazi township, Durban. South Africa HIV Med. 2014;15(6):367–72.
51. Lessells RJ, Mutevedzi PC, Cooke GS, Newell M-L. Retention in HIV care for individuals not yet eligible for antiretroviral therapy: rural KwaZulu-Natal. South Africa JAIDS. 2011;56(3):e79.
52. Mugglin C, Estill J, Wandeler G, Bender N, Egger M, Gsponer T, et al. Loss to programme between HIV diagnosis and initiation of antiretroviral therapy in sub-Saharan Africa: systematic review and meta-analysis. Tropical Med Int Health. 2012;17(12):1509–20.
53. Gómez-Olivé FX, Angotti N, Houle B, Klipstein-Grobusch K, Kabudula C, Menken J, et al. Prevalence of HIV among those 15 and older in rural South Africa. AIDS Care. 2013;25(9):112–8.
54. Clark SJ, Gómez-Olivé FX, Houle B, Thorogood M, Klipstein-Grobusch K, Angotti N, et al. Cardiometabolic disease risk and HIV status in rural South Africa: establishing a baseline. BMC Public Health. 2015;15(1):1.
55. Collinson MA, White MJ, Bocquier P, McGarvey ST, Afolabi SA, Clark SJ, et al. Migration and the epidemiological transition: insights from the Agincourt sub-district of northeast South Africa. Glob Health Action. 2014;7

Additional file 1: Table 1. Multinomial logistic regression of death by cause, sex, age, and time period

Variable	Relative Risk Ratio	95% CI	p-value
HIV/TB			
Sex			
Female	1.00	–	–
Male	1.10	[0.86,1.40]	0.442
Age Groups			
0-4	1.00	–	–
5-14	0.09	[0.05,0.16]	< 0.001
15-49	0.78	[0.59,1.05]	0.102
50-64	1.81	[1.28,2.56]	< 0.001
65+	3.47	[2.51,4.81]	< 0.001
Time Period			
1993-1997	1.00	–	–
1998–2000	2.00	[1.41,2.83]	< 0.001
2001–2003	3.66	[2.69,5.00]	< 0.001
2004–2007	2.78	[2.05,3.78]	< 0.001
2008–2010	1.44	[1.02,2.04]	0.040
2011–2013	0.78	[0.51,1.19]	0.249
Interactions between Sex and Age			
Male ***X*** age 5–14	1.13	[0.75,1.71]	0.562
Male ***X*** age 15–49	0.92	[0.76,1.11]	0.372
Male ***X*** age 50–64	1.92	[1.53,2.41]	< 0.001
Male ***X*** age 65+	2.72	[2.15,3.44]	< 0.001
Interactions between Sex and Time			
Male ***X*** 1998–2000	0.94	[0.72,1.22]	0.644
Male ***X*** 2001–2003	0.86	[0.68,1.08]	0.202
Male ***X*** 2004–2007	0.84	[0.67,1.04]	0.107
Male ***X*** 2008–2010	0.98	[0.78,1.24]	0.877
Male ***X*** 2011–2013	0.88	[0.68,1.13]	0.298
Interactions between Age and Time			
Age 5-14 ***X*** 1998–2000	0.39	[0.14,1.08]	0.069
Age 5-14 ***X*** 2001–2003	0.77	[0.36,1.63]	0.493
Age 5-14 ***X*** 2004–2007	1.11	[0.54,2.27]	0.778
Age 5-14 ***X*** 2008–2010	2.80	[1.36,5.78]	0.005
Age 5-14 ***X*** 2011–2013	1.86	[0.76,4.55]	0.177
Age 15-49 ***X*** 1998–2000	1.14	[0.79,1.66]	0.485
Age 15-49 ***X*** 2001–2003	1.45	[1.04,2.02]	0.029
Age 15-49 ***X*** 2004–2007	2.23	[1.61,3.10]	< 0.001
Age 15-49 ***X*** 2008–2010	2.72	[1.88,3.94]	< 0.001
Age 15-49 ***X*** 2011–2013	2.87	[1.85,4.45]	< 0.001

Age 50-64 ***X*** 1998–2000	0.51	[0.32,0.84]	0.008
Age 50-64 ***X*** 2001–2003	0.79	[0.53,1.19]	0.260
Age 50-64 ***X*** 2004–2007	1.45	[0.98,2.13]	0.062
Age 50-64 ***X*** 2008–2010	1.72	[1.12,2.65]	0.014
Age 50-64 ***X*** 2011–2013	1.94	[1.18,3.21]	0.010
Age 65+ ***X*** 1998–2000	0.62	[0.41,0.96]	0.032
Age 65+ ***X*** 2001–2003	0.41	[0.28,0.61]	< 0.001
Age 65+ ***X*** 2004–2007	0.54	[0.37,0.78]	0.001
Age 65+ ***X*** 2008–2010	0.91	[0.60,1.38]	0.649
Age 65+ ***X*** 2011–2013	1.32	[0.81,2.15]	0.269
Other Communicable Causes			
Sex			
Female	1.00	–	–
Male	0.90	[0.69,1.16]	0.408
Age Groups			
0-4	1.00	–	–
5-14	0.06	[0.04,0.11]	< 0.001
15-49	0.11	[0.07,0.15]	< 0.001
50-64	0.21	[0.12,0.36]	< 0.001
65+	1.01	[0.70,1.44]	0.975
Time Period			
1993-1997	1.00	–	–
1998–2000	1.40	[1.03,1.91]	0.033
2001–2003	1.84	[1.37,2.46]	< 0.001
2004–2007	2.04	[1.57,2.65]	< 0.001
2008–2010	1.89	[1.45,2.47]	< 0.001
2011–2013	1.36	[1.03,1.81]	0.032
Interactions between Sex and Age			
Male ***X*** age 5–14	1.08	[0.70,1.65]	0.734
Male ***X*** age 15–49	0.80	[0.64,0.99]	0.038
Male ***X*** age 50–64	2.05	[1.46,2.89]	< 0.001
Male ***X*** age 65+	1.64	[1.25,2.17]	< 0.001
Interactions between Sex and Time			
Male ***X*** 1998–2000	1.09	[0.75,1.59]	0.638
Male ***X*** 2001–2003	1.16	[0.82,1.65]	0.402
Male ***X*** 2004–2007	1.32	[0.97,1.80]	0.078
Male ***X*** 2008–2010	1.48	[1.09,2.01]	0.013
Male ***X*** 2011–2013	1.10	[0.80,1.51]	0.551
Interactions between Age and Time			
Age 5-14 ***X*** 1998–2000	0.72	[0.32,1.60]	0.418
Age 5-14 ***X*** 2001–2003	0.45	[0.19,1.05]	0.063
Age 5-14 ***X*** 2004–2007	0.62	[0.31,1.23]	0.168
Age 5-14 ***X*** 2008–2010	1.02	[0.53,1.97]	0.944

Age 5-14 ***X*** 2011–2013	0.79	[0.36,1.71]	0.547
Age 15-49 ***X*** 1998–2000	0.92	[0.55,1.53]	0.738
Age 15-49 ***X*** 2001–2003	0.93	[0.58,1.48]	0.747
Age 15-49 ***X*** 2004–2007	1.19	[0.79,1.79]	0.409
Age 15-49 ***X*** 2008–2010	1.59	[1.05,2.38]	0.027
Age 15-49 ***X*** 2011–2013	2.88	[1.90,4.36]	< 0.001
Age 50-64 ***X*** 1998–2000	0.78	[0.36,1.73]	0.546
Age 50-64 ***X*** 2001–2003	0.81	[0.40,1.65]	0.562
Age 50-64 ***X*** 2004–2007	0.77	[0.40,1.47]	0.429
Age 50-64 ***X*** 2008–2010	1.23	[0.66,2.28]	0.517
Age 50-64 ***X*** 2011–2013	2.31	[1.26,4.26]	0.007
Age 65+ ***X*** 1998–2000	0.67	[0.40,1.14]	0.144
Age 65+ ***X*** 2001–2003	0.39	[0.22,0.67]	0.001
Age 65+ ***X*** 2004–2007	0.38	[0.24,0.61]	< 0.001
Age 65+ ***X*** 2008–2010	0.57	[0.36,0.90]	0.017
Age 65+ ***X*** 2011–2013	1.32	[0.85,2.06]	0.212
Non-communicable Causes			
Sex			
Female	1.00	–	–
Male	1.37	[0.88,2.12]	0.159
Age Groups			
0-4	1.00	–	–
5-14	0.48	[0.20,1.16]	0.104
15-49	2.22	[1.20,4.11]	0.011
50-64	15.05	[8.12,27.90]	< 0.001
65+	76.85	[42.68,138.36]	< 0.001
Time Period			
1993-1997	1.00	–	–
1998–2000	1.78	[0.82,3.84]	0.144
2001–2003	2.47	[1.22,5.00]	0.012
2004–2007	1.58	[0.78,3.23]	0.205
2008–2010	1.93	[0.95,3.92]	0.071
2011–2013	1.57	[0.75,3.29]	0.234
Interactions between Sex and Age			
Male ***X*** age 5–14	0.74	[0.36,1.52]	0.406
Male ***X*** age 15–49	0.89	[0.58,1.36]	0.590
Male ***X*** age 50–64	1.09	[0.71,1.67]	0.707
Male ***X*** age 65+	1.05	[0.69,1.59]	0.818
Interactions between Sex and Time			
Male ***X*** 1998–2000	0.93	[0.71,1.22]	0.592
Male ***X*** 2001–2003	1.09	[0.84,1.41]	0.513
Male ***X*** 2004–2007	1.22	[0.97,1.53]	0.093
Male ***X*** 2008–2010	0.99	[0.79,1.25]	0.934

Male ***X*** 2011–2013	1.02	[0.81,1.28]	0.872
Interactions between Age and Time			
Age 5-14 ***X*** 1998–2000	0.43	[0.12,1.58]	0.205
Age 5-14 ***X*** 2001–2003	0.24	[0.06,0.91]	0.035
Age 5-14 ***X*** 2004–2007	0.74	[0.25,2.20]	0.590
Age 5-14 ***X*** 2008–2010	0.54	[0.17,1.75]	0.304
Age 5-14 ***X*** 2011–2013	0.46	[0.13,1.65]	0.233
Age 15-49 ***X*** 1998–2000	0.94	[0.42,2.10]	0.877
Age 15-49 ***X*** 2001–2003	0.78	[0.37,1.63]	0.507
Age 15-49 ***X*** 2004–2007	1.40	[0.67,2.94]	0.368
Age 15-49 ***X*** 2008–2010	1.31	[0.63,2.76]	0.469
Age 15-49 ***X*** 2011–2013	1.81	[0.84,3.90]	0.130
Age 50-64 ***X*** 1998–2000	0.72	[0.32,1.63]	0.427
Age 50-64 ***X*** 2001–2003	0.70	[0.33,1.47]	0.344
Age 50-64 ***X*** 2004–2007	1.18	[0.56,2.49]	0.659
Age 50-64 ***X*** 2008–2010	1.03	[0.49,2.19]	0.929
Age 50-64 ***X*** 2011–2013	1.17	[0.54,2.54]	0.691
Age 65+ ***X*** 1998–2000	0.63	[0.29,1.37]	0.240
Age 65+ ***X*** 2001–2003	0.45	[0.22,0.92]	0.028
Age 65+ ***X*** 2004–2007	0.78	[0.38,1.60]	0.499
Age 65+ ***X*** 2008–2010	0.85	[0.41,1.73]	0.646
Age 65+ ***X*** 2011–2013	1.05	[0.50,2.22]	0.888
Injuries			
Sex	1.00	–	–
Female	0.80	[0.42,1.51]	0.486
Male			
Age Groups			
0-4	1.00	–	–
5-14	0.37	[0.15,0.89]	0.026
15-49	1.09	[0.55,2.14]	0.804
50-64	2.46	[1.11,5.44]	0.026
65+	2.15	[0.89,5.16]	0.087
Time Period			
1993-1997	1.00	–	–
1998–2000	1.1	[0.42,2.87]	0.851
2001–2003	0.77	[0.26,2.24]	0.630
2004–2007	0.65	[0.24,1.77]	0.395
2008–2010	0.7	[0.26,1.93]	0.493
2011–2013	1.18	[0.50,2.79]	0.699
Interactions between Sex and Age			
Male ***X*** age 5–14	2.98	[1.37,6.47]	0.006
Male ***X*** age 15–49	5.13	[2.79,9.43]	< 0.001
Male ***X*** age 50–64	4.03	[1.97,8.21]	< 0.001

Male ***X*** age 65+	4.17	[1.97,8.83]	< 0.001
Interactions between Sex and Time			
Male ***X*** 1998–2000	0.94	[0.55,1.63]	0.835
Male ***X*** 2001–2003	1.07	[0.65,1.78]	0.780
Male ***X*** 2004–2007	1.16	[0.72,1.87]	0.539
Male ***X*** 2008–2010	1.13	[0.67,1.93]	0.642
Male ***X*** 2011–2013	1.33	[0.80,2.21]	0.273
Interactions between Age and Time			
Age 5-14 ***X*** 1998–2000	1.06	[0.31,3.58]	0.923
Age 5-14 ***X*** 2001–2003	2.09	[0.59,7.35]	0.252
Age 5-14 ***X*** 2004–2007	1.36	[0.40,4.71]	0.623
Age 5-14 ***X*** 2008–2010	0.80	[0.20,3.11]	0.742
Age 5-14 ***X*** 2011–2013	0.69	[0.22,2.16]	0.519
Age 15-49 ***X*** 1998–2000	0.87	[0.32,2.35]	0.780
Age 15-49 ***X*** 2001–2003	1.60	[0.54,4.77]	0.397
Age 15-49 ***X*** 2004–2007	1.79	[0.64,4.97]	0.266
Age 15-49 ***X*** 2008–2010	1.24	[0.44,3.49]	0.677
Age 15-49 ***X*** 2011–2013	0.71	[0.30,1.71]	0.446
Age 50-64 ***X*** 1998–2000	0.81	[0.26,2.55]	0.717
Age 50-64 ***X*** 2001–2003	1.27	[0.38,4.26]	0.704
Age 50-64 ***X*** 2004–2007	1.02	[0.32,3.25]	0.974
Age 50-64 ***X*** 2008–2010	0.90	[0.28,2.91]	0.854
Age 50-64 ***X*** 2011–2013	0.57	[0.21,1.57]	0.276
Age 65+ ***X*** 1998–2000	0.95	[0.27,3.34]	0.935
Age 65+ ***X*** 2001–2003	2.05	[0.57,7.34]	0.271
Age 65+ ***X*** 2004–2007	1.96	[0.59,6.52]	0.274
Age 65+ ***X*** 2008–2010	0.91	[0.24,3.37]	0.882
Age 65+ ***X*** 2011–2013	0.53	[0.16,1.73]	0.295
Indeterminate			
Sex			
Female	1.00	–	–
Male	1.13	[0.83,1.53]	0.434
Age Groups			
0-4	1.00	–	–
5-14	0.09	[0.05,0.19]	< 0.001
15-49	0.29	[0.20,0.42]	< 0.001
50-64	1.56	[1.03,2.35]	0.036
65+	5.89	[4.17,8.34]	< 0.001
Time Period			
1993-1997	1.00	–	–
1998–2000	1.76	[1.18,2.63]	0.006
2001–2003	1.49	[0.99,2.25]	0.057
2004–2007	1.46	[1.00,2.13]	0.048

2008–2010	0.68	[0.43,1.10]	0.116
2011–2013	0.78	[0.49,1.25]	0.307
Interactions between Sex and Age			
Male ***X*** age 5–14	0.97	[0.57,1.65]	0.909
Male ***X*** age 15–49	1.97	[1.48,2.61]	< 0.001
Male ***X*** age 50–64	1.84	[1.31,2.60]	< 0.001
Male ***X*** age 65+	1.27	[0.95,1.71]	0.111
Interactions between Sex and Time			
Male ***X*** 1998–2000	0.86	[0.61,1.22]	0.392
Male ***X*** 2001–2003	1.07	[0.77,1.49]	0.685
Male ***X*** 2004–2007	1.07	[0.80,1.44]	0.649
Male ***X*** 2008–2010	1.13	[0.79,1.62]	0.494
Male ***X*** 2011–2013	0.93	[0.63,1.36]	0.693
Interactions between Age and Time			
Age 5-14 ***X*** 1998–2000	0.72	[0.27,1.93]	0.520
Age 5-14 ***X*** 2001–2003	0.89	[0.34,2.32]	0.818
Age 5-14 ***X*** 2004–2007	1.87	[0.85,4.13]	0.120
Age 5-14 ***X*** 2008–2010	1.01	[0.32,3.16]	0.987
Age 5-14 ***X*** 2011–2013	1.93	[0.74,5.02]	0.179
Age 15-49 ***X*** 1998–2000	0.76	[0.47,1.24]	0.273
Age 15-49 ***X*** 2001–2003	1.18	[0.74,1.88]	0.487
Age 15-49 ***X*** 2004–2007	1.40	[0.91,2.15]	0.126
Age 15-49 ***X*** 2008–2010	1.70	[1.01,2.86]	0.046
Age 15-49 ***X*** 2011–2013	1.07	[0.62,1.82]	0.815
Age 50-64 ***X*** 1998–2000	0.49	[0.27,0.88]	0.017
Age 50-64 ***X*** 2001–2003	0.72	[0.41,1.25]	0.238
Age 50-64 ***X*** 2004–2007	1.03	[0.63,1.67]	0.912
Age 50-64 ***X*** 2008–2010	0.97	[0.53,1.79]	0.927
Age 50-64 ***X*** 2011–2013	0.50	[0.25,0.98]	0.045
Age 65+ ***X*** 1998–2000	0.62	[0.38,0.99]	0.044
Age 65+ ***X*** 2001–2003	0.65	[0.40,1.04]	0.075
Age 65+ ***X*** 2004–2007	0.53	[0.34,0.83]	0.005
Age 65+ ***X*** 2008–2010	0.73	[0.42,1.27]	0.266
Age 65+ ***X*** 2011–2013	0.63	[0.36,1.10]	0.104

PAPER III

Socioeconomic differences in mortality in the antiretroviral therapy era in Agincourt, rural South Africa, 2001–13: a population surveillance analysis

Chodziwadziwa W Kabudula, Brian Houle, Mark A Collinson, Kathleen Kahn, Francesc Xavier Gómez-Olivé, Stephen Tollman, Samuel J Clark

Summary

Background Understanding the effects of socioeconomic disparities in health outcomes is important to implement specific preventive actions. We assessed socioeconomic disparities in mortality indicators in a rural South African population over the period 2001–13.

Methods We used data from 21 villages of the Agincourt Health and socio-Demographic Surveillance System (HDSS). We calculated the probabilities of death from birth to age 5 years and from age 15 to 60 years, life expectancy at birth, and cause-specific and age-specific mortality by sex (not in children <5 years), time period, and socioeconomic status (household wealth) quintile for HIV/AIDS and tuberculosis, other communicable diseases (excluding HIV/AIDS and tuberculosis) and maternal, perinatal, and nutritional causes, non-communicable diseases, and injury. We also quantified differences with relative risk ratios and relative and slope indices of inequality.

Findings Between 2001 and 2013, 10 414 deaths were registered over 1 058 538 person-years of follow-up, meaning the overall crude mortality was 9·8 deaths per 1000 person-years. We found significant socioecomonic status gradients for mortality and life expectancy at birth, with outcomes improving with increasing socioeconomic status. An inverse relation was seen for HIV/AIDS and tuberculosis mortality and socioeconomic status that persisted from 2001 to 2013. Deaths from non-communicable diseases increased over time in both sexes, and injury was an important cause of death in men and boys. Neither of these causes of death, however, showed consistent significant associations with household socioeconomic status.

Interpretation The poorest people in the population continue to bear a high burden of HIV/AIDS and tuberculosis mortality, despite free antiretroviral therapy being made available from public health facilities. Associations between socioeconomic status and increasing burden of mortality from non-communicable diseases is likely to become prominent. Integrated strategies are needed to improve access to and uptake of HIV testing, care, and treatment, and management of non-communicable diseases in the poorest populations.

Funding Wellcome Trust, South African Medical Research Council, and University of the Witwatersrand, South Africa.

Lancet Glob Health 2017; 5: e924–35

This online publication has been corrected. The corrected version first appeared at thelancet.com/lancetgh on September 11, 2017

See Comment page e851

MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa (C W Kabudula MSc, B Houle PhD, M A Collinson PhD, Prof K Kahn PhD, F X Gómez-Olivé PhD, Prof S Tollman PhD, Prof S J Clark, PhD)**; School of Demography, Australian National University, Canberra, ACT, Australia** (B Houle)**; CU Population Center, Institute of Behavioral Science, University of Colorado at Boulder, Boulder, CO, USA** (B Houle, Prof S J Clark)**; INDEPTH Network, Accra, Ghana** (M A Collinson, Prof K Kahn, F X Gómez-Olivé, Prof S Tollman, Prof S J Clark)**; Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden** (Prof K Kahn, Prof S Tollman)**; Department of Science and Technology/ Medical Research Council, South African Population Research Infrastructure Network (SAPRIN), Acornhoek, South Africa** (M A Collinson)**; and Department of Sociology, The Ohio State University, Columbus, OH, USA** (Prof S J Clark)

Correspondence to:
Mr Chodziwadziwa W Kabudula, MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of the Witwatersrand, PO Box 2, Acornhoek 1360, South Africa
chodziwadziwa.kabudula@wits.ac.za

Introduction

The distribution of health outcomes varies by social factors, such as marital status, ethnic origin, and socioeconomic status.[1,2] For example, a review by Link and Phelan[3] showed that socioeconomic status has a positive association with life expectancy and a negative association with overall, infant, and perinatal mortality. McKinnon and colleagues[4] also reported a negative association between household socioeconomic status and neonatal mortality in many low-income and middle-income countries, through use of data from demographic and health surveys done between 1997 and 2012. Social disparities in population health outcomes are sustained because social conditions, such as knowledge, money, power, prestige, and beneficial social connections, allow individuals to avoid health-related risks, adopt protective strategies, and access medical facilities and services.[3,5–7] Understanding the magnitude of social disparities in health outcomes is important to implement specific actions to reduce them. In many sub-Saharan African settings, however, evidence of socioeconomic differences in health is limited because of the requirement for complex information systems, longitudinal studies with sufficiently large samples, and detailed information on health outcomes and social characteristics.

Over the past two decades, complex and rapidly evolving health transitions have occurred in South Africa. Most important has been the steady and substantial increase in overall mortality due to communicable diseases from the mid-1990s to the mid-2000s, peaking around 2005–07 owing to the HIV/AIDS epidemic.[8–14] After the widespread introduction of free antiretroviral therapy (ART) available from public health facilities, AIDS-related mortality declined.[11,15,16] At the same time, however, modernisation and changes in social and economic development (eg, increases in the proportion of households that owned wealth-associated assets, such as stoves, fridges, and televisions[17]) have resulted in the adoption of lifestyle

Research in context

Evidence before this study

We searched PubMed and Google Scholar for studies on mortality and associated differences in socioeconomic status in South Africa, using the search terms “mortality”, “death”, “socioeconomic”, “wealth”, and “South Africa” without any language or date restrictions. Several studies showed that the emergence of the HIV/AIDS epidemic substantially increased overall mortality and the contribution of communicable diseases to the overall mortality burden and reduction in life expectancy from the mid-1990s to the mid-2000s. Later studies have shown that HIV/AIDS-related mortality has been declining since antiretroviral therapy (ART) became widely available through public health services, but limited information was available on the distribution of mortality by socioeconomic status, particularly in resource-poor rural areas.

Added value of this study

Our evidence describes the distribution of mortality in a resource-poor rural area of northeast South Africa by household socioeconomic status before and after free ART became available. HIV/AIDS-related mortality reduced and life expectancy at birth improved, but individuals from the poorest households continue to bear the greatest burden of HIV/AIDS and tuberculosis mortality. Additionally, the mortality burden from non-communicable diseases is rising, and associations with household socioeconomic status are likely to become prominent. These findings might reflect the situation in other resource-poor rural settings with high HIV/AIDS disease burdens and increasing risk of non-communicable diseases in South Africa and southern Africa.

Implications of all the available evidence

Integrated health-care planning and programme delivery strategies are needed to increase access to and uptake of HIV testing, linkage to care and ART, and prevention and treatment of non-communicable diseases among the poorest individuals in resource-poor settings with high burdens of HIV/AIDS and rising burdens of non-communicable disease risk factors. The aim should be to reduce socioeconomic inequalities in mortality where disease burden is high, and to achieve further reductions in overall mortality.

For more on **Agincourt HDSS data** see http://www.agincourt.co.za

For more on **INDEPTH Network Data Repository** see http://www.indepth-ishare.org/

practices that expose South Africans to risk factors for non-communicable diseases and injury. The mortality profile in South Africa over the past two decades has been dominated by communicable diseases, maternal, perinatal, and nutritional causes, non-communicable diseases, and injury.[11,15,16,18–27] Information on how mortality patterns are changing in relation to socioeconomic status, however, has been limited, particularly in rural areas. We used a high-quality and methodologically consistent longitudinal dataset that provides detailed information on health outcomes by indicators of socioeconomic status to assess changes in mortality in a poor rural South African population over the period 2001–13.

Methods

Setting and data sources

We used data from the ongoing Agincourt Health and socio-Demographic Surveillance System (HDSS), which was established in 1992.[28,29] Agincourt is located in a resource-poor rural setting in Bushbuckridge Municipality in northeast South Africa, close to the Mozambique border. The Agincourt HDSS has generated detailed longitudinal data on births, deaths, and migration and complementary data covering health and socioeconomic indicators. The study area included 21 villages spread over 402 km^2 until 2006,[22] and was extended to 26 villages in 2007 and to 31 villages from 2010 to 2012.[17] Most people speak Shangaan. About a third of the population is made up of immigrants from Mozambique, who arrived in the area in the early to mid-1980s, and their descendants. Data have been collected annually since 1999. Detailed documentation describing the Agincourt HDSS data and an anonymised database containing data from 10% of the surveillance households are available for public access. The Agincourt HDSS core demographic data are also routinely deposited for public access in the INDEPTH Network Data Repository. In this study we have used only data from the original 21 villages to maximise the duration of follow-up at the village level. These customised data are available on request to interested researchers.

Ethics approval was obtained from the Human Research Ethics Committee (Medical) of the University of the Witwatersrand, Johannesburg, South Africa, for surveillance activities in the Agincourt HDSS (protocols M960720 and M110138) and for the analyses reported in this study (protocol M120488). Informed verbal consent was obtained at every surveillance visit from the head of the household or another eligible adult in the household. The person giving consent was noted in the household roster, and the details and date of the process were recorded by the responsible fieldworker.

Causes of death

For every death recorded from 2001 to 2013, we used the InterVA-4 probabilistic model (version 4.03) to assign the most probable cause, rather than the more traditional, clinically oriented underlying cause. This model enables a standardised, automated assignment of cause of death that is much quicker and more consistent than physician assessment, and is particularly useful for assessing changes over time and across settings. It assigns each death to a maximum of three likely causes, with associated likelihoods based on information about signs and symptoms of illness or injury collected through verbal autopsy interviews.[30] In the annual surveillance updates of the Agincourt HDSS, caregivers of individuals who had died since the previous visit were interviewed with a questionnaire in Shangaan that had been locally validated.[29,31] Thus, timing of the interviews ranged from 1 to 11 months after death. The cause of death was categorised as indeterminate when inadequate information was obtained for the model to assign a cause of death. The causes of death generated by the InterVA-4 model are based on the WHO 2012 verbal autopsy standards and correspond to the International Classification of Diseases, tenth edition.[30]

We categorised the most probable causes of death into five broad groups: HIV/AIDS and tuberculosis; other communicable diseases (excluding HIV/AIDS and tuberculosis) and maternal, perinatal, and nutritional causes; non-communicable diseases; injury; and indeterminate. The first four categories are consistent with the burden of disease classification system used in South Africa.[27] We combined HIV/AIDS and tuberculosis because HIV is an underlying cause in most tuberculosis deaths and distinguishing those that are HIV related from those that are not is difficult with the verbal autopsy method.[23]

Socioeconomic status

We measured socioeconomic status with an absolute household wealth index computed from a list of household asset indicators that were grouped in the following categories: construction materials in the main dwelling; type of toilet facilities and sources of water; sources of energy; ownership of modern assets; and livestock.[17,23,32] For each household, after categorisation, asset indicators were assigned weights, with higher values corresponding to higher socioeconomic status. The value assigned to each item was divided by the highest value for all households to obtain normalised values that fell in the range of 0–1. The normalised values within each category were summed to obtain category-specific values, normalised by the same method, then summed to produce an overall household wealth index value that fell in the range 0–5. Once constructed, the wealth index was divided into household wealth quintiles, in which the first quintile represented the poorest households and the fifth the richest households. Data on household asset indicators were collected in 2001, 2003, 2005, 2007, 2009, 2011, and 2013.

Statistical analysis

For each individual we organised data into a person-year file that contained one record for each full year lived, similar to the methods of Houle and colleagues[23,33,34] and Kabudula and colleagues.[26] We included only records for completely observed person-years plus the year in which the individual died irrespective of whether the person-year was complete. Covariates recorded were sex, age (<5, 5–14, 15–49, 50–64, or ≥65 years), time period (2001–03, 2004–07, 2008–10, and 2011–13), date of death, likely cause of death, and household wealth quintile. For covariates that change over time, such as age and household wealth quintile, we used the value at the beginning of the relevant person-year. For completed person-years the death indicator was set to 0, and it was set to 1 in records where there was a death during the year. Time periods were split across years to contextualise the dynamics of the HIV/AIDS epidemic and the roll-out of services for prevention of mother-to-child transmission and ART.

We used the person-year file to calculate the probabilities of death from birth to age 5 years and from age 15 to 60 years, life expectancy at birth, and age-specific and cause-specific mortality by sex (excluding children <5 years), time period, and household wealth quintile. Thereafter, we estimated relative and absolute socioeconomic differences in the mortality indicators with the relative index of inequality (RII) and the slope index of inequality (SII), respectively (appendix).[35] These measures take into account the whole socioeconomic distribution and the effects on mortality indicators of a person moving from the lowest to the highest quintile.[35,36] RII=1 and SII=0 imply no difference between the lower and higher ends of the socioeconomic continuum. RII values greater than 1 and positive SII values imply greater mortality at the lower end of the continuum, and RII values less than 1 and negative SII values imply greater mortality at the higher end. We fitted separate models for each time period and sex (except for children <5 years) to calculate RIIs and SIIs for mortality in children and adults and life expectancy at birth, with the modified ridit score (appendix)[37,38] as the independent variable. To calculate RIIs and SIIs for cause-specific mortality, we fitted separate models for each cause-of-death category, time period, and sex, with the modified ridit score and age group as independent variables. We also fitted models with two-way interaction terms between the modified ridit score and time period to assess trends in socioeconomic differences in the mortality indicators over time.

See Online for appendix

We also calculated relative risk ratios and 95% CIs to investigate associations between relative inequalities and household wealth quintile, which we obtained from multinomial logistic regression models,[39–43] with cause of death as an indicator of mortality used as the dependent variable and household wealth quintiles, sex, age group, and time period as independent variables.

Although socioeconomic status can be measured at the individual level with factors such as education and occupation,[44] samples are necessarily restricted to people who have reached a certain age to make the indicators meaningful (eg, age beyond which individuals are unlikely to advance their eduction further or working age). Instead, we used unadjusted household socioeconomic

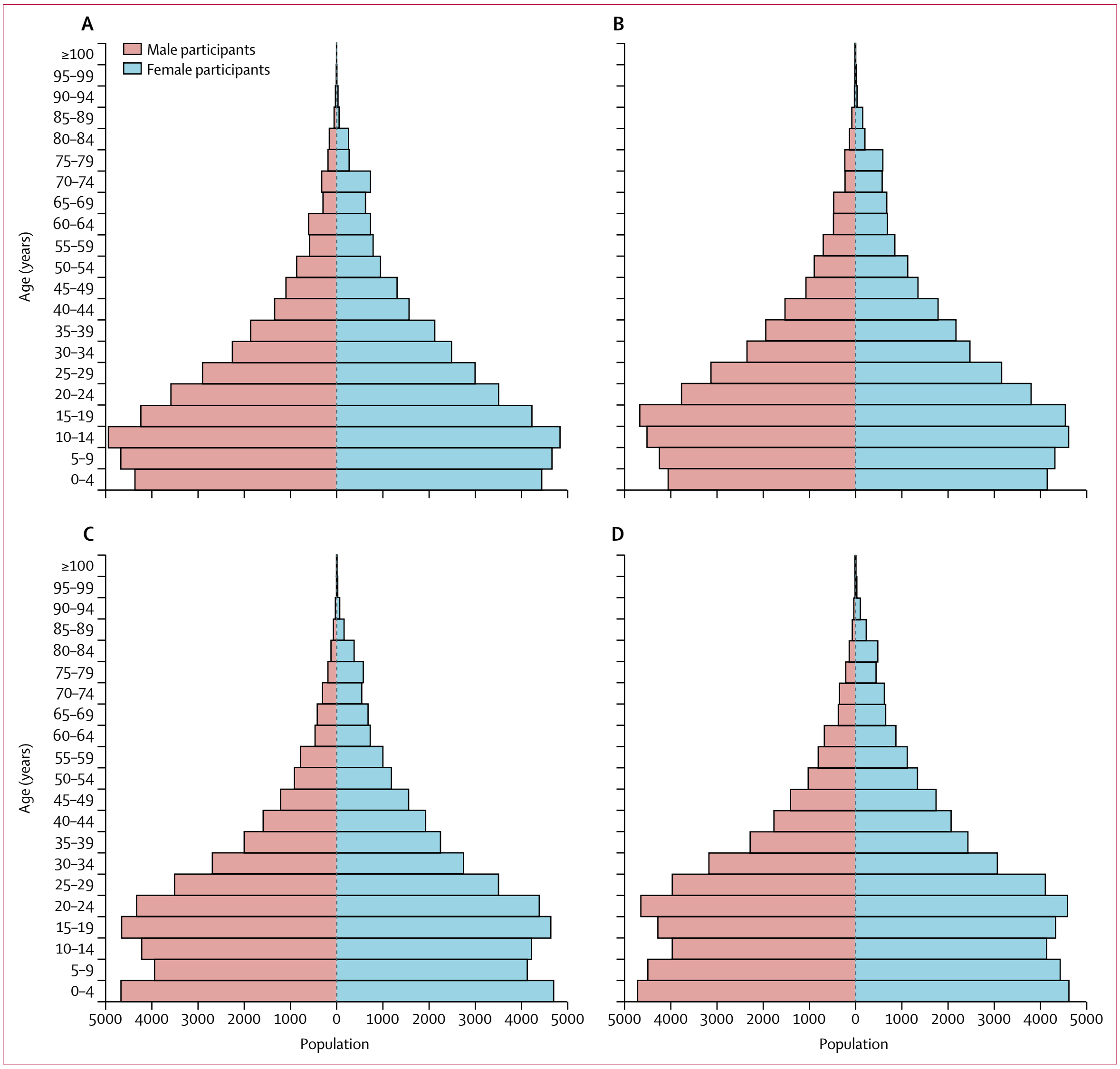


Figure 1: **Age distribution in the original 21 villages of the Agincourt Health and socio-Demographic Surveillance System**
(A) Population, July 1, 2001. (B) Population, July 1, 2005. (C) Population, July 1, 2009. (D) Population, July 1, 2013.

status to maximise the sample size because these data are collected more frequently than individual data and because all individuals in the household are affected by the household environment. Household socioeconomic status provides a good cumulative indicator of material living standards,[44,45] which strongly affect individual household members.

Data on household asset indicators used for calculating the household wealth index were collected in alternate years from 2001 onwards and, therefore, we used multiple imputation to minimise the loss of data due to missing values. We used partial mean matching (based on the nearest two neighbours) to generate five imputed datasets and derive parameter estimates and SEs by averaging across the imputations and adjusting for variance. As done by Houle and colleagues,[33] the imputations are generated from a household-year data set that includes counts of men, boys, women, and girls, Mozambicans and South Africans, individuals aged younger than 20 years, 20–59 years, and 60 years and older, and 1–2-year lags of household wealth index.

We did all analyses with Stata version 14.1. Estimates with p values less than 0·05 were taken to be significant.

Role of the funding source

The funders had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Results

In 2001, the population assessed by the Agincourt HDSS in the original 21 villages was 70 809 people in 11 818 households. The population had increased to 71 830 people in 12 302 households by 2005, 75 603 people in 13 460 households by 2009, and 79 912 in 14 692 households in 2013. The age structure of the population in selected years is shown in figure 1 and table 1. Changes in the distribution of education attainment and key asset indicators are shown in table 1. Between 2001 and 2013, 10 414 deaths were recorded in 1 058 538 person-years of follow-up in 2001–13 (table 2). Information from verbal autopsy interviews was available for 93·5% of these deaths, of which the InterVA-4 model classified 435 (4·5%) as indeterminate. The verbal autopsy information for the 672 (6·5%) remaining deaths was mainly missing due to inability to contact suitable respondents.

Overall mortality reduced steadily over the study period among children younger than 5 years and increased in adults from 2001 to 2007 but reduced steadily thereafter. Overall life expectancy at birth decreased from 2001 to 2007 but then increased until the end of the study (figure 2, appendix). Adult mortality and life expectancy from birth were consistently better for women than men (figure 2, appendix).

Except at the height of the HIV/AIDS epidemic in 2004–07 and in the early years after ART introduction in 2008–10, we found a strong inverse relation between mortality in children younger than 5 years and household socioeconomic status (figure 2, appendix). In 2001–03 the probability of death from birth to age 5 years was 90·95 (95% CI 73·23–108·66) per 1000 person-years in the poorest households, which was significantly higher than in the richest households (53·98, 40·53–67·43). In 2011–13, both values had substantially reduced (42·81, 32·57–53·05 *vs* 19·46, 12·26–26·67 per 1000 person-years), but the difference remained significant. An inverse relation with socioeconomic status was also seen for women and men, and was significant for women from 2001 to 2007 and for men throughout the study period (figure 2, appendix). Although overall adult mortality remained higher in women in the poorest households in 2008–10 (probability of death from age 15 to 60 years 440·31, 95% CI 350·56–530·05) and in 2011–13 (325·96, 266·43–385·48) than in the richest households (322·57, 266·50–378·64 and 265·01, 224·25–305·77, respectively), the difference was not significant.

Substantial differences in life expectancy at birth associated with household wealth were evident for women in the periods 2001–03 and 2004–07, with the lowest life expectancy being seen in the poorest wealth quintile, but the difference progressively narrowed and became non-significant from 2008–10 (figure 2, appendix). The lowest life expectancy at birth was also seen in the poorest quintile for men, but significant differences persisted throughout the study period (figure 2 , appendix).

	2001 (n=70 809)	2005 (n=71 830)	2009 (n=75 603)	2013 (n=79 912)
Age group (years)				
<5	12·4%	11·4%	12·4%	11·7%
5–14	26·9%	24·6%	21·9%	21·3%
15–64	56·4%	59·2%	61·0%	62·3%
≥65	4·3%	4·8%	4·7%	4·7%
Education attainment among population aged ≥20 years				
No schooling	28·9%	23·9%	18·7%	14·2%
Higher education	4·8%	5·6%	5·9%	7·7%
Matriculated	8·8%	13·6%	18·4%	25·4%
Living conditions				
Dwelling materials				
Brick walls	1·5%	2·5%	6·2%	4·6%
Cement walls	75·9%	86·1%	89·1%	93·8%
Tiled roof	3·3%	6·0%	10·3%	15·8%
Corrugated iron roof	90·7%	90·9%	88·8%	83·8%
Tiled floor	0·5%	1·9%	4·8%	15·0%
Cement floor	90·7%	93·9%	93·8%	84·4%
Toilet facility				
Inside dwelling	0·2%	0·1%	0·6%	2·1%
Modern or flush toilet	0·2%	0·2%	0·2%	2·1%
Water supply				
Piped inside dwelling	0·9%	0·6%	1·6%	0·5%
Piped in the yard	18·1%	17·4%	28·0%	33·8%
Electricity				
For lighting	70·8%	90·5%	95·3%	97·0%
For cooking	13·2%	17·7%	36·0%	45·8%
Modern assets				
Mobile telephone	43·3%	82·2%	95·3%	98·8%
Television	59·2%	65·9%	78·0%	88·0%
Satellite television	0·3%	0·5%	6·0%	19·9%
Landline telephone	3·6%	1·8%	1·2%	0·9%
Motor car	17·5%	17·5%	22·0%	23·5%
Refrigerator	46·4%	64·5%	80·8%	90·2%
Electric or gas stove	40·9%	52·2%	76·3%	86·2%

Table 1: **Changes in distribution of age, education, and asset indicators over time**

In all time periods, the relative and absolute inequalities for summary mortality outcomes were inversely associated with household socioeconomic status (table 3). All RIIs for mortality were greater than 1, indicating greater mortality at the lower end of the socioeconomic continuum. For children younger than 5 years, the RIIs decreased in 2004–07 compared with those in 2001–03, but increased steadily thereafter, although the differences over time were not significant. For adults, the RIIs were significant within time periods, but the differences over time were not significant for men or women. Among women, however, the RIIs for mortality decreased steadily from the 2001–03 time period to the 2011–13 time period, whereas among men the values fluctuated.

	2001–03	2004–07	2008–10	2011–13	2001–13
Women and girls					
Person-years	110 608	155 062	138 883	145 799	550 352
Number of deaths					
All	1019 (100%)	1651 (100%)	1229 (100%)	1096 (100%)	4995 (100%)
HIV/AIDS and tuberculosis	505 (49·6%)	825 (50%)	476 (38·7%)	286 (26·1%)	2092 (41·9%)
Other communicable, maternal, perinatal, or nutritional causes	126 (12·4%)	219 (13·3%)	220 (17·9%)	237 (21·6%)	802 (16·1%)
Non-communicable	238 (23·4%)	391 (23·7%)	424 (34·5%)	452 (41·2%)	1 505 (30·1%)
Injuries	38 (3·7%)	46 (2·8%)	29 (2·4%)	31 (2·8%)	144 (2·9%)
Indeterminate	44 (4·3%)	93 (5·6%)	48 (3·9%)	41 (3·7%)	226 (4·5%)
Verbal autopsy interview not done	68 (6·7%)	77 (4·7%)	32 (2·6%)	49 (4·5%)	226 (4·5%)
Crude mortality per 1000 person-years	9·2	10·6	8·8	7·5	9·1
Men and boys					
Person-years	102 972	143 188	127 695	134 331	508 186
Number of deaths					
All	1115 (100%)	1833 (100%)	1363 (100%)	1108 (100%)	5419 (100%)
HIV/AIDS and tuberculosis	480 (43%)	755 (41·2%)	528 (38·7%)	300 (27·1%)	2063 (38·1%)
Other communicable, maternal, perinatal, or nutritional causes	126 (11·3%)	236 (12·9%)	271 (19·9%)	221 (19·9%)	854 (15·8%)
Non-communicable	230 (20·6%)	420 (22·9%)	337 (24·7%)	352 (31·8%)	1339 (24·7%)
Injuries	119 (10·7%)	160 (8·7%)	102 (7·5%)	127 (11·5%)	508 (9·4%)
Indeterminate	40 (3·6%)	74 (4%)	49 (3·6%)	46 (4·2%)	209 (3·9%)
Verbal autopsy interview not done	120 (10·8%)	188 (10·3%)	76 (5·6%)	62 (5·6%)	446 (8·2%)
Crude mortality per 1000 person-years	10·8	12·8	10·7	8·2	10·7
All					
Person-years	213 580	298 250	266 578	280 130	1 058 538
Number of deaths					
All	2134 (100%)	3484 (100%)	2592 (100%)	2204 (100%)	10 414 (100%)
HIV/AIDS and tuberculosis	985 (46·2%)	1580 (45·4%)	1004 (38·7%)	586 (26·6%)	4155 (39·9%)
Other communicable, maternal, perinatal, or nutritional causes	252 (11·8%)	455 (13·1%)	491 (18·9%)	458 (20·8%)	1 656 (15·9%)
Non-communicable	468 (21·9%)	811 (23·3%)	761 (29·4%)	804 (36·5%)	2844 (27·3%)
Injuries	157 (7·4%)	206 (5·9%)	131 (5·1%)	158 (7·2%)	652 (6·3%)
Indeterminate	84 (3·9%)	167 (4·8%)	97 (3·7%)	87 (3·9%)	435 (4·2%)
Verbal autopsy interview not done	188 (8·8%)	265 (7·6%)	108 (4·2%)	111 (5·0%)	672 (6·5%)
Crude mortality per 1000 person-years	10·0	11·7	9·7	7·9	9·8

Percentages may not sum to 100 because of rounding.

***Table 2*: Numbers of person-years and deaths, by time period and cause**

All SIIs for mortality were positive, indicating greater mortality at the lower end of the socioeconomic continuum. The SIIs for children younger than 5 years decreased in 2004–07 compared with those in 2001–03, but increased steadily thereafter, although the differences over time were not significant. The SIIs for adult mortality among women were significant within time periods, and steady decline over the entire study period meant that the differences were also significant over time. Among men, the SII values fluctuated and no significant difference was seen over time.

Relative inequalities in life expectancy at birth narrowed over time for men and women, but to a greater degree in women (table 3). The difference over time, therefore, was significant among women but not men. The SII values for life expectancy at birth also decreased steadily for women and men, again more so and significantly over time for women and in a non-significant fluctuating pattern for men.

The predicted probabilities of dying from different causes according to household wealth quintiles, adjusted for age, sex, and time period, are shown in figure 3, with

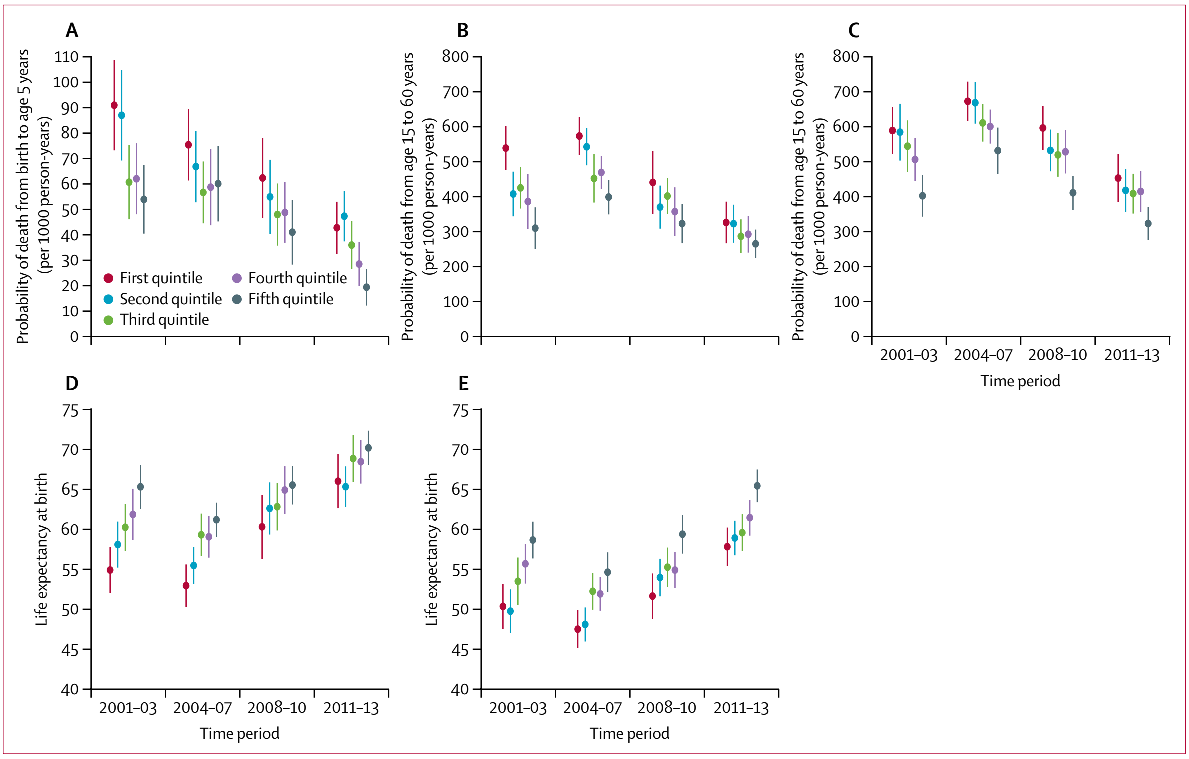


***Figure 2*: Differences in mortality and life expectancy at birth by household wealth and time period**

(A) Mortality in children younger than 5 years. (B) Mortality in women. (C) Mortality in men. (D) Life expectancy at birth in women. (E) Life expectancy at birth in men.

	2001–03	2004–07	2008–10	2011–13	p value*
Relative inequalities (RII, 95% CI)					
Mortality					
Children <5 years	2·06 (1·50 to 2·82)	1·37 (1·05 to 1·78)	1·62 (1·38 to 1·90)	2·38 (1·45 to 3·91)	0·075
Women	1·81 (1·33 to 2·45)	1·55 (1·30 to 1·84)	1·39 (1·13 to 1·71)	1·29 (1·16 to 1·44)	0·211
Men	1·54 (1·29 to 1·84)	1·33 (1·22 to 1·45)	1·43 (1·15 to 1·78)	1·38 (1·12 to 1·71)	0·629
Life expectancy at birth (years)					
Women	0·82 (0·79 to 0·84)	0·85 (0·79 to 0·91)	0·91 (0·88 to 0·93)	0·92 (0·88 to 0·96)	0·001
Men	0·81 (0·76 to 0·86)	0·84 (0·79 to 0·90)	0·86 (0·81 to 0·92)	0·86 (0·82 to 0·91)	0·414
Absolute inequalities (SII, 95% CI)					
Mortality					
Children <5 years	49·30 (26·18 to 72·41)	18·76 (1·99 to 35·53)	24·04 (15·57 to 32·51)	32·47 (18·42 to 46·52)	0·057
Women	237·12 (113·72 to 360·51)	208·11 (124·70 to 291·53)	122·63 (45·72 to 199·54)	75·55 (44·07 to 107·03)	0·031
Men	231·08 (152·01 to 310·16)	175·60 (125·07 to 226·12)	186·93 (80·34 to 293·53)	132·28 (50·08 to 214·48)	0·423
Life expectancy at birth (years)					
Women	–12·22 (–14·06 to –10·38)	–9·82 (–13·67 to –5·98)	–6·26 (–8·18 to –4·35)	–5·70 (–8·90 to –2·50)	0·004
Men	–11·37 (–14·66 to –8·08)	–8·96 (–12·32 to –5·60)	–8·17 (–11·90 to –4·44)	–8·93 (–12·20 to –5·67)	0·593

Mortality indicator values for socioeconomic groups are regressed on modified ridit scores, representing the relative ranks of groups in the cumulative distribution of household socioeconomic statuses, in generalised linear models. RII is the relative effect on the mortality indicator of moving from the lowest socioeconomic group to the highest. RII=1 implies that mortality in the lower and higher ends of the socioeconomic continuum do not differ, RII >1 implies greater mortality at the lower end, and RII <1 implies greater mortality at the higher end. SII is the absolute effect on the mortality indicator of moving from the lowest socioeconomic group through to the highest. SII=0 indicates that mortality at the lower and higher ends of the socioeconomic continuum do not differ, a positive SII indicates greater mortality at the lower end, and a negative SII indicates greater mortality at the higher end. RII and SII estimates are obtained from separate models for each time period and sex (except for in children <5 years), with only the modified ridit score as the independent variable. RII=relative index of inequality. SII=slope index of inequality. *For comparison of the RII or SII values in the different time periods (measured through two-way interaction terms between time periods and modified ridit scores).

***Table 3*: Relative and absolute socioeconomic inequalities in summary mortality indicators**

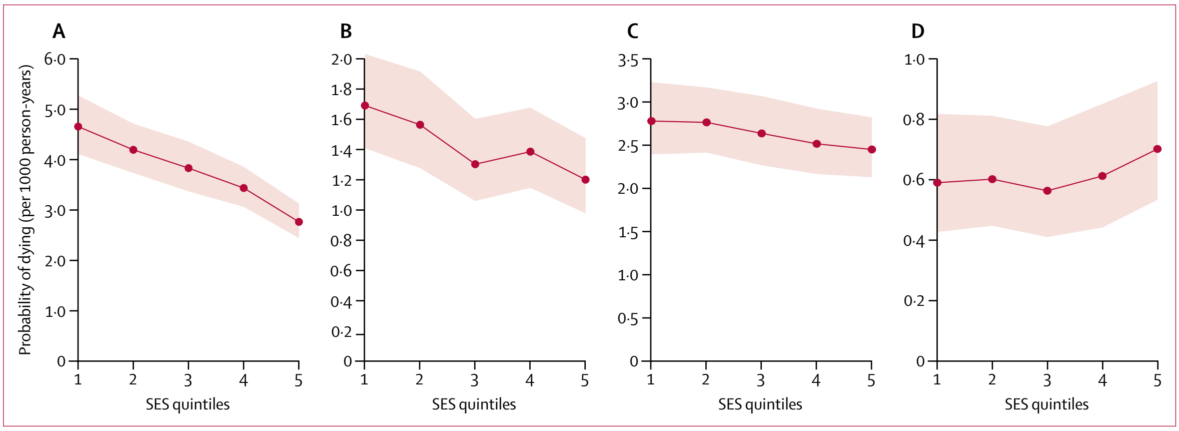


***Figure 3*: Annual probability of dying, by cause and quintile of household wealth for 2001–13**

(A) HIV/AIDS and tuberculosis. (B) Other communicable diseases (excluding HIV/AIDS and tuberculosis) and maternal, perinatal, and nutritional causes. (C) Non-communicable diseases. (D) Injuries. Data are predicted summed annual probabilities of death per 1000 person-years, by cause of death and household wealth quintiles that were estimated by multinomial logistic regression. All measures are adjusted for age, sex, and time period. SES=socioeconomic status.

corresponding relative risk ratios presented in the appendix. We found a strong and significant inverse relation between household wealth and death from HIV/AIDS and tuberculosis ($p<0\cdot0001$). A significant inverse relation was also seen between household wealth and death from other communicable diseases and maternal, perinatal, and nutritional causes ($p=0\cdot009$), although this was of a smaller magnitude and less consistent than that for HIV/AIDS and tuberculosis. Mortality from non-communicable diseases showed a non-significant inverse relation with household wealth, whereas mortality from injuries showed a non-significant positive relation with household wealth.

For men, boys, women, and girls, relative and absolute inequalities in mortality from HIV/AIDS and tuberculosis showed a persistent and significant inverse relation with socioeconomic status, with the highest values being in 2008–10 for men and boys and in 2004–07 for women and girls (table 4). For both sexes, the RIIs and SIIs associated with other causes of death fluctuated between being significant and non-significant over the period 2001–13. The RIIs for other communicable diseases and maternal, perinatal, and nutritional causes showed significant inverse relations with socioeconomic status only in 2011–13 for men and boys and in 2008–13 for women and girls. The SIIs for this cause-of-death category showed significant inverse relations with socioeconomic status only in 2011–13 for both sexes. For non-communicable diseases, no effect of socioeconomic status was seen on RIIs or SIIs in any period for women and girls, but a significant inverse relation was seen for men and boys in the RIIs for 2004–07 and 2008–10 and in the SIIs for 2004–07 and 2011–13. For injuries, only the RIIs for men and boys in 2004–07 and 2008–10 showed significant relations with socioeconomic status, but the relation was inverse in the earlier period and positive in the later period. No differences in SII and RII over time were significant for any cause of death except for the RII for injuries in men and boys.

Discussion

In a rural South African population, we found that socioeconomic disparities in mortality and life exepectancy at birth have evolved over the period 2001–13. Our findings update and improve those from earlier studies of socioeconomic differences in mortality in the Agincourt HDSS study population.[12,23,46] We included years in our analysis that cover the period before and after free ART was introduced. ART was first available from three district hospitals around the study area in 2004 and 2005.[23,47] From 2007, ART was also available within the study area from a privately funded community health centre specialising in HIV and tuberculosis care and treatment services, and run in partnership with the Department of Health (the Bhubezi Community Health Centre). Extension to public-sector primary-health-care facilities occurred in 2008 and 2009, and ART has been widely available since 2010.[47] We used a later version of the InterVA model than in previous studies, which strengthened cause-of-death assignment, and our analytical approach allowed us to estimate the relative and absolute socioeconomic inequalities in mortality across household wealth quintiles and to account for changes over time in the distribution of socioeconomic status. Our study additionally complements the second National Burden of Disease Study in South Africa,[27] which focused on differences between ethnic groups and provinces in mortality by focusing on socioeconomic differences at the local level in a resource-poor rural area and used data from rigorous longitudinal population surveillance.

Over the period 2001–13, the proportion of the population in Agincourt that lived in households owning assets associated with modern wealth increased substantially and polarisation in socioeconomic status declined,[17] although differences remained. Nevertheless, the population has undergone diverse health transitions

	2001–03	2004–07	2008–10	2011–13	p value*
Relative inequalities (RII, 95% CI)					
Men and boys					
HIV/AIDS and tuberculosis	1·67 (1·45 to 1·94)	1·74 (1·31 to 2·32)	4·70 (2·47 to 8·96)	2·09 (1·16 to 3·76)	0·125
Other communicable, maternal, perinatal, or nutritional causes	1·41 (0·92 to 2·16)	1·16 (0·79 to 1·70)	1·09 (0·85 to 1·39)	1·78 (1·28 to 2·49)	0·667
Non-communicable	0·84 (0·43 to 1·67)	1·54 (1·40 to 1·68)	1·49 (1·17 to 1·89)	1·36 (0·95 to 1·95)	0·334
Injuries	0·89 (0·50 to 1·61)	1·45 (1·11 to 1·89)	0·40 (0·22 to 0·73)	1·03 (0·52 to 2·08)	0·011
Women and girls					
HIV/AIDS and tuberculosis	1·89 (1·29 to 2·78)	2·73 (1·95 to 3·80)	1·42 (1·13 to 1·77)	2·53 (1·66 to 3·86)	0·118
Other communicable, maternal, perinatal, or nutritional causes	1·83 (0·83 to 4·03)	1·03 (0·62 to 1·70)	1·56 (1·13 to 2·17)	1·65 (1·10 to 2·48)	0·668
Non-communicable	0·96 (0·58 to 1·59)	1·06 (0·81 to 1·38)	0·99 (0·84 to 1·17)	0·84 (0·63 to 1·12)	0·648
Injuries	1·32 (0·45 to 3·90)	1·02 (0·25 to 4·18)	0·87 (0·37 to 2·06)	0·86 (0·25 to 2·93)	0·932
Absolute inequalities (SII, 95% CI)					
Men and boys					
HIV/AIDS and tuberculosis	4·94 (3·02 to 6·86)	5·67 (2·87 to 8·46)	8·17 (2·47 to 13·86)	3·71 (1·10 to 6·32)	0·603
Other communicable, maternal, perinatal, or nutritional causes	0·61 (–0·58 to 1·81)	0·75 (–0·38 to 1·88)	0·37 (–0·56 to 1·31)	2·45 (1·33 to 3·57)	0·643
Non-communicable	–0·36 (–6·29 to 5·57)	4·26 (1·13 to 7·40)	3·88 (–0·56 to 8·32)	4·28 (0·70 to 7·86)	0·658
Injuries	–0·11 (–0·79 to 0·58)	0·35 (–0·14 to 0·84)	–0·58 (–1·20 to 0·04)	0·11 (–0·52 to 0·74)	0·325
Women and girls					
HIV/AIDS and tuberculosis	3·28 (1·11 to 5·44)	5·04 (2·38 to 7·70)	1·58 (0·76 to 2·40)	1·94 (0·94 to 2·94)	0·186
Other communicable, maternal, perinatal, or nutritional causes	0·27 (–1·57 to 2·10)	0·16 (–0·99 to 1·32)	0·93 (–0·01 to 1·87)	1·27 (0·29 to 2·24)	0·705
Non-communicable	0·53 (–1·70 to 2·77)	0·87 (–0·70 to 2·45)	0·14 (–1·16 to 1·45)	–0·92 (–3·24 to 1·40)	0·619
Injuries	0·13 (–0·25 to 0·51)	0·02 (–0·29 to 0·34)	–0·02 (–0·22 to 0·19)	–0·03 (–0·31 to 0·25)	0·941

Cause-specific mortality is regressed on modified ridit scores representing the relative ranks of the socioeconomic groups in the cumulative distribution of household socioeconomic and age, in generalised linear models. RII is the relative effect on mortality of moving from the lowest socioeconomic group to the highest. RII=1 implies that mortality at the lower and higher ends of the socioeconomic continuum do not differ, RII >1 implies greater mortality at the lower end, and RII <1 implies greater mortality at the higher end. SII is the absolute effect on mortality of moving from the lowest socioeconomic group to the highest. SII=0 implies that mortality at the lower and higher ends of the socioeconomic continuum do not differ, a positive SII implies greater mortality at the lower end, and a negative SII implies greater mortality at the higher end. RII and SII estimates are obtained from separate models for each cause of death category, time period, and sex, with the modified ridit score and age group as the independent variables. RII=relative index of inequality. SII=slope index of inequality. *For comparison of RII or SII values in the different time periods (measured through the two-way interaction terms between time periods and modified ridit scores).

***Table 4*: Relative and absolute inequalities in mortality by cause of death and sex**

because of these changes. Our key findings are the significant relative and absolute socioeconomic gradients for mortality among children younger than 5 years, mortality in men and women, and life expectancy at birth throughout the 13-year study period, with outcomes being best in the wealthiest households. Despite ART being widely available and provided free of charge at public health facilities and HIV/AIDS-related mortality declining in recent years, the relative and absolute measures showed that significant inverse gradients in HIV/AIDS and tuberculosis mortality by household wealth have persisted. Although the proportion and number of deaths from non-communicable diseases are increasing, no significant difference associated with socioeconomic status was found. By contrast, in an earlier study, a persistent and significant inverse relation was reported between deaths from non-communicable diseases and household socioeconomic status for the period 2001–09.[23] This inconsistency between findings, however, might be due to version 4.0 of the InterVA model being used in the previous study. We applied version 4.0 to assign causes of death for the data used in this study and reproduced the significant inverse relation. Of note, our finding of a persistent significant inverse gradient between HIV/AIDS and tuberculosis mortality and household wealth, which was also reported in the earlier study, was not affected by which version of the InterVA model was used.

Several factors might explain the robustness of the finding that HIV/AIDS-related mortality is inversely related to socioeconomic status. First, a cross-sectional study of HIV prevalence in the Agincourt HDSS population in 2010–11 found a lower probability of being HIV positive among people living in households in the wealthiest socioeconomic status quintile than among those living in households in the poorest quintile.[48] Second, a negative gradient seems to exist in the availability of and access to resources that enable HIV-infected individuals to adopt strategies to improve their health and avoid HIV/AIDS-related mortality. No household in the Agincourt HDSS study area is more than 10 km from a primary health care facility, but the persistence of the significant inverse relation we saw between socioeconomic status and HIV/AIDS and tuberculosis mortality suggests that barriers to accessing HIV care and treatment services persist for individuals living in households with low socioeconomic status. Associated costs, such as transport to the health facility, probably hinder such individuals from accessing care and treatment services, despite ART being free of charge. Abgrall and del Amo[49] found that socioeconomic factors also affect retention in care and adherence to ART, which in turn affects survival for people living with HIV/AIDS. These factors might also contribute to the inverse socioeconomic gradient in HIV/AIDS-related mortality in the Agincourt population.

The overall mortality in this study was unexpectedly low for a poor rural setting, although the overall estimates for mortality among children younger than 5 years, mortality in adults, and life expectancy at birth in 2011–13 period were consistent with the unexpectedly low 2012 overall average estimates for Limpopo province reported in the second National Burden of Disease Study in South Africa.[27,50] Although the Agincourt HDSS study area is in Mpumalanga province, it is adjacent to and was previously within Limpopo province, from 1994 to 2005. Hence, the similarity in overall mortality is not too surprising, although the factors affecting mortality are not easy to explain.

Our study has several limitations. We acknowledge that using a household wealth index constructed from information on ownership of household assets is not the only way to measure socioeconomic status. Therefore, our findings might only partly reflect the evolution of socioeconomic disparities in mortality indicators. As in earlier studies, such as that by Houle and colleagues,[23] the data we used did not include individual-level measures of HIV seroprevalence or access to HIV care and treatment services. We are, therefore, unable to determine the magnitude of excess HIV/AIDS-related mortality among individuals from poor households that resulted specifically from increased risk of infection and barriers to care and treatment. Future analyses based on information generated by linking data in the Agincourt HDSS and the local health-care facilities will allow us to further refine our understanding of the mechanisms underlying our main findings.

Beyond HIV/AIDS and tuberculosis commonly occurring comcomitantly, difficulty in distinguishing HIV-related from non-HIV-related tuberculosis deaths with the verbal autopsy method made estimating the contribution of HIV/AIDS mortality alone to the socioeconomic gradient difficult. Substantially increased rates of tuberculosis, other communicable diseases, and non-communicable diseases were seen in the Agincourt population during the peak of HIV/AIDS mortality.[11] This pattern could make the socioeconomic differences in cause-specific mortality we identified less certain, although perhaps not substantially so, as another study showed that the InterVA-4 model had high specificity for HIV/AIDS-related mortality in relation to serostatus.[51] The data we used came from one geographically defined resource-poor rural area in South Africa, but the Agincourt area has similarities with other rural areas in South Africa. Therefore, although not directly transferable, our findings are likely to be relevant to other populations, especially those living in the north and northeast of the country, including areas bordering other countries. Our findings also highlight the need to include socioeconomic differences in assessments of health outcomes at the local level, even in resource-poor rural areas, because individuals in different socioeconomic positions might have different health and mortality profiles. Finally, we did not have sufficiently robust measures to assess social capital in the study area, and were unable to assess effects of this factor on risk of mortality.

HIV/AIDS and tuberculosis mortality in Agincourt is associated with disparities in socioeconomic status that does not seem to have changed over the period 2001–13, despite widespread availability and provision of free ART at public health facilities. This finding suggests that individuals from the poorest households continue to bear a disproportionately high burden of increased mortality and shortened lives related to the long-standing HIV/AIDS epidemic. The burden of mortality from non-communicable diseases is rising, and the association with household socioeconomic status is likely to become prominent. Integrated health-care planning and programme delivery strategies are needed to increase access to and uptake of HIV testing, linkage to care and ART, and prevention and treatment of non-communicable diseases among the poorest individuals to reduce the inequalities in cause-specific and overall mortality.

Contributors

CWK, BH, MAC, ST, and SJC conceived the study. CWK analysed the data with the support of BH. MAC and FXG-O oversaw the surveillance activities in the Agincourt Health and socio-Demographic Surveillance System. KK led the InterVA analysis of causes of death. MAC, KK, ST, and SJC provided overall guidance to the conduct of the study. CWK wrote the paper, which all authors reviewed and approved for submission.

Declaration of interests

We declare no competing interests.

Acknowledgments

The MRC/Wits Rural Public Health and Health Transitions Research Unit, which runs the Agincourt Health and socio-Demographic Surveillance System, was supported by the Wellcome Trust (grants 058893/Z/99/A, 069683/Z/02/Z, 085477/Z/08/Z, and 085477/B/08/Z), the South African Medical Research Council, and University of the Witwatersrand, Johannesburg South Africa. We thank the respondents, field staff, and management of the Agincourt Health and socio-Demographic Surveillance System for contributing to the production of the data.

References

1 Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. *Health Aff (Millwood)* 2002; **21:** 60–76.

2 Braveman P, Tarimo E. Social inequalities in health within countries: not only an issue for affluent nations. *Soc Sci Med* 2002; **54:** 1621–35.

3 Link BG, Phelan J. Social conditions as fundamental causes of disease. *J Health Soc Behav* 1995; Spec No: 80–94.

4 McKinnon B, Harper S, Kaufman JS, Bergevin Y. Socioeconomic inequality in neonatal mortality in countries of low and middle income: a multicountry analysis. *Lancet Glob Health* 2014; **2:** e165–73.

5 Link BG, Phelan JC. Fundamental sources of health inequalities. In: Mechanic D, Rogut LB, Colby DC, Knickman JR, eds. Policy challenges in modern health care. New Brunswick, NJ: Rutgers University Press, 2004.

6 Link BG, Phelan JC, Miech R, Westin EL. The resources that matter: fundamental social causes of health disparities and the challenge of intelligence. *J Health Soc Behav* 2008; **49:** 72–91.

7 Phelan JC, Link BG, Tehranifar P. Social conditions as fundamental causes of health inequalities theory, evidence, and policy implications. *J Health Soc Behav* 2010; **51** (1 suppl): S28–40.

8 Karim SSA, Churchyard GJ, Karim QA, Lawn SD. HIV infection and tuberculosis in South Africa: an urgent need to escalate the public health response. *Lancet* 2009; **374:** 921–33.

9 Department of Economic and Social Affairs Population Division. World population prospects: the 2010 revision. New York, NY: United Nations, 2011.

10 Bradshaw D, Laubscher R, Dorrington R, Bourne DE, Timaeus IM. Unabated rise in number of adult deaths in South Africa. *S Afr Med J* 2004; **94:** 278–79.

11 Kabudula CW, Tollman S, Mee P, et al. Two decades of mortality change in rural northeast South Africa. *Glob Health Action* 2014; **7:** 25596.

12 Kahn K, Garenne ML, Collinson MA, Tollman SM. Mortality trends in a new South Africa: hard to make a fresh start. *Scand J Public Health Suppl* 2007; **69:** 26–34.

13 Zwang J, Garenne M, Kahn K, Collinson M, Tollman SM. Trends in mortality from pulmonary tuberculosis and HIV/AIDS co-infection in rural South Africa (Agincourt). *Trans R Soc Trop Med Hyg* 2007; **101:** 893–98.

14 Tollman SM, Kahn K, Garenne M, Gear JS. Reversal in mortality trends: evidence from the Agincourt field site, South Africa, 1992–1995. *AIDS* 1999; **13:** 1091–97.

15 Herbst AJ, Cooke GS, Bärnighausen T, KanyKany A, Tanser F, Newell ML. Adult mortality and antiretroviral treatment roll-out in rural KwaZulu-Natal, South Africa. *Bull World Health Organ* 2009; **87:** 754–62.

16 Herbst AJ, Mafojane T, Newell ML, et al. Verbal autopsy-based cause-specific mortality trends in rural KwaZulu-Natal, South Africa, 2000–2009. *Popul Health Metr* 2011; **9:** 47.

17 Kabudula CW, Houle B, Collinson MA, Kahn K, Tollman S, Clark S. Assessing changes in household socioeconomic status in rural South Africa, 2001–2013: a distributional analysis using household asset indicators. *Soc Indic Res* 2016; published online June 28. DOI:10.1007/s11205-016-1397-z.

18 Bradshaw D, Groenewald P, Laubscher R, et al. Initial burden of disease estimates for South Africa, 2000. *S Afr Med J* 2003; **93:** 682–88.

19 Bradshaw D, Nannan N, Groenewald P, et al. Provincial mortality in South Africa, 2000-priority-setting for now and benchmark for the future. *S Afr Med J* 2005; **95:** 496–503.

20 Groenewald P, Bradshaw D, Daniels J, et al. Local-level mortality surveillance in resource-limited settings: a case study of Cape Town highlights disparities in health. *Bull World Health Organ* 2010; **88:** 444–51.

21 Hosegood V, Vanneste AM, Timæus IM. Levels and causes of adult mortality in rural South Africa: the impact of AIDS. *AIDS* 2004; **18:** 663–71.

22 Tollman SM, Kahn K, Sartorius B, Collinson MA, Clark SJ, Garenne ML. Implications of mortality transition for primary health care in rural South Africa: a population-based surveillance study. *Lancet* 2008; **372:** 893–901.

23 Houle B, Clark SJ, Gómez-Olivé FX, Kahn K, Tollman SM. The unfolding counter-transition in rural South Africa: mortality and cause of death, 1994–2009. *PLoS One* 2014; **9:** e100420.

24 Bradshaw D, Schneider M, Dorrington R, Bourne DE, Laubscher R. South African cause-of-death profile in transition—1996 and future trends. *S Afr Med J* 2002; **92:** 618–23.

25 Kahn K, Tollman SM, Garenne M, Gear JSS. Who dies from what? Determining cause of death in South Africa's rural north-east. *Trop Med Int Health* 1999; **4:** 433–41.

26 Kabudula CW, Houle B, Collinson MA, et al. Progression of the epidemiological transition in a rural South African setting: findings from population surveillance in Agincourt, 1993–2013. *BMC Public Health* 2017; **17:** 424.

27 Pillay-van Wyk V, Msemburi W, Laubscher R, et al. Mortality trends and differentials in South Africa from 1997 to 2012: second National Burden of Disease Study. *Lancet Glob Health* 2016; **4:** e642–53.

28 Kahn K, Collinson MA, Gómez-Olivé FX, et al. Profile: Agincourt Health and Socio-demographic Surveillance System. *Int J Epidemiol* 2012; **41:** 988–1001.

29 Kahn K, Tollman SM, Collinson MA, et al. Research into health, population and social transitions in rural South Africa: data and methods of the Agincourt Health and Demographic Surveillance System. *Scand J Public Health* 2007; **35** (69 suppl): 8–20.

30 Byass P, Chandramohan D, Clark SJ, et al. Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool. *Glob Health Action* 2012; **5:** 19281.

31 Kahn K, Tollman SM, Garenne M, Gear JSS. Validation and application of verbal autopsies in a rural area of South Africa. *Trop Med Int Health* 2000; **5:** 824–31.

32 Houle B, Stein A, Kahn K, et al. Household context and child mortality in rural South Africa: the effects of birth spacing, shared mortality, household composition and socio-economic status. *Int J Epidemiol* 2013; **42:** 1444–54.

33 Houle B, Pantazis A, Kabudula C, Tollman S, Clark SJ. Social patterns and differentials in the fertility transition in the context of HIV/AIDS: evidence from population surveillance, rural South Africa, 1993–2013. *Popul Health Metr* 2016; **14:** 10.

34 Houle B, Clark SJ, Kahn K, Tollman S, Yamin AE. The impacts of maternal mortality and cause of death on children's risk of dying in rural South Africa: evidence from a population based surveillance study (1992–2013). *Reprod Health* 2015; **12:** S7.

35 Mackenbach JP, Kunst AE. Measuring the magnitude of socio-economic inequalities in health: an overview of available measures illustrated with two examples from Europe. *Soc Sci Med* 1997; **44:** 757–71.

36 Wagstaff A, Paci P, Van Doorslaer E. On the measurement of inequalities in health. *Soc Sci Med* 1991; **33:** 545–57.

37 Allison PD. Event history analysis: regression for longitudinal event data. Thousand Oaks, CA: Sage, 1984.

38 Allison PD. Discrete-time methods for the analysis of event histories. *Sociol Methodol* 1982; **13:** 61–98.

39 Allison PD. Survival analysis using SAS®: a practical guide, 2nd edn. London: SAS Institute, 2010.

40 Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *J Am Stat Assoc* 1988; **83:** 414–25.

41 Ernstsen L, Strand BH, Nilsen SM, Espnes GA, Krokstad S. Trends in absolute and relative educational inequalities in four modifiable ischaemic heart disease risk factors: repeated cross-sectional surveys from the Nord-Trøndelag Health Study (HUNT) 1984-2008. *BMC Public Health* 2012; **12**(1): 266.

42 Wamala S, Blakely T, Atkinson J. Trends in absolute socioeconomic inequalities in mortality in Sweden and New Zealand. A 20-year gender perspective. *BMC Public Health* 2006; **6**(1): 164.

43 Van Hook J, Altman CE. Using discrete-time event history fertility models to simulate total fertility rates and other fertility measures. *Popul Res Policy Rev* 2013; **32:** 585–610.

44 Galobardes B, Shaw M, Lawlor DA, Lynch JW, Smith GD. Indicators of socioeconomic position (part 1). *J Epidemiol Community Health* 2006; **60:** 7–12.

45 Howe LD, Galobardes B, Matijasevich A, et al. Measuring socio-economic position for epidemiological studies in low-and middle-income countries: a methods of measurement in epidemiology paper. *Int J Epidemiol* 2012; **41:** 871–86.

46 Sartorius B, Kahn K, Collinson MA, Sartorius K, Tollman SM. Dying in their prime: determinants and space-time risk of adult mortality in rural South Africa. *Geospatial Health* 2013; **7:** 237.

47 Mee P, Collinson MA, Madhavan S, et al. Evidence for localised HIV related micro-epidemics associated with the decentralised provision of antiretroviral treatment in rural South Africa: a spatio-temporal analysis of changing mortality patterns (2007–2010). *J Glob Health* 2014; **4:** 010403.

48 Gómez-Olivé FX, Angotti N, Houle B, et al. Prevalence of HIV among those 15 and older in rural South Africa. *AIDS Care* 2013; **25:** 1122–28.

49 Abgrall S, Del Amo J. Effect of sociodemographic factors on survival of people living with HIV. *Curr Opin HIV AIDS* 2016; **11:** 501–06.

50 Msemburi W, Pillay-van Wyk V, Dorrington R, et al. Second national burden of disease study for South Africa: cause-of-death profile for South Africa, 1997–2010. Cape Town: South African Medical Research Council, 2016.

51 Byass P, Calvert C, Miiro-Nakiyingi J, et al. InterVA-4 as a public health tool for measuring HIV/AIDS mortality: a validation study from five African countries. *Glob Health Action* 2013; **6:** 22448.

THE LANCET
Global Health

Supplementary appendix

Supplement to: Kabudula CW, Houle B, Collinson MA, et al. Socioeconomic differences in mortality in the antiretroviral therapy era in Agincourt, rural South Africa, 2001–13: a population surveillance analysis. *Lancet Glob Health* 2017; **5:** e924–35.

APPENDIX

Statistical analysis

Calculation of RII and SII

RIIs and SIIs were obtained from generalised linear models by regressing the mortality indicator (under-five mortality, adult mortality, life expectancy at birth, and cause-specific mortality rate) of each socioeconomic group on a modified 'ridit' score,[1,2] representing the relative rank of that group in the cumulative distribution of household SES. The modified ridit score ranges from 0 (richest) to 1 (poorest) and is calculated by arranging the household wealth quintiles in order from richest to poorest and assigning a cumulative proportion of the total population to each quintile. Thereafter, half the proportion of the total population in Q5 is taken as the modified ridit score for Q5 and half the proportion of the total population in Q4 added to the proportion in Q5 as the modified ridit score for Q4 and so on. For example, if the proportion of the total population in Q5 is 20%, the modified ridit score takes the value 0·1 (0·2/2) and if the total population in Q4 is 18% then the modified ridit score for Q4 takes the value 0·29 (0·2+[0·18/2]) and so forth. The beta or slope coefficient on modified ridit scores from each model expresses RII when the logarithm link function is used, and SII when the identity link function is used.

[1] Ernstsen L, Strand BH, Nilsen SM, Espnes GA, Krokstad S. Trends in absolute and relative educational inequalities in four modifiable ischaemic heart disease risk factors: repeated cross-sectional surveys from the Nord-Trøndelag Health Study (HUNT) 1984–2008. *BMC Public Health* 2012; **12:** 266.

[2] Wamala S, Blakely T, Atkinson J. Trends in absolute socioeconomic inequalities in mortality in Sweden and New Zealand. A 20-year gender perspective. *BMC Public Health* 2006; **6:** 164.

Table S1: Selected summary mortality indicators by household wealth, Agincourt, South Africa, 2001-2013

		Q1 (Low SES)	**Q2**	**Q3**	**Q4**	**Q5 (High SES)**	**Overall**
Under five mortality rate: Both sexes							
	2001–2003	90·95(73·23,108·66)	86·97(69·24,104·69)	60·73(46·15,75·31)	62·06(48·12,76·00)	53·98(40·53,67·43)	68·02(61·44,75·26)
	2004–2007	75·40(61·39,89·42)	66·84(52·81,80·87)	56·71(44·56,68·86)	58·75(43·79,73·72)	60·12(45·31,74·94)	61·86(56·62,67·58)
	2008–2010	62·38(46·68,78·08)	54·93(40·32,69·54)	48·02(35·81,60·22)	48·84(36·95,60·74)	41·07(28·36,53·79)	49·94(45·02,55·39)
	2011–2013	42·81(32·57,53·05)	47·35(37·47,57·23)	36·02(26·55,45·49)	28·56(19·90,37·22)	19·46(12·26,26·67)	33·47(29·45,38·03)
Adult mortality rate: Females							
	2001–2003	538·46(475·24,601·69)	407·33(343·47,471·19)	424·81(365·73,483·90)	385·78(306·83,464·74)	309·70(250·20,369·21)	400·05(372·28,429·11)
	2004–2007	572·89(518·36,627·42)	542·29(489·22,595·35)	451·92(382·91,520·93)	468·84(421·27,516·41)	398·54(349·01,448·06)	472·63(449·47,496·38)
	2008–2010	440·31(350·56,530·05)	369·85(308·34,431·36)	401·39(350·20,452·58)	356·91(287·36,426·46)	322·57(266·50,378·64)	374(349·26,399·91)
	2011–2013	325·96(266·43,385·48)	322·56(268·21,376·91)	286·50(238·15,334·85)	292·31(239·88,344·74)	265·01(224·25,305·77)	296·31(274·4,319·56)
Adult mortality rate: Males							
	2001–2003	587·70(521·07,654·33)	583·21(502·00,664·42)	542·96(469·02,616·91)	505·08(444·16,566·00)	401·39(341·78,461·00)	514·62(483·02,547·04)
	2004–2007	671·20(614·84,727·57)	667·03(607·44,726·63)	609·51(556·16,662·85)	599·13(550·47,647·78)	530·29(464·28,596·29)	607·24(583,631·59)
	2008–2010	595·04(532·46,657·61)	530·96(471·28,590·64)	518·09(455·89,580·29)	527·08(465·06,589·10)	409·80(361·20,458·40)	508·96(481·98,536·55)
	2011–2013	451·90(383·61,520·20)	416·69(354·75,478·63)	407·57(350·67,464·47)	413·64(354·63,472·65)	322·19(274·40,369·99)	398·25(370·96,426·8)
Life expectancy at birth: Females							
	2001–2003	54·94(52·08,57·80)	58·12(55·24,61·00)	60·28(57·34,63·22)	61·89(58·69,65·10)	65·34(62·58,68·10)	60·86(59·47,62·26)
	2004–2007	52·98(50·31,55·64)	55·51(53·19,57·82)	59·34(56·69,61·99)	59·09(56·49,61·69)	61·22(59·08,63·37)	58·41(57·28,59·53)
	2008–2010	60·33(56·34,64·32)	62·64(59·38,65·90)	62·85(59·89,65·81)	64·93(61·96,67·91)	65·55(63·13,67·97)	63·63(62·43,64·83)
	2011–2013	66·04(62·66,69·42)	65·35(62·81,67·90)	68·87(65·94,71·80)	68·48(65·75,71·21)	70·21(68·05,72·36)	68·02(66·83,69·21)
Life expectancy at birth: Males							
	2001–2003	50·33(47·50,53·16)	49·73(46·99,52·47)	53·48(50·51,56·44)	55·65(53·19,58·12)	58·63(56·33,60·93)	54·01(52·78,55·23)
	2004–2007	47·48(45·11,49·86)	48·08(45·95,50·21)	52·22(49·92,54·51)	51·89(49·80,53·99)	54·60(52·11,57·09)	51·26(50·33,52·2)
	2008–2010	51·62(48·78,54·46)	53·95(51·60,56·30)	55·23(52·76,57·69)	54·87(52·63,57·12)	59·35(56·93,61·77)	55·33(54·26,56·41)
	2011–2013	57·79(55·39,60·19)	58·88(56·72,61·05)	59·54(57·24,61·85)	61·42(59·18,63·66)	65·40(63·34,67·45)	60·86(59·78,61·93)

Table S2: Multinomial logistic regression of death by cause, sex, age, time period and household wealth quintile.

Variable	Relative Risk Ratio	95% CI	p-value
HIV/AIDS and TB			
Sex			
Female	1	–	–
Male	1·13	[1·06,1·20]	<0·0001
Age Groups			
0-4	1	–	–
5-14	0·12	[0·10,0·16]	<0·0001
15-49	1·60	[1·43,1·78]	<0·0001
50-64	3·29	[2·89,3·75]	<0·0001
65+	3·30	[2·86,3·79]	<0·0001
Time Period			
2001–2003	1	–	–
2004–2007	1·10	[1·01,1·19]	0·022
2008–2010	0·76	[0·70,0·83]	<0·0001
2011–2013	0·42	[0·38,0·46]	<0·0001
Wealth Quintile			
Q1 (Lowest)	1	–	–
Q2	0·90	[0·78,1·04]	0·150
Q3	0·81	[0·69,0·94]	0·009
Q4	0·71	[0·62,0·80]	<0·0001
Q5 (Highest)	0·56	[0·49,0·64]	<0·0001
***Other comm/mat/peri/nutr Causes**			
Sex			
Female	1	–	–
Male	1·19	[1·08,1·32]	0·001
Age Groups			
0-4	1	–	–
5-14	0·05	[0·04,0·06]	<0·0001
15-49	0·15	[0·13,0·17]	<0·0001
50-64	0·35	[0·29,0·42]	<0·0001
65+	0·73	[0·62,0·86]	<0·0001
Time Period			
2001–2003	1	–	–
2004–2007	1·26	[1·07,1·47]	0·005
2008–2010	1·46	[1·24,1·70]	<0·0001
2011–2013	1·32	[1·12,1·54]	0·001
Wealth Quintile			
Q1 (Lowest)	1	–	–
Q2	0·95	[0·76,1·18]	0·612
Q3	0·80	[0·67,0·96]	0·019
Q4	0·86	[0·72,1·04]	0·113

Q5 (Highest)	0·76	[0·62,0·93]	0·009
Non-communicable Causes			
Sex			
Female	1	–	–
Male	1·47	[1·36,1·58]	<0·0001
Age Groups			
0-4	1	–	–
5-14	0·20	[0·13,0·32]	<0·0001
15-49	2·62	[2·05,3·35]	<0·0001
50-64	15·24	[11·86,19·59]	<0·0001
65+	57·78	[45·33,73·63]	<0·0001
Time Period			
2001–2003	1	–	–
2004–2007	1·17	[1·04,1·32]	0·008
2008–2010	1·28	[1·14,1·44]	<0·0001
2011–2013	1·30	[1·16,1·46]	<0·0001
Wealth Quintile			
Q1 (Lowest)	1	–	–
Q2	1·03	[0·91,1·17]	0·635
Q3	0·99	[0·80,1·22]	0·928
Q4	0·94	[0·79,1·12]	0·469
Q5 (Highest)	0·89	[0·75,1·06]	0·189
Injuries			
Sex			
Female	1	–	–
Male	4·13	[3·42,4·98]	<0·0001
Age Groups			
0-4	1	–	–
5-14	0·79	[0·49,1·26]	0·320
15-49	3·77	[2·57,5·52]	<0·0001
50-64	4·74	[3·04,7·37]	<0·0001
65+	6·08	[3·82,9·68]	<0·0001
Time Period			
2001–2003	1	–	–
2004–2007	0·89	[0·72,1·10]	0·269
2008–2010	0·65	[0·51,0·82]	<0·0001
2011–2013	0·75	[0·60,0·94]	0·011
Wealth Quintile			
Q1 (Lowest)	1	–	–
Q2	1·01	[0·70,1·45]	0·967
Q3	0·93	[0·61,1·40]	0·696
Q4	0·98	[0·66,1·46]	0·932
Q5 (Highest)	1·08	[0·82,1·43]	0·563
Indeterminate			

Sex			
Female	1	–	–
Male	1·80	[1·59,2·04]	<0·0001
Age Groups			
0-4	1	–	–
5-14	0·13	[0·10,0·18]	<0·0001
15-49	0·56	[0·47,0·67]	<0·0001
50-64	1·88	[1·52,2·32]	<0·0001
65+	4·16	[3·42,5·06]	<0·0001
Time Period			
2001–2003	1	–	–
2004–2007	1·09	[0·94,1·27]	0·271
2008–2010	0·59	[0·49,0·71]	<0·0001
2011–2013	0·44	[0·36,0·54]	<0·0001
Wealth Quintile			
Q1 (Lowest)	1	–	–
Q2	0·94	[0·76,1·17]	0·586
Q3	0·82	[0·66,1·01]	0·062
Q4	0·92	[0·75,1·11]	0·376
Q5 (Highest)	0·80	[0·66,0·98]	0·033

***The abbreviation “Other comm/mat/peri/nutr causes” stands for “other communicable (excluding HIV/AIDS and TB), maternal, perinatal, and nutritional causes of death”**

PAPER IV

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa

Chodziwadziwa W Kabudula[1*], Benjamin D Clark[2], Francesc Xavier Gómez-Olivé[1], Stephen Tollman[1,3,4], Jane Menken[1,5] and Georges Reniers[6]

Abstract

Background: Health and Demographic Surveillance Systems (HDSS) have been instrumental in advancing population and health research in low- and middle- income countries where vital registration systems are often weak. However, the utility of HDSS would be enhanced if their databases could be linked with those of local health facilities. We assess the feasibility of record linkage in rural South Africa using data from the Agincourt HDSS and a local health facility.

Methods: Using a gold standard dataset of 623 record pairs matched by means of fingerprints, we evaluate twenty record linkage scenarios (involving different identifiers, string comparison techniques and with and without clerical review) based on the Fellegi-Sunter probabilistic record linkage model. Matching rates and quality are measured by their sensitivity and positive predictive value (PPV). Background characteristics of matched and unmatched cases are compared to assess systematic bias in the resulting record-linked dataset.

Results: A hybrid approach of deterministic followed by probabilistic record linkage, and scenarios that use an extended set of identifiers including another household member's first name yield the best results. The best fully automated record linkage scenario has a sensitivity of 83.6% and PPV of 95.1%. The sensitivity and PPV increase to 84.3% and 96.9%, respectively, when clerical review is undertaken on 10% of the record pairs. The likelihood of being linked is significantly lower for females, non-South Africans and the elderly.

Conclusion: Using records matched by means of fingerprints as the gold standard, we have demonstrated the feasibility of fully automated probabilistic record linkage using identifiers that are routinely collected in health facilities in South Africa. Our study also shows that matching statistics can be improved if other identifiers (e.g., another household member's first name) are added to the set of matching variables, and, to a lesser extent, with clerical review. Matching success is, however, correlated with background characteristics that are indicative of the instability of personal attributes over time (e.g., surname in the case of women) or with misreporting (e.g., age).

Keywords: Health and Demographic Surveillance System (HDSS), Record linkage, Health facilities, South Africa, Population surveillance

* Correspondence: chodziwadziwa.kabudula@wits.ac.za
[1] MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
Full list of author information is available at the end of the article

Background

Health and Demographic Surveillance Systems (HDSS) enumerate populations in geographically well-defined areas and prospectively collect detailed information on vital events including births, deaths, and migrations, as well as complementary data covering health, social and economic indicators [1-3]. These data allow for population-based investigations of population and health dynamics and their determinants in low- and middle- income countries where vital registration systems are often weak [2]. However, the scope of analysis possible with datasets from most HDSSs is constrained by the lack of integration with other administrative data, including those emanating from health facilities. For example, HDSS data have demonstrated reductions in overall mortality levels in HIV/AIDS affected African populations following the expansion of antiretroviral therapy programs [4-6], but residual AIDS mortality remains important. In order to achieve further reductions in mortality levels, it is important to understand whether individuals dying of AIDS have had any contact with the health facilities and the nature of that contact (e.g., diagnosis, in care awaiting treatment initiation, on first line treatment). Unfortunately, this is difficult without linking HDSS and health facility data. The best measures currently available on health care utilization rely on retrospective reports from living patients or from relatives or caretakers of the deceased. Data from health facilities alone do not address these types of research and policy questions either as they fail to account for individuals who never make contact with the health facility.

Record linkage of electronic patient records based on conventional personal identifiers is a cost-effective means for integrating information from different sources [7]. This approach has been applied extensively to generate datasets for epidemiological studies in higher income settings (e.g., United States of America [8,9], Wales [10], Australia [11-13], Italy [14], Canada [15], Netherlands [16] and the United Kingdom [17]) but it is much less common in African populations or in the context of HDSS[a]. Obstacles to record linkage in these settings include the lack of unique and ubiquitous identification systems (e.g., national insurance or social security number), variation in the transcription of names, imprecision in the reporting of dates, and other data quality related issues.

In this study, we assess the feasibility of record linkage with conventional personal identifiers (e.g., name, age, address) between an HDSS and a health facility in South Africa using data from the Agincourt HDSS and patient attendance records from a local government health facility. Our study is unusual because we first construct a gold standard dataset of records matched by means of fingerprints and subsequently use it to assess the coverage and accuracy of various record linkage scenarios. Finally, we compare the background characteristics of matched and unmatched cases, and evaluate compositional differences in the linked and full dataset.

There are three reasons why we pursue record linkage on conventional personal identifiers as opposed to record linkage on fingerprints. First, fingerprints are known to have a very high specificity but relatively low sensitivity [18]. This property renders fingerprint-matched records a good gold standard for evaluating other record linkage approaches, but makes it less desirable as a record linkage solution itself. Other biometric identifiers (e.g., iris scan and facial recognition) may outperform fingerprints in that regard. Second, record linkage on the basis of fingerprints (or any other biometric) would require the HDSS to collect and store fingerprints for all its residents, and we chose to assess the utility of a cheaper method. Third, fingerprint-based record linkage would require that fingerprint collection becomes part of the patient administration systems in all health facilities. Since many health facilities in low- and middle- income countries do not have computerized health management information systems, this is unlikely to become a realistic solution in the short term.

Methods

Datasets

Three datasets are used in this study. The first dataset (dataset1) consists of identifiers of 93,507 individuals who were under surveillance by the Agincourt HDSS at any time between 1 August 2009 and 1 August 2010. The Agincourt HDSS encompasses 27 villages spread over 420 km^2 of semi-arid scrubland in rural northeast South Africa in the Bushbuckridge sub-district of Ehlanzeni district, Mpumalanga Province [19,20]. The population under surveillance is largely Xitsonga-speaking with one-third being former Mozambican refugees who arrived in the 1980s- and their descendants.

The second dataset (dataset2) consists of identifiers and fingerprints of 2,865 individuals aged 18 years and above from two villages in the Agincourt HDSS. The fingerprints were collected during a mini-census in which 6,185 residents aged 18 years and above were visited in their homes between November 2008 and April 2009. Verbal informed consent was obtained to collect fingerprints and to link the Agincourt HDSS database record to any visits to Agincourt Health Centre (AHC), which is one of eight local health facilities within the Agincourt HDSS. Between two and four fingerprints were collected from each individual who agreed to participate in the study. A large number of the individuals from whom fingerprints could not be collected were absent during the household visits (circular labor migration is very common in the area). Among the individuals who were found at home (2,965 individuals), only 45 individuals refused participation, and technical problems with the collection of fingerprints (often due to scars or cuts on the finger) accounted for 55 cases. Details about the community-based fingerprint collection are presented elsewhere [21].

The third dataset (dataset3) consists of identifiers and fingerprints that were collected as part of a pilot electronic patient registration system at the reception desk of the AHC. This electronic patient registration system was managed by SAP Meraka Unit for Technology Development (UTD) and the School of Public Health from the University of the Witwatersrand [22]. The data were collected between August 2008 and August 2010. Identifiers were collected from 10,790 individuals and fingerprints from 3,633 of them. At least two fingerprints were collected from 93.6% of these 3,633 individuals. Fingerprints were not collected for extended periods of time at the AHC because of technical problems that the personnel at the reception desk could not independently resolve.

Identifiers included in dataset3 are those that are routinely collected at the AHC such as first name, surname, sex, date of birth, and place of residence, and attributes that we added to the patient registration for the purpose of this study (e.g., the first and surname of another household member). National ID number and telephone number were also on the list of identifiers to be collected but were not consistently reported by individuals attending the AHC. In anticipation of this (and future) record linkage studies we collect National ID number and telephone number(s) during the annual Agincourt HDSS census update since 2007 and 2011 respectively. Additionally, we have included the collection of *other names* for all individuals in the annual Agincourt HDSS census update since 2011.

Gold standard dataset

We constructed a dataset of matched individuals from the Agincourt HDSS and the AHC by linking individuals' fingerprints in dataset2 with the fingerprints in dataset3. Matching of the fingerprints was performed using the SAGEM MorphoSmart Compact Biometric Module (CBM) with a threshold of 5 as recommended by the manufacturer [23]. The threshold can be varied from 0 to 10 with higher thresholds producing less false positive cases and lower thresholds producing fewer false negatives. The threshold of 5 has a false acceptance rate (FAR = 1-Specificity) <0.01% [23].

The matching of fingerprints from the 2,865 individuals in the two target villages of the Agincourt HDSS with those captured from the 3,633 individuals that visited the AHC resulted in 623 matched record pairs. At least two fingerprints were matched in 393 (63.08%) cases.

Record linkage with conventional personal identifiers

We use two approaches for linking individuals in dataset1 with individuals in dataset3. In the first approach we exclusively use probabilistic record linkage methods. In the second approach we use a hybrid strategy whereby we first link records deterministically and thereafter match the remaining records using probabilistic methods. Deterministic record linkage designates a pair of records from two data sources as belonging to the same individual when they match on a unique identifier such as fingerprints, a social security or national identification number, or a set of conventional personal identifiers (e.g., the combination of first name, last name and date of birth) [24-27]. Probabilistic record linkage classifies a pair of records from two data sources as belonging to the same individual based on the statistical probability that common identifiers drawn from the two data sources belong to the same individual [28-33]. Whereas deterministic linkage is most suitable when unique identifiers are available and the quality of the data are high, probabilistic linkage yields better results when unique identifiers are lacking or in situations where there is variation in reporting or transcription of personal identifiers [24,29,34-36].

We first define 15 probabilistic record linkage scenarios (S1-S15) based on different combinations of personal identifiers or linking variables (first name, surname, day of birth, month of birth, year of birth, village and first name and surname of another household member), and various string comparison techniques to accommodate typographical errors and spelling variation in first and surnames. The string comparison techniques used are the Jaro-Winkler (JW) string comparator [37], the Soundex phonetic encoding and the Double Metaphone phonetic encoding [38]. Details about these probabilistic linkage scenarios are given in Table 1.

Thereafter, we create another scenario (S16 in Table 1), which first matches records deterministically using National ID number or a combination of telephone number and first name, and subsequently matches the remaining cases using the scenario that yields the maximum sensitivity and positive predictive value (PPV) among the first 15 probabilistic linkage scenarios.

Since the number of possible record pair comparisons in two data files to be linked is enormous - equal to the product of the number of records on each file (over 1 billion record pairs in our case) - we use a technique called "blocking" to restrict the comparison space to blocks or pockets of record pairs where one or more variables match exactly [31]. Blocking is useful for reducing computing time, but may decrease the sensitivity if blocking variables are measured with error. In order to minimize the effect of errors in blocking variables, we

Table 1 Linkage scenarios by identifiers and string comparison techniques applied to names

		String comparison techniques applied to first and surnames					
		Exact	JW ≥ 0.7	JW ≥ 0.9	DM	Soundex	JW ≥ 0.9 or DM or soundex
Identifiers used	Routinely collected identifiers*	S1	S2	S3	S4	S5	S6
	Routinely collected identifiers + household member first name	S7	S8	S9	S10	S11	S12
	Routinely collected identifiers + household member first name and surname			S13	S14	S15	
	Deterministic linkage on National ID Number or telephone number followed by best of S1-S15**						S16
	S16 + clerical review of 5%, 10%, 15%, and 20% of record pairs above and below the threshold value above which record pairs are automatically accepted as matches						S17-S20

*Routinely collected identifiers = first name, last name, sex, day of birth, month of birth, year of birth and village; JW = Jaro-Winkler; DM = double metaphone code.
**The best of the 15 probabilistic linkage scenarios is the one that yields the maximum sensitivity and PPV.

use three blocking schemes: exact match on sex and year of birth (BS-1), exact match on sex and village (BS-2) and exact match on the first letter of the first name and surname and age difference of not more than 10 years (BS-3). We combine linked record pairs from the different blocks and extract a unique set of linked record pairs as a combination of all distinct record pairs and the record pair with the highest matching score (see below) in cases where a record from dataset3 is matched to multiple records in dataset1.

A key step in probabilistic linkage is the estimation of weights to indicate the contribution of each identifier to the probability of accurately designating a pair of records from two different sources as either a match or non-match [27,30,31]. For each common identifier, i, available in the two data sources, the process involves first estimating the probability that the identifier agrees given that the two records belong to the same individual, denoted by m_i, and the probability that the identifier agrees given that the two records do not belong to the same individual, denoted by u_i [30,31,33]. The m_i values depend on measurement and reporting error in an identifier whereas the u_i values depend on the number of distinct values of an identifier and their frequencies [32,39]. Identifiers collected and recorded with good quality in both datasets have higher m_i values. On the other hand, identifiers with many different values are less likely to agree by chance, and hence, have lower u_i values. In record pairs where identifier i agrees, the identifier is assigned a weight value of $\log_{2}\frac{m_i}{u_i}$ and where identifier i disagrees a weight value of $\log_{2}\frac{1-m_i}{1-u_i}$ is assigned. Thereafter each record pair is classified as a match or non-match depending on whether the sum of the weights on all the identifiers used (matching score) is above or below a threshold value above which record pairs are automatically accepted as matches.

For each scenario, we estimate m_i and u_i probabilities from the datasets to be linked using an Expectation Maximization (EM) algorithm [31,40,41] based on the Fellegi-Sunter model [42]. Following Méray et al. [39] and Tromp et al. [43], we use an estimate of the proportion of true matches among all possible record pair combinations to determine a scenario-specific threshold matching score above which record pairs are automatically accepted as matches.

Finally, we create four more scenarios (S17-S20 in Table 1) that use scenario S16 as the starting point and add clerical review for a selection of record pairs immediately above and below the threshold value. These scenarios allocate 5% (S17), 10% (S18), 15% (S19) and 20% (S20) of record pairs immediately above and below the threshold value in scenario S16 to clerical review. Two reviewers independently review the targeted record pairs and classify each of them as a match or non-match. When the two reviewers disagree, a third reviewer adjudicates over the match status.

There are four possible outcomes from record linkage: true matches (true positives), true non-matches (true negatives), mismatches (false positives) and false non-matches (false negatives) [44]. Coverage and accuracy of each linkage scenario can thus be assessed by four indices: sensitivity, specificity, PPV and negative predictive value (NPV). Sensitivity is the proportion of true matches that are produced by the linkage algorithm, specificity is the proportion of true non-matches, PPV is the proportion of matches produced by the linkage algorithm that are true matches and NPV is the proportion of non-matches produced by the linkage algorithm that are true non-matches [45]. However, as the number of true non-matches are often very large, specificity and NPV are not very informative [34]. Therefore, we report sensitivity and PPV for each linkage scenario against the gold standard.

Bias in the record-linked dataset

Because record linkage may produce mismatches and missed matches it is recommended that linked and unlinked records are assessed for systematic bias [46,47]. We thus select cases for which we know the true match status from the gold standard dataset and regress the record linkage outcome on individual characteristics using a logistic model. Age, sex, residency status in the Agincourt HDSS, nationality, level of education, employment status and household wealth quintile are considered as predictors of accurate linkage. Wealth quintiles are derived from data on ownership of assets such as cattle, car, and cell phone as well as access to amenities including drinking water and sanitation using principal components analysis [48]. In addition to this individual-level assessment of factors associated with linkage success, we also compare the distribution of background characteristics in the gold standard and record linked datasets using Pearson Chi squared tests.

Implementation

We implemented the record linkage with conventional personal identifiers in Microsoft SQL Server 2008. The EM algorithm used to estimate the *m* and *u* probabilities and the proportion of true matches among all possible record pair combinations is implemented in Microsoft C# and integrated into Microsoft SQL Server as a common language runtime (CLR) function. The Soundex algorithm is a Microsoft SQL Server built-in function. The JW and Double Metaphone algorithms were integrated into Microsoft SQL Server as CLR functions. The JW algorithm is part of the SimMetrics library and its source code is freely available [49]. The source code for the Double Metaphone algorithm is also freely available [50]. Data analysis is conducted in Stata version 12.

Ethical approval

The study received ethical approvals from the University of the Witwatersrand Human Research Ethics Committee (Clearance number: M071141) and the Mpumalanga Provincial Department of Health Research and Ethics Committee.

Results

The level of completeness of the identifiers used as linking variables in the various scenarios is higher in the data from the Agincourt HDSS compared to that from the AHC (Table 2). Village, another household member's first and surname, National ID number and telephone number are often missing in the AHC dataset. None of these characteristics are routinely recorded in health facilities.

Table 2 Completeness of identifiers from both sources

Identifier	Percentage of individuals with complete information	
	From Agincourt HDSS (*n* = 93 507)	From Agincourt Health Centre (*n* = 10790)
First name	100.00	100.00
Surname	100.00	100.00
Other first name	35.57	6.14
Sex	100.00	99.95
Date of birth	100.00	100.00
Village	100.00	81.17
Household member first name	98.48	77.29
Household member surname	98.48	76.60
ID number	67.14	1.55
Telephone number	37.48	26.67

Figure 1 plots the sensitivity against PPV for each of the record linkage scenarios. Scenarios solely based on identifiers that are routinely collected in health facilities (S1-S6) have sensitivity ranging from 57.30% to 74.64%, and PPV ranging from 81.69% to 91.72%. Adding another household member's first name to the set of matching variables (S7-S12) considerably improved sensitivity (range: 66.13% to 81.35%) and PPV (range: 89.76% to 94.94%). However, adding another household member's last name (S13-S15) to the set of identifying variables leads to deterioration in the matching rates and accuracy. The string comparison methods that produce the best results are the JW with a threshold value of 0.9, the Double Metaphone and Soundex. Differences between these three are small. Scenarios where we consider an exact match on names or a JW score above 0.7 have a markedly lower sensitivity and PPV.

With sensitivity of 81.38% and PPV of 94.94%, scenario S12 produces the best results among the purely probabilistic linkage scenarios. Matching statistics further improve by first matching records deterministically using National ID number or telephone number and first name, and subsequently matching the remaining records with probabilistic methods using the criteria set forth in scenario S12. This hybrid record linkage approach (S16) increases sensitivity to 83.63% and PPV to 95.07%. The improvement in matching statistics is only marginal, however, and probably due to the fact that these attributes have a substantial number of missing values in either one or both datasets.

The inclusion of clerical review in the linkage process results in modest improvements in PPV. Allocating 5% of the record pairs below and above the threshold value in scenario S16 to clerical review (S17) yields the best results in terms of maximizing both sensitivity (84.27%) and PPV (96.86%). The other scenarios involving clerical review produce small gains in PPV, but are considerably more labour intensive. For example, for scenario S17,

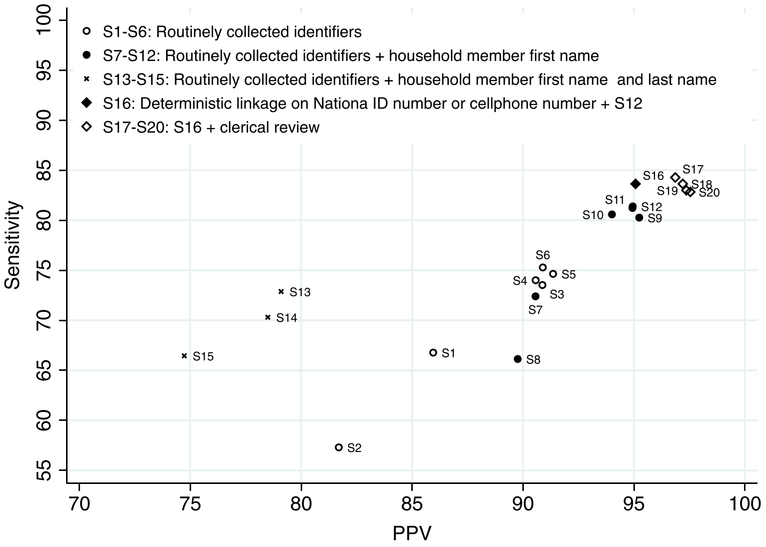


Figure 1 Sensitivity and positive predictive values (PPVs) in various linkage scenarios. See Table 1 for a description of the scenarios.

1131 record pairs were reviewed and it took the two reviewers an average of 5 hours each to complete the task whereas for scenario S20, 3492 record pairs were reviewed, which required an average of 15 hours per reviewer.

In Table 3, we present a number of background characteristics of individuals and their association with matching success. The records come from the gold standard dataset in which record pairs are matched using fingerprints, and match success in record linkage scenarios based on conventional personal identifiers is the outcome of interest. This analysis is conducted for three of the scenarios defined in Table 1: (i) the best fully automated scenario that uses only personal identifiers that are routinely collected in health facilities (S6), (ii) the best fully automated record linkage scenario based on an extended set of personal identifiers and wherein deterministic and probabilistic linkage methods are combined (S16), and (iii) S17, which is equivalent to S16 with the addition of clerical review of 5% of the record pairs with a matching score immediately above and below the threshold value.

Background characteristics associated with a lower matching likelihood in a multivariable model are female gender, old age, and low socioeconomic status (being below the highest wealth quintile). The coefficients for age indicate that matching rates deteriorate above age 50 (significantly above age 65), which suggests that reporting of personal identifiers in older respondents may not be as reliable. Being non South African is associated with lower matching success only in scenario S17 whereas having received less than primary education is associated with lower matching success in both scenarios S6 and S17. Interestingly, the scenarios that produce the best matching statistics (S16 and S17) do not necessarily produce samples of matched records that are less biased (i.e., significant predictors of matching success are similar across the three scenarios in Table 3).

Although matched and non-matched records differ in terms of some of their background characteristics, the distribution of background characteristics in the fingerprint linked dataset and the dataset generated via record linkage on conventional personal identifiers is quite similar for all the three scenarios considered here (Table 4). The reason is that the algorithms will select an individual with similar personal attributes (gender, age, etc.), even if it is not an exact match.

Discussion

We have evaluated the coverage and quality of record linkage in rural South Africa between the Agincourt HDSS and patient administration records from a health facility in its vicinity. We created a gold standard dataset of records matched by means of fingerprints and use it to evaluate the performance of 20 record linkage scenarios with conventional personal identifiers. The various record linkage scenarios can be distinguished by four attributes. First, one set of scenarios uses only personal identifiers that are routinely collected in health facilities (first name, surname, date of birth, sex and village) whereas another set of scenarios uses an extended set of identifiers (adding another household member's names, national ID number and telephone number). Second, some scenarios use purely probabilistic methods of record linkage, whereas others follow a hybrid approach where records are first matched deterministically using National ID number or telephone number and first name, and the remainder are retained for

Table 3 Background characteristics associated with successful matching in the dataset of records matched by means of fingerprints

Variable	n	Linkage scenario 6		Linkage scenario 16		Linkage scenario 17	
		Matched	Multivariable	Matched	Multivariable	Matched	Multivariable
		n (%)	OR (95% CI)	*n* (%)	OR (95% CI)	*n* (%)	OR (95% CI)
	623	492 (79.0)		551 (88.4)		552 (88.6)	
Sex							
Female	511	395 (77.3)	1	445 (87.1)	1	447 (87.5)	1
Male	112	97 (86.6)	2.86 (1.41-5.82)*	106 (94.6)	4.38 (1.52-12.61)*	105 (93.8)	3.34 (1.25-8.97)*
Age							
18-34	334	284 (85.0)	1	308 (92.2)	1	308 (92.2)	1
35-49	125	100 (80.0)	0.99 (0.53-1.84)	112 (89.6)	0.84 (0.36-1.93)	115 (92.0)	1.21 (0.5-2.92)
50-64	89	66 (74.2)	0.76 (0.35-1.66)	78 (87.6)	0.75 (0.27-2.14)	77 (86.5)	0.75 (0.27-2.12)
65+	75	42 (56.0)	0.35 (0.15-0.85)*	53 (70.7)	0.21 (0.07-0.63)*	52 (69.3)	0.25 (0.08-0.74)*
Ethnicity							
Other	96	67 (70.0)	1	76 (79.2)	1	75 (78.1)	1
South African	527	425 (80.7)	1.3 (0.71-2.37)	475 (90.1)	1.82 (0.88-3.77)	477 (90.5)	2.1 (1.02-4.33)*
Residence status							
Permanent	574	450 (78.4)	1	506 (88.1)	1	507 (88.3)	1
Temporary and other	49	42 (85.7)	1.63 (0.54-4.88)	45 (91.8)	1.28 (0.28-5.89)	45 (91.8)	1.4 (0.31-6.44)
Highest level of education							
None	97	54 (55.7)	1	71 (73.2)	1	69 (71.1)	1
Some primary	191	144 (75.4)	1.46 (0.76-2.83)	164 (85.8)	1.16 (0.51-2.63)	166 (87.0)	1.43 (0.64-3.22)
Post primary	302	267 (88.4)	2.73 (1.18-6.36)*	288 (95.4)	2.62 (0.87-7.92)	288 (95.4)	3.05 (1.01-9.24)*
Employment							
Not working	514	413 (80.4)	1	462 (89.8)	1	460 (89.5)	1
Working	93	70 (75.3)	0.68 (0.37-1.25)	79 (85.0)	0.53 (0.25-1.14)	81 (87.1)	0.71 (0.32-1.58)
Wealth quintile							
Lowest	44	28 (63.6)	1	33 (75.0)	1	34 (77.3)	1
Second	84	62 (73.8)	1.48 (0.63-3.49)	75 (89.3)	2.42 (0.84-6.98)	73 (90.0)	1.63 (0.57-4.62)
Middle	125	100 (80.0)	1.89 (0.82-4.37)	108 (86.4)	1.60 (0.6-4.25)	110 (88.0)	1.58 (0.58-4.36)
Fourth	172	136 (79.1)	1.81 (0.8-4.11)	152 (88.3)	2.08 (0.78-5.54)	150 (87.2)	1.47 (0.55-3.93)
Highest	184	159 (86.4)	2.9 (1.24-6.75)*	174 (94.5)	4.4 (1.51-12.84)*	175 (95.1)	4.03 (1.34-12.17)*
Goodness-of-fit							
Pseudo R^2, Wald χ^2 (*p*-value)		0.11, 56.89 (<0.0001)		0.16, 51.94 (<0.0001)		0.16, 53.76 (<0.0001)	

Statistical significance: * = *p*-value < 0.05.

probabilistic record linkage. Third, we use different string comparison metrics for names. Finally, we define purely automated record linkage scenarios as well as scenarios involving clerical review of a subset of record pairs.

Record linkage scenarios with the most satisfying results are those that follow a hybrid approach of deterministic followed by probabilistic record linkage, and those that use an extended set of identifiers including another household member's first name, National ID number and telephone number. Worth noting is that another household member's first name is a substantially better matching variable than his or her surname as the latter is often the same as that of the person to be linked and does not add much new information. In terms of string comparison metrics, the best results are obtained in scenarios that use a combination of Soundex, Double Metaphone and a Jaro-Winkler score above 0.9 (see also [51]).

Fully automated record linkage based on a set of personal identifiers that are routinely collected at health facilities (S6 in Table 1) has a sensitivity of 75.28% and PPV of 90.89%. The best fully automated record linkage scenario based on an extended set of identifiers and following a hybrid deterministic-probabilistic approach (S16), yields a sensitivity of 83.63% and PPV of 95.07%.

Table 4 Distribution of background characteristics in the dataset matched by means of fingerprints compared to three datasets of records matched using conventional personal identifiers

Variable	Matched on fingerprint (n = 623)	Matched with scenario 6 (n = 492)		Matched with scenario 16 (n = 551)		Matched with scenario 17 (n = 552)	
	n (%)	*n* (%)	*p*-value*	*n* (%)	*p*-value*	*n* (%)	*p*-value*
Sex							
Female	511 (82.0)	395 (80.3)		445 (80.8)		447 (81.0)	
Male	112 (18.0)	97 (19.7)	0.460	106 (19.2)	0.579	105 (19.0)	0.645
Age							
18-34	334 (53.6)	284 (57.7)		308 (55.9)		308 (55.8)	
35-49	125 (20.1)	100 (20.3		112 (20.3)		115 (20.8)	
50-64	89 (14.3)	66 (13.4)		78 (14.2)		77 (14.0)	
65+	75 (12.0)	42 (8.5)	0.240	53 (9.6)	0.601	52 (9.4)	0.528
Ethnicity							
Other	96 (15.4)	67 (13.6)		76 (13.8)		75 (13.6)	
South African	527 (84.6)	425 (86.4)	0.401	475 (86.2)	0.434	477 (86.4)	0.377
Residence status							
Permanent	574 (92.1)	450 (91.5)		506 (91.8)		507 (91.8)	
Temporary and other	48 (7.7)	42 (8.5)	0.595	45 (8.2)	0.617	45 (8.2)	0.618
Highest level of education							
None	97 (15.6)	54 (11.0)		71 (12.9)		69 (12.5)	
Some primary	191 (30.7)	144 (29.3)		164 (29.8)		166 (30.1)	
Post primary	302 (48.5)	267 (54.3)	0.098	288 (52.3)	0.491	288 (52.2)	0.426
Employment							
Not working	514 (82.5)	413 (83.9)		462 (83.4)		460 (83.3)	
Working	93 (14.9)	70 (14.2)	0.660	79 (14.3)	0.643	81 (14.7)	0.795
Wealth quintile							
Lowest	44 (7.1)	28 (5.7)		33 (6.0)		34 (16.2)	
Second	84 (13.5)	62 (12.6)		75 (13.6)		73 (13.2)	
Middle	125 (20.1)	100 (20.3)		108 (19.6)		110 (19.9)	
Fourth	172 (27.6)	136 (27.6)		152 (27.6)		150 (21.2)	
Highest	184 (29.5)	159 (32.3)	0.753	174 (31.58)	0.912	175 (31.7)	0.952

**p*-value using chi-squared test computed separately for records in each scenario compared to records matched by means of fingerprints.

The sensitivity and PPV increase to 84.27% and 96.86%, respectively, when clerical review is performed on 10% of the record pairs around the matching score threshold of scenario S16. Even though these results are very encouraging, it is likely that they could be improved further by more comprehensive collection of National ID numbers and telephone numbers in both the Agincourt HDSS and the health facility.

Matching rates are significantly worse for women (compared to men), for former Mozambican refugees (compared to native South Africans), and for the poorly educated and older respondents. The association between these background characteristics and matching rates is similar in all record linkage scenarios, irrespective of their sensitivity and PPV. The lower matching success for women may be because some of them change names upon marriage and may be known by their husband's name in one data source and registered under their maiden name in another data source. As for older respondents, the lower matching success could be a result of poorer reporting with age or an effect of older generations not having accurate information on some of their identifiers such as date of birth. The lower matching success for Mozambicans could be related to their legal status, but we have no means of verifying this. These analyses of the individual-level correspondence in matching success are thus indicative of systematic bias in all of the record linkage scenarios considered here. It is also worth noting, however, that the distributions of socio-demographic background characteristics in the gold standard and record-linked datasets are very similar, which suggests that record-linked datasets may still be used for assessing equitable uptake of services.

Conclusion

Using records matched by means of fingerprints as the gold standard, we have demonstrated the feasibility of fully automated probabilistic record linkage using identifiers that are routinely collected in health facilities in South Africa. Our study also shows that matching statistics can be improved if other identifiers (e.g., another household member's first name) are added to the set of matching variables, and, to a lesser extent, with clerical review. Matching success is, however, correlated with background characteristics that are indicative of the instability of personal attributes over time (e.g., surname in the case of women) or with misreporting of attributes (e.g., age).

Endnotes

[a]Some HDSS that have been built around a health facility or manage a health facility as part of their research operation (e.g., the Kilifi HDSS or the Masaka HDSS). In these studies, the data systems are well integrated.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CWK and GR designed and executed the study, conducted the analyses, and wrote the first draft of the manuscript. BDC and FXGO provided assistance with programming and the organization of the fieldwork. All co-authors made substantial contributions to the study design and manuscript preparation. All authors approved the final version of the manuscript.

Acknowledgments

This study was supported by grants from the National Institute on Aging (NIA, R24AG032112); the Wellcome Trust, UK (085477/Z/08/Z); the Bill and Melinda Gates Foundation (BMGF- OPP1082114); and the Anglo-American Chairman's Fund. We acknowledge the support of the MRC/Wits Rural Public Health and Health Transitions Unit (Agincourt), University of the Witwatersrand; SAP Meraka Unit for Technology Development; Staff of Agincourt Health Centre, our field team for the community-based fingerprint collection, and participating communities. We are also grateful to Ryan Wagner and Rhian Twine for their thoughtful review and feedback on this manuscript.

Author details

[1]MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), School of Public Health, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [2]Department of Ecology, Evolution and Environmental Biology, Columbia University, New York, USA. [3]Umeå Centre for Global Health Research, Division of Epidemiology and Global Health, Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. [4]INDEPTH Network, Accra, Ghana. [5]Institute of Behavioral Science, University of Colorado, Boulder, Colorado, USA. [6]Department of Population Health, London School of Hygiene and Tropical Medicine, London, UK.

Received: 7 February 2014 Accepted: 16 May 2014
Published: 24 May 2014

References

1. Sankoh O: **Global health estimates: stronger collaboration needed with low-and middle-income countries.** *PLoS Med* 2010, **7**(11):e1001005.
2. Sankoh O, Byass P: **The INDEPTH Network: filling vital gaps in global epidemiology.** *Int J Epidemiol* 2012, **41**(3):579–588.
3. Ye Y, Wamukoya M, Ezeh A, Emina J, Sankoh O: **Health and demographic surveillance systems: a step towards full civil registration and vital statistics system in sub-Sahara Africa?** *BMC Public Health* 2012, **12**(1):741.
4. Jahn A, Floyd S, Crampin AC, Mwaungulu F, Mvula H, Munthali F, McGrath N, Mwafilaso J, Mwinuka V, Mangongo B, Fine PEM, Zaba B, Glynn JR: **Population-level effect of HIV on adult mortality and early evidence of reversal after introduction of antiretroviral therapy in Malawi.** *Lancet* 2008, **371**(9624):1603–1611.
5. Herbst AJ, Mafojane T, Newell ML: **Verbal autopsy-based cause-specific mortality trends in rural KwaZulu-Natal, South Africa, 2000–2009.** *Popul Health Metrics* 2011, **9**(1):47.
6. Bor J, Herbst AJ, Newell M-L, Bärnighausen T: **Increases in adult life expectancy in rural South Africa: valuing the scale-up of HIV treatment.** *Science* 2013, **339**(6122):961–965.
7. Jutte DP, Roos LL, Brownell MD: **Administrative record linkage as a tool for public health research.** *Annu Rev Public Health* 2011, **32:**91–108.
8. Holian J, Mallick MJ, Zaremba CM: **Maternity and infant care, race and birth outcomes.** *J Health Soc Policy* 2004, **18**(4):1–11.
9. Holian J: **Live birth and infant death record linkage.** *J Health Soc Policy* 2000, **12**(1):1–10.
10. Lyons R, Jones K, John G, Brooks C, Verplancke J-P, Ford D, Brown G, Leake K: **The SAIL databank: linking multiple health and social care datasets.** *BMC Med Inform Decis Mak* 2009, **9**(1):3.
11. Amin J, Law MG, Bartlett M, Kaldor JM, Dore GJ: **Causes of death after diagnosis of hepatitis B or hepatitis C infection: a large community-based linkage study.** *Lancet* 2006, **368**(9539):938–945.
12. Falster K, Wand H, Donovan B, Anderson J, Nolan D, Watson K, Watson J, Law MG: **Hospitalizations in a cohort of HIV patients in Australia, 1999–2007.** *AIDS* 2010, **24**(9):1329.
13. Holman CAJ, Bass AJ, Rouse IL, Hobbs MS: **Population-based linkage of health records in Western Australia: development of a health services research linked database.** *Aust N Z J Public Health* 2008, **23**(5):453–459.
14. Franceschi S, Dal Maso L, Arniani S, Crosignani P, Vercelli M, Simonato L, Falcini F, Zanetti R, Barchielli A, Serraino D, Rezza G: **Risk of cancer other than Kaposi's sarcoma and non-Hodgkin's lymphoma in persons with AIDS in Italy. Cancer and AIDS Registry Linkage Study.** *Br J Cancer* 1998, **78**(7):966–970.
15. Chen J, Fair M, Wilkins R, Cyr M: **Maternal education and fetal and infant mortality in Quebec. Fetal and Infant Mortality Study Group of the Canadian Perinatal Surveillance System.** *Health Reports/Statistics Canada, Canadian Centre for Health Information* 1998, **10**(2):53–64.
16. Tromp M, Meray N, Ravelli ACJ, Reitsma JB, Bonsel GJ: **Medical record linkage of anonymous registries without validated sample linkage of the Dutch perinatal registries.** *Stud Health Tech Informat* 2005, **116:**125–130.
17. Chard T, Penney G, Chalmers J: **The risk of neonatal death in relation to birth weight and maternal hypertensive disease in infants born at 24–32 weeks.** *Eur J Obstet Gynecol Reprod Biol* 2001, **95**(1):114–118.
18. Cole S: **History of fingerprint pattern recognition.** In *Automatic Fingerprint Recognition Systems.* Edited by Ratha N, Bollle R. New York: Springer; 2004:1–25.
19. Kahn K, Collinson MA, Gómez-Olivé FX, Mokoena O, Twine R, Mee P, Afolabi SA, Clark BD, Kabudula CW, Khosa A, Khoza S, Shabangu MG, Silaule B, Tibane JB, Wagner RG, Garenne ML, Clark SJ, Tollman SM: **Profile: Agincourt Health and Socio-demographic Surveillance System.** *Int J Epidemiol* 2012, **41**(4):988–1001.
20. Kahn K, Tollman SM, Collinson MA, Clark SJ, Twine R, Clark BD, Shabangu M, Gomez-Olive FX, Mokoena O, Garenne ML: **Research into health, population and social transitions in rural South Africa: Data and methods of the Agincourt Health and Demographic Surveillance System.** *Scand J Publ Health* 2007, **35**(69 suppl):8–20.
21. Serwaa-Bonsu A, Herbst AJ, Reniers G, Ijaa W, Clark B, Kabudula C, Sankoh O: **First experiences in the implementation of biometric technology to link data from Health and Demographic Surveillance Systems with health facility data.** *Glob Health Action* 2010, **3:**2120.
22. CSIR: **Using ICT to support rural clinics in managing chronic lifestyle diseases.** In *Sciencescope.* Pretoria: CSIR; 2009:57–58.
23. **MorphoSmart Overview.** [https://www.yumpu.com/en/document/view/10855857/morphosmarttm-overview]
24. Li B, Quan H, Fong A, Lu M: **Assessing record linkage between health care and Vital Statistics databases using deterministic methods.** *BMC Health Serv Res* 2006, **6**(1):48.

25. Machado CJ: **A literature review of record linkage procedures focusing on infant health outcomes.** *Cadernos de Saúde Pública* 2004, **20:**362–371.
26. Maso LD, Braga C, Franceschi S: **Methodology used for software for automated linkage in Italy (SALI).** *Comput Biomed Res* 2001, **34**(6):395.
27. Victor TW, Mera RM: **Record linkage of health care insurance claims.** *J Am Med Inform Assoc* 2001, **8**(3):281–288.
28. Howe GR: **Use of computerized record linkage in cohort studies.** *Epidemiol Rev* 1998, **20**(1):112–121.
29. Beauchamp A, Tonkin AM, Kelsall H, Sundararajan V, English DR, Sundaresan L, Wolfe R, Turrell G, Giles GG, Peeters A: **Validation of de-identified record linkage to ascertain hospital admissions in a cohort study.** *BMC Med Res Methodol* 2011, **11**(1):42.
30. Cook L, Olson L, Dean J: **Probabilistic record linkage: relationships between file sizes, identifiers, and match weights.** *Methods Inf Med* 2001, **40**(3):196–203.
31. Jaro MA: **Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida.** *J Am Stat Assoc* 1989, **84**(406):414–420.
32. Jaro MA: **Probabilistic linkage of large public health data files.** *Stat Med* 1995, **14**(5–7):491–498.
33. Nitsch D, Morton S, DeStavola BL, Clark H, Leon DA: **How good is probabilistic record linkage to reconstruct reproductive histories? Results from the Aberdeen children of the 1950 s study.** *BMC Med Res Methodol* 2006, **6**(1):15.
34. Clark D: **Practical introduction to record linkage for injury research.** *Inj Prev* 2004, **10**(3):186.
35. Pacheco AG, Saraceni V, Tuboi SH, Moulton LH, Chaisson RE, Cavalcante SC, Durovni B, Faulhaber JC, Golub JE, King B, Schechter M, Harrison LH: **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**(11):1326–1332.
36. Rosman DL: **The feasibility of linking hospital and police road crash casualty records without names.** *Accid Anal Prev* 1996, **28**(2):271–274.
37. Winkler WE: **String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.** In *Proceedings of the Section on Survey Research Methods.* Alexandria: American Statistical Association; 1990:354–359.
38. Philips L: **The double metaphone search algorithm.** *C Plus Plus Users J* 2000, **18**(6):38–43.
39. Méray 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**(9):883–e881.
40. Grannis SJ, Overhage JM, Hui S, McDonald CJ: **Analysis of a probabilistic record linkage technique without human review.** *J Am Med Informat Assoc* 2003, **2003:**259–263.
41. Herzog TN, Scheuren F, Winkler WE: *Data quality and record linkage techniques.* Heidelberg: Springer; 2007.
42. Fellegi IP, Sunter AB: **A theory for record linkage.** *J Am Stat Assoc* 1969, **64**(328):1183–1210.
43. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB: **Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage.** *J Clin Epidemiol* 2011, **64**(5):565–572.
44. Karmel R, Rosman D: **Linkage of health and aged care service events: comparing linkage and event selection methods.** *BMC Health Serv Res* 2008, **8**(1):149.
45. Blakely T, Salmond C: **Probabilistic record linkage and a method to calculate the positive predictive value.** *Int J Epidemiol* 2002, **31**(6):1246–1252.
46. Bentley JP, Ford JB, Taylor LK, Irvine KA, Roberts CL: **Investigating linkage rates among probabilistically linked birth and hospitalization records.** *BMC Med Res Methodol* 2012, **12**(1):149.
47. Megan B, Damien J, Vijaya S, Sue E, David P, Ian S, Caroline B: **Data Linkage: A powerful research tool with potential problems.** *BMC Health Serv Res* 2010, **10:**346.
48. Kolenikov S, Angeles G: **Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer?** *Rev Income Wealth* 2009, **55**(1):128–165.
49. **SimMetrics.** [http://sourceforge.net/projects/simmetrics]
50. Implement Phonetic ("Sounds-like") Name Searches with Double Metaphone Part V: **NET Implementation.** [http://www.codeproject.com/Articles/4624/Implement-Phonetic-quot-Sounds-like-quot-Name-Sear]
51. Snae C: **A comparison and analysis of name matching algorithms.** *Int J Appl Sci Eng Technol* 2007, **4**(1):252–257.

doi:10.1186/1471-2288-14-71
Cite this article as: Kabudula *et al.*: **The promise of record linkage for assessing the uptake of health services in resource constrained settings: a pilot study from South Africa.** *BMC Medical Research Methodology* 2014 **14**:71.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Mr CW Kabudula

CLEARANCE CERTIFICATE	**M120488**
PROJECT	Tracking Changes in Mortality Patterns and Associated Socioeconomic Diffrentials in Rural South Africa
INVESTIGATORS	Mr CW Kabudula.
DEPARTMENT	School of Public Health
DATE CONSIDERED	04/05/2012
+DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 04/05/2012 **CHAIRPERSON** ....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable
cc: Supervisor : Prof S Tollman

DECLARATION OF INVESTIGATOR(S)
To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...

2012-06-09

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

FACULTY OF
HEALTH SCIENCES

PLAGIARISM DECLARATION TO BE SIGNED BY ALL HIGHER DEGREE STUDENTS

SENATE PLAGIARISM POLICY: APPENDIX ONE

I _*Chodziwadziwa Whiteson Kabudula*_(Student number: _369065______) am a student registered for the degree of Doctor of Philosophy in the academic year __2017_______.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where I have explicitly indicated otherwise.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: ________________________________ Date:__20th December 2017___________