Addressing Health Inequalities in South Africa

Policy Insights and the Role of Improved Efficiency

Eyob Zere Asbu

Thesis Presented for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Public Health and Primary Health Care

UNIVERSITY OF CAPE TOWN

August 2002

DECLARATION

I, **Eyob Zere Asbu**, hereby declare that the work on which this thesis is based is my original work (except where acknowledgements indicate otherwise), and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the University of Cape Town to reproduce for the purpose of research either the whole or any portion of the contents in any matter whatsoever.

Eyob Zere Asbu

August 2002

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES AND BOXES

LIST OF ACRONYMS

AIDS	Acquired immune deficiency syndrome
ALS	Average length of stay
ANC	African National Congress
BTR	Bed turnover ratio
CMR	Child mortality rate
CSS	Central Statistical services
DOH	Department of Health
GDP	Gross Domestic Product
GEAR	Growth, Employment and Redistribution Strategy
HDI	Human development index
HIV	Human immune deficiency Virus
HST	Health Systems Trust
IES	Income and expenditure survey
IMR	Infant mortality rate
LSDS	Living standards and development survey
ML	Maximum likelihood
MMR	Maternal mortality ratio
MPI	Malmquist productivity index
PDE	Patient day equivalent
PHC	Primary Health Care
PPP	Purchasing power parity
PYLL	Potential years of life lost
QALY	Quality-adjusted life year
RDP	Reconstruction and Development Programme
RII	Relative index of inequality

RSA	Republic of South Africa
SA	South Africa
SADHS	South Africa Demographic and Health Survey
SALDRU	South African Labour and Development Research Unit
TFP	Total factor productivity
U5MR	Under-five mortality rate
UNDP	United Nations Development Programme
UNICEF	United Nations Children's Fund
WHO	World Health Organization

ACKNOWLEDGEMENTS

I am profoundly indebted to Professor Diane McIntyre, Director of the Health Economics Unit, University of Cape Town, for her superb supervision of my dissertation from its inception up to the end.

Special appreciation is due to all my family members and friends for their continuous moral support throughout my study period. The staff of the Health Economics Unit, University of Cape Town, deserve special mention for their wonderful support.

I gratefully acknowledge the University of Cape Town for its partial financial assistance. The World Institute for Development Economics Research, United Nations University, is also acknowledged for granting me the opportunity of a PhD internship.

I am solely responsible for any errors of omission or commission.

Eyob Zere Asbu

ABSTRACT

This study attempts to assess the equity and technical efficiency aspects of the South African health system. It empirically assesses the *status quo* and trends in equity as it relates to child morbidity and mortality and self-reported illness and utilization of different service providers in adulthood. Furthermore, an assessment of the technical efficiency and productivity of a sample of public sector hospitals is conducted. This is meant to explore the size of potential efficiency gains that is tantamount to the injection of additional resources, which are highly needed for addressing inequities in a scenario where mobilization of additional resources from the public purse is seriously constrained as a result of poor economic performance, stringent fiscal policies and competing priorities, among other things.

Secondary data are used in the analyses. These include data from the Living Standards and Development Survey (LSDS) of 1993, conducted jointly by the World Bank and the South African Labour and Development Research Unit at the University of Cape Town, and data from the October Household Survey (OHS) series (OHS 1995 and OHS 1998) that are conducted annually by Statistics South Africa. For the analysis of hospital efficiency, data are obtained from annual statistical publications of provincial health departments. The equity analysis is done using concentration indices (and curves). In the adult population, standardized concentration indices are computed to rule out a possible confounding effect of the demographic variables, age and gender. Furthermore, utilization of services is standardized for need as measured by self-reported acute or chronic illness. Additionally, to identify some factors, which may be associated with inequities in child health, probit models are estimated. Data envelopment analysis (DEA) and DEA-based Malmquist productivity index are used to examine the state of hospital technical efficiency and productivity respectively. With the limited data available a tobit regression is also run to identify factors influencing the technical efficiency of hospitals.

Overall, the findings of this study indicate that the huge income-related inequalities in health and health care that existed prior to the change of the political system in 1994 have been reduced significantly in the years after the installation of the new government. Analyses of the LSDS 1993 indicate significant pro-rich inequities in all the dimensions of equity in health and health care utilization examined in this study. Under-five mortality and child malnutrition manifest pro-rich inequalities of high magnitude. In the adult population, as is seen in many other studies, pro-poor inequities are seen in self-reported acute illness. This paradoxical pro-poor finding is, however, changed to pro-rich inequalities in the OHS 1995 and 1998 data.

Inequalities in under-five mortality in the OHS 1998 data that do not show when income is used as a measure of socio-economic status (SES) are prominently seen when SES is proxied by race and residential location. This implies that the apparent bridging of inequities seen when income is used as a measure of SES may not enable us to definitively assert the absence of socio-economic inequities in health.

Utilization statistics from all data sets indicate pro-poor horizontal inequities in the use of primary and other public health facilities, implying an appropriate targeting of public sector health care resources.

The data clearly show that considerable health and health system inequities remain in South Africa. In order to rapidly address these inequities, additional resources are required to improve health and other health-promoting services in currently under-served areas and for specific disadvantaged groups. However, given the macro-economic context, the allocation of additional resources to the health sector is unlikely.

The hospital sector, which absorbs the lion's share of the public health resources, seems to be plagued by high degrees of technical inefficiency. With the prevailing high levels of technical inefficiency and the adverse economic realities of the country, it would be difficult to mobilize additional resources needed for addressing existing inequities. Hence it is of paramount importance to address the existing technical inefficiencies in the hospital sector.

Finally the study recommends that to address the inequities that besiege the country's health system, policies that transcend the health sector are needed and that there is an urgent need to rectify existing inefficiencies.

CHAPTER 1
OVERVIEW OF THE STUDY

1.1. INTRODUCTION

Most countries in sub-Saharan Africa inherited from their colonial masters health systems that were grossly inefficient and inequitable. The systems primarily catered for the minority and predominantly focussed on urban-based, high technology hospital care. Over the years since independence, equity and efficiency have been accorded prominent places in their health policies. However, to date, the achievement of these goals has been extremely limited.

The case of South Africa is no different from the above scenario. The legacy of apartheid policies is typically characterized by the prevalence of high degrees of inequality in terms of income and access to resources. There are also indications that inefficiencies –both technical and allocative – have been among the major maladies of the system (see for example McMurchy 1996). Thus post-apartheid South Africa's health system, plagued by these problems, has to toil to minimize the legacies of a socially unjust political regime.

Systematic inequalities are not only limited to income and access to resources. There are also high levels of inequality in the country's health and development indicators. Life expectancy, and rates of mortality at all stages of the life-course show inequalities between the different population/race groups. Human development indices for the various population groups show the dual nature of the South African society, where characteristics of developing and developed worlds co-exist. Cognizant of the problems besieging the country's health system, the new government that came to power in 1994 put health equity high on its agenda of social development.

In a country such as South Africa that is characterized by high levels of inequalities in health, access to health care, and high levels of income inequality and poverty, the task of redressing inequities through levelling up rather than levelling out is preferred. A levelling up approach is required if the apartheid legacy of massive inequities is to be addressed in the short- to medium-term rather than over many decades. This requires the injection of significant amounts of additional resources into the health system (it is assumed that the role of health care is to improve health and reduce inequalities in health), as the restructuring necessary to resolve inequities through levelling out (redistribution of currently available resources) takes considerable time, or if done quickly, may destroy the existing health service infrastructure. In addition to political commitment, this presupposes the existence of a vibrant economy.

The availability of additional resources for the health system in order to redress past inequities is, however, seriously constrained given the country's macro-economic environment. Economic growth has been slackening, and the government's macroeconomic restructuring plan, the GEAR (Growth, Employment and Redistribution) Strategy, which advocates for deficit reduction implies that a substantial increase in public spending is an unlikely event. To make matters worse, the HIV/AIDS epidemic is straining the meagre resources for health care that are already overstretched. Resurging diseases such as malaria and acute diarrhoeal diseases (e.g. cholera) are also taking their toll on resources. By and large, resources needed to redress inequities in health and health care are seen to be dwindling. In such a situation, it is of paramount importance to explore the possibilities of reaping efficiency gains that may sometimes be comparable to the injection of huge amounts of additional health care resources.

In the era of evidence-based policy it is imperative to have hard facts on the *status quo* regarding equity and efficiency of the system. The magnitude of systematic inequalities in

health and health care, as well as inefficiencies in South Africa's health system, has not been studied extensively using robust quantitative methods. Hence, there is a dire need to have a quantitative measurement of the magnitude, determinants and trends of equity and efficiency to contribute to the existing evidence base that is necessary to guide the formulation of effective policies and interventions This forms the rationale for undertaking an empirical study of the equity and efficiency aspects of the South African health system.

To this end it is important to answer such questions as:

1. Are there significant systematic inequalities in health and health care related to the socio-economic status of a person at the different stages in the life course?
2. What is the nature and magnitude of these inequities?
3. What are some of the factors that are likely to exacerbate or mitigate the inequities?
4. What has been the trend of these systematic inequalities over the years?
5. What is the potential to release resources from efficiency savings from within the health system to contribute towards the endeavours to redress inequities? What is the nature and form of technical inefficiency prevalent, and what are its correlates? Was a total factor productivity growth seen over the years after the installation of the new government?

These are the main issues that this study attempts to address. Answers to these questions are expected to generate insights that may contribute towards the practice of evidence-based health equity and efficiency policies. It should at the outset be noted that the study does not attempt to examine the trade-offs between equity and efficiency in consumption. The analysis of technical efficiency (efficiency in production) and productivity is included to illustrate the

fact that within the context of Africa, the health sector itself can be a source of substantial amounts of resources that are needed to enhance the redress of inequities.

1.2. AIM AND OBJECTIVES

1.2.1. AIM

The aim of this study is to undertake an empirical assessment of the *status quo*, trends and influencing factors of equity and efficiency in the South African Health System using robust, state-of-the-art techniques of measurement. This is done with a view to supplement efforts aimed at generating an evidence-base for resource allocation policies and interventions that are needed to redress the huge backlogs of inequity inherited from an unjust political system.

1.2.2. SPECIFIC OBJECTIVES

The specific objectives of the study are to:

i. Assess the magnitude, trends and determinants of equity in child survival as measured by infant and under-five mortality;

ii. Assess the magnitude, trend and influencing factors of equity in child health as measured by under-five child malnutrition: stunting, underweight and wasting;

iii. Evaluate equity in self-reported adult illness and assess the degree of avoidable inequalities;

iv. Examine the horizontal equity of service utilization, that is, whether people in equal need receive equal treatment or not;

v. Estimate the state of technical efficiency and its correlates (both pure technical and scale efficiency) of a sample of public sector hospitals with a view to assessing the extent of potential efficiency gains that may be ploughed back into the system to contribute towards efforts to redress inequities; and

vi. Investigate the total factor productivity of a sample of public sector hospitals and assess trends in efficiency and technology.

vii. Consider the possible policy implications of the evidence on equity and technical efficiency.

1.3. SIGNIFICANCE OF THE STUDY

This study makes a number of important new contributions to the analysis of equity and technical efficiency in the South African health sector, particularly in the following areas:

i. Examination of the state of equity in health will help to quantify the degree of inequity in health existing at the various stages of a person's life cycle using both measures of morbidity and mortality. The technique of measurement used (concentration curve and index) produces a *composite* measure of overall inequity, which is easy to comprehend and undertake comparisons over two time periods to evaluate the effects of policies and interventions in an objective and robust way. This is a case of quantitative policy analysis that helps intensify the move towards evidence-based health policy. To date, most of the studies on health equity conducted in South Africa have used partial measures of inequality such as rate-ratios, which try to compare only two categories of people at a time (e.g., income quintile 1 versus income quintile 5), disregarding the categories in the middle. The consequence of the neglect of the middle groups is that we don't get a comprehensive picture of the systematic inequality, and therefore, our policy evaluation may not reflect the whole reality.

ii. The multivariate analysis of some of the factors, which influence the state of equity, will help to identify a set of variables that can be manoeuvred as policy levers in strategies to combat inequity. Most studies to date in South Africa have used univariate analyses. In situations where experimental is scarce for varied reasons, the

use of multivariate techniques is valuable to validate the findings and prescriptions of the univariate methods.

iii. The component on equity in adult illness tries to measure inequities after taking into account the confounding effect of age and sex, that is it employs age-sex standardized measures of inequity. Furthermore the assessment of equity in service utilization links utilization of services/providers to *need* as measured by self-reported illness with the objective of measuring horizontal inequity. This implies assessing whether people in equal need get equal treatment or not. Again studies previously conducted in South Africa have not excluded the confounding effect of age and sex, and thus, it is possible that in some situations in which inequities are regarded to exist, it may purely be the effect of age and sex. Furthermore, if the need for health care is not taken into account, it may not be possible to obtain a reliable measure of inequity in utilization. Most of the available studies in South Africa are lacking in this aspect.

iv. Although the concentration index (and curve) has been used relatively widely in measuring equity in health in developed countries as well as some developing countries, its use in sub-Saharan Africa is at a very nascent stage. Hence, its application to South African data will contribute to the impetus to use the technique within the context of African health systems.

v. The measurement of the technical efficiency of hospitals, which consume the lion's share of health care resources, will help bring into focus the fact that efficiency gains could be regarded as a significant source of scarce resources in a system whose resource constraints are mounting by the day. Although admittedly there may be an absolute lack of resources in the public health sector, the relative lack of resources

that stems from inefficiency in production is expected to be equally important. Currently, the latter is in most instances not noticed. Hence, the study will demonstrate with empirical evidence that the injection of more and more additional resources will not satiate the system's demand for resources if the system's efficiency is not improved.

vi. The technique of data envelopment analysis (DEA) has gained currency in assessing the efficiency of decision-making units, especially those in the public sector whose operation is not dictated by the motive of profit maximization. Its use in scrutinizing the efficiency of health facilities is also increasing. However, to date, the application of this methodology to developing country health systems, particularly in sub-Saharan Africa is extremely rare. Most health system efficiency studies use ratio analyses, which can only assess efficiency in a piecemeal manner. Ratios do not give a comprehensive view of efficiency, and cannot reliably indicate the amount of potential efficiency savings and the re-organization of production technology that is required to operate efficiently. Hence, this study, by demonstrating the merits of this frontier method of analysis (i.e. DEA) will pave the way for more use of techniques of microeconomic efficiency.

1.4. ORGANIZATION OF THE REPORT

The remaining sections of the dissertation are organized as follows:

Chapter 2 presents a profile of the study country. Health systems form an integral part of a country's macro-economic system, and therefore are influenced by decisions taken by other sectors of the economy. The chapter thus attempts to review the socio-economic, demographic and political context within which the South African health system operates. Furthermore, an overview of the sector itself, including the epidemiological profile, and health

care financing and delivery is given. This is expected to provide a wider perspective to the research problem.

Chapter 3 discusses in detail the concept of equity, its various definitions, philosophical foundations and issues of measurement. This will help readers understand the definitions of equity used as well as issues involved in measurement in the empirical analyses to follow in the next three chapters. This chapter is important as it lays the conceptual and analytical framework for the equity-related empirical work.

Chapter 4 assesses equity in child survival, through examining systematic inequalities related to the very sensitive indicators of infant and under-five mortality. Concentration indices are computed after the study population is classified using a measure of socio-economic status. Furthermore, a probit model is estimated to identify some factors that may impinge on equity in mortality in early life.

Chapter 5 examines equity in early life with reference to under-five child malnutrition and its correlates. It is felt that mortality alone may not capture the whole picture of inequities in childhood, as it may be possible to find the absence of systematic inequalities in mortality. However, even if children are spared from death, their quality of health/life may be gravely compromised. An arrested growth and development during this period is likely to perpetuate inequalities in income and health in later adult life, thus attenuating government initiatives to redress socio-economic inequalities inherited from the past. This chapter therefore will help investigate equity from different perspectives.

Chapter 6 deals with the assessment of inequities in adult health. Inequities may manifest differently at the various stages in the life-course. To have a relatively comprehensive picture,

it is important to examine the issue at all stages of the life cycle. The chapter investigates inequities in self-reported acute and chronic adult illness using both standardized and unstandardized concentration indices, and measures the extent of avoidable inequality. Furthermore, equity in utilization of services and different health providers is measured. In particular, the extent of horizontal inequities is highlighted.

Chapter 7 discusses the microeconomics and measurement of technical efficiency and productivity. It specifically focuses on the techniques of Data Envelopment Analysis (DEA) and the DEA-based Malmquist productivity index, and the issues surrounding the measurement of productive efficiency and productivity.

Chapter 8 deals with analysing the technical efficiency and productivity of a sample of public sector hospitals using data envelopment analysis and the DEA-based Malmquist productivity index (MPI). This is intended to demonstrate that the often-neglected efficiency savings can also be a significant source of resources from within the health system. Endeavours to redress backlogs of inequities in health and health care have immense resource requirements, in addition to the need for the re-allocation of resources. The injection of sizable amounts of resources is, however, a tough task given the country's macro-economic situation. This chapter will therefore highlight the potential for efficiency savings to supplement government financing initiatives. Analysis of productivity is done using panel input and output data for a sample of Western Cape hospitals. The MPI, unlike partial measures of productivity, gives a measure of total factor productivity (TFP), thus making it the preferred option in a multiple input/output hospital industry. It also has the appeal that it enables the decomposition of change in productivity into its component parts: efficiency change and technical change.

The final chapter of the dissertation provides a summary of the findings, conclusion and policy-relevant recommendations.

CHAPTER 2
PROFILE OF THE STUDY COUNTRY

2.1. SOCIO-DEMOGRAPHIC PROFILE

South Africa (SA) is a country in Sub-Saharan Africa (SSA), situated at the southern tip of the continent. It has a surface area of a little over 1.2 million square kilometres and is administratively divided into nine provinces. The population for 2001 is estimated at about 44.3 million (Day and Gray 2002). Considering the 1996 population of about 40.6 million (Statistics South Africa 1996), an annual average growth rate of 1.8 percent is calculated. Decomposition by population group shows that Africans comprise about 77 percent of the population, the remaining 23 percent being shared by the other three population groups, namely, Whites (10.4 percent), Coloureds (8.7 percent) and Indians (2.5 percent) (Day and Gray 2002).

South Africa is a country that has recently emerged from decades of apartheid rule, which was characterised by a polarised human development of the various population groups. Table 2.1 presents the basic health and development indicators of SA compared to those of Sub-Saharan Africa (SSA) and other upper middle-income countries with which SA's economic performance is comparable.

Table 2.1
Health and development indicators: SA, SSA and other upper middle-income countries[1]

Indicator	SA	Other SSA	Other upper middle income countries
GNP per capita 1999 (US$)	3,160	500	4,900
Average annual growth of GNP per capita, 1998-99 (%)	-0.9	-0.3	0.7
Total fertility rate (1996)	2.9	5.6	2.6
Life expectancy at birth, 1998 (M/F), (years)	61/66	49/52	67/74
Population with access to sanitation facilities, 1995 (%)	46	37	64
Population with access to safe water, 1995 (%)	70	45	76
Adult illiteracy rate (% of people 15 and above) 1998 (M/F)	15/16	32/49	9/11
Under-five mortality rate per 1000 (1998)	83	151	35

Source: World Bank (1998a, 2000)

As can be seen from the above table, SA's social and economic indicators surpass by far the averages for SSA. However, comparison with other countries classified as upper middle income reveals that SA's indicators are, in most cases, lagging behind its reference group.

The aggregated statistics for South Africa cited in the above table do not clearly indicate the disparity existing between the various population groups or geographical localities. The gap in standards of living and socio-economic indicators between the various population groups and geographical localities is so large that SA harbours characteristics of both the developed and developing worlds. Any social policy, including health, which does not take account of these differentials, will not bring about an improvement in the lot of the overwhelming majority.

The infant mortality rate (IMR) of Whites is 11.4 (HST 2000), a figure which is about a third of that of the upper middle income countries with which South Africa is classified and a little higher than that of the high income countries (in 1996, the average IMRs for the upper middle and high income countries respectively were 31 and 6) (World Bank 1999). On the other hand,

[1] The World Bank classifies countries into three income groups according to 1990 GNP per capita: low income (US$ 755 or less); lower middle-income (US$ 756 – 2,995); upper middle-income (US$ 2,996 – 9,265), and high income (US$ 9,266 or more

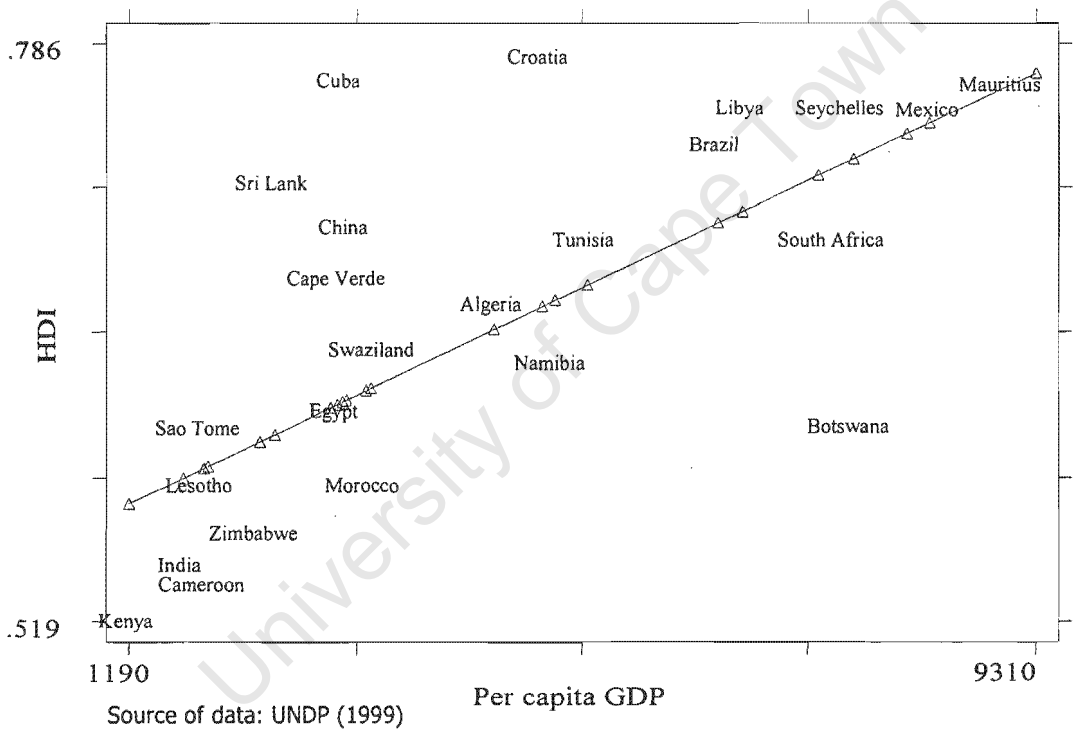
the IMR of Africans is 47.0 (HST 2000), a figure which is more than four times that of their White counterparts and which is significantly higher than the average for the lower middle income countries (World Bank 1999). In 1989, the maternal mortality rate (MMR) per 100,000 was 8.0 for Whites and 58.0 for Africans. Similarly, while the life expectancy of White women was 77 years, it was only 67 years for Africans.

In 1996, South Africa ranked 100th on the human development index (HDI)[2] scale, with an HDI of 0.649. However, when this was disaggregated by province, the Northern province had an HDI of 0.450 that would be equivalent to the country ranked 134th. On the other hand, the Western Cape province, where only 17 percent of the population was black, had an HDI value of 0.791, equivalent to the country ranked 62nd (United Nations Development Programme 1996). Decomposition of the HDI by population group reveals the duality of the South African society in terms of human development. In 1991 the HDI figures for the African, Coloured, Indian and White population groups respectively were: 0.500, 0.663, 0.836 and 0.901 (Central Statistical Services 1998). The indices for the Indians and the Whites fall under the classification of *high human development*, a group that includes all the developed countries. In contrast, those of the Africans and coloureds fall under the group of *medium human development* (UNDP 1999).

Even though SA falls into the group of countries classified as *medium human development*, its performance in terms of the HDI is not in keeping with its real GDP per capita. The difference between its rank in real GDP per capita and rank in HDI (-41) points to this fact (*ibid*). Countries with such a discrepancy between their GDP per capita and HDI ranks are less successful in translating income into better lives for their people (UNDP 2000). This is an

important indication of the pervasiveness of a significant degree of inefficiency and inequity in the macro-economy. This also implies that countries in its reference group (whose GDP per capita may sometimes be significantly less than that of SA) excel over SA in translating income into human development as can be seen from the figure below.

Figure 2.1
Per capita GDP Vs HDI: SA's performance
relative to those of 'medium human development index' countries



Source of data: UNDP (1999)

Spatial disparities within South Africa also exist in the human development indicators. To illuminate the extent of the spatial differences it is worth presenting some data as depicted in Table 2.2.

[2] The Human Development Index (HDI) is a composite measure composed of the GNP per capita, longevity, level of literacy and school enrolment. It measures average achievements in basic human development. It ranges from zero to one.

Table 2.2
Selected Human Development Indicators by Province

Province	Indicator					
	Per capita income[1] (Rand)	Life expectancy at birth	Total fertility rate[2], 1998	Literacy rate[1] (%), 1996	Infant mortality rate, 1998	Under-five mortality rate (1998)
Eastern Cape	1,358	63.7	3.5	59.0	61.2	80.5
Free State	2,419	64.4	2.2	62.7	36.8	50.0
Gauteng	4,992	65.8	2.3	80.6	36.3	45.3
Kwazulu-Natal	1,910	64.3	3.3	61.2	52.1	74.5
Mpumalanga	2,164	64.1	3.1	57.0	47.3	63.7
Northern Cape	2,865	64.3	2.7	58.9	41.8	55.5
Northern Province	725	63.5	3.9	53.0	37.2	52.3
North West	1,789	63.9	2.4	58.3	36.8	45.3
Western Cape	4,188	64.9	2.3	78.7	8.4	13.2
South Africa	**2,566**	**64.4**	**2.9**	**65.8**	**45.4**	**59.4**

[1] Health Systems Trust (1998)
[2] DoH *et al* (1998)

The above data highlight the wide gap existing in the levels of well being among the various social groups and geographical areas in SA. This indicates that within the South African context, aggregate data and averages conceal a lot of important information that could possibly misguide policy. Thus, there is always a need for summary statistics for each population group so as to have a real understanding of the deep-rooted problems.

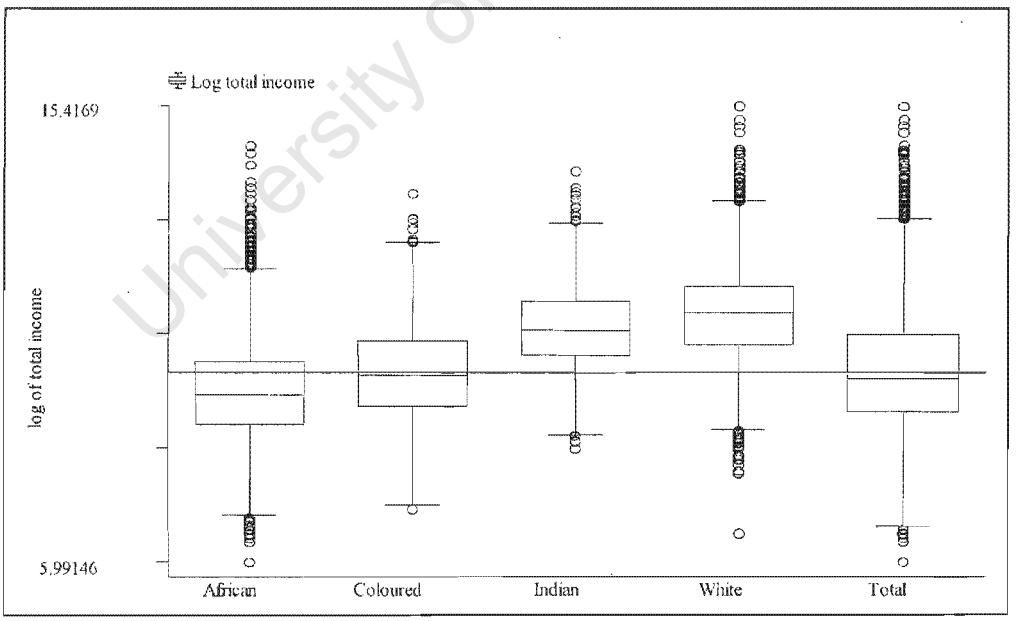
2.2. SOCIAL STRUCTURE, POVERTY AND INEQUALITIES

Prior to 1994, SA had a racially structured political and administrative system, favouring Whites. The exclusionary policies of a system of racial discrimination have left a huge legacy of poverty and inequality. In the period 1986-95, the richest quintile amassed 63.3 percent of the

country's income, whereas the share of the poorest quintile was only 3.3 percent (World Bank, 1998b). In the period 1980-1994 the real GDP per capita (PPP$) of the poorest quintile was 516. In contrast, the richest quintile's per capita GDP (1997) was more than 19 times that of the poorest category (UNDP 1999). In 1993, 23.7 percent of the population lived below $1 a day and about 50.2 percent below $2 a day (World Bank 1998a)[3].

Household income and expenditure figures for the four population groups computed from the 1995 Income and Expenditure Survey (IES 95) manifest wide variations. The mean household income of the Whites is more than four times that of the Africans and more than three times that of the Coloured population. The box plots in Figure 2.2 illustrate the wide variations in household income among the various population groups.

Figure 2.2
Total household income by population group

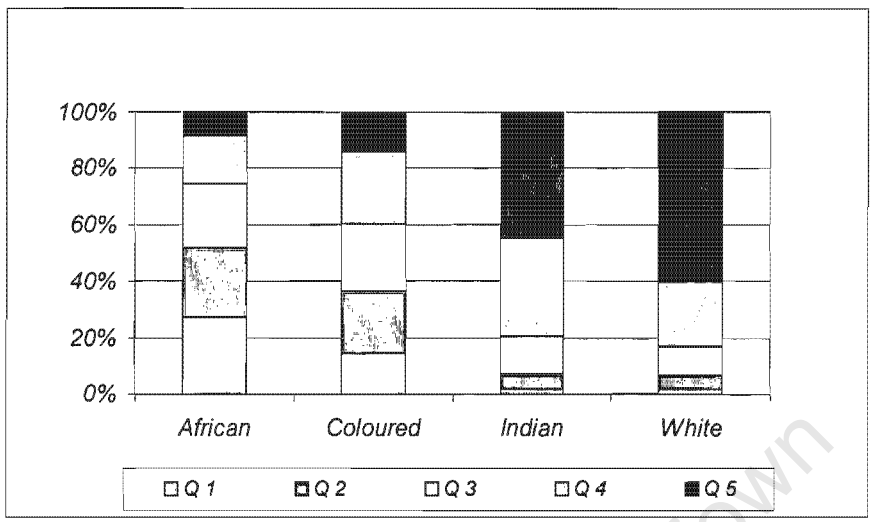


[3] Population below $1 a day and $2 a day are the percentages of the population living on less than $1 a day and $2 a day at 1985 international prices, adjusted for purchasing power parity.

It can be seen from the figure above that the lowest household income quartile of the White and Indian population groups has a mean income that is more than that of the richest quartile of the African population group. In other words, the mean income of the upper quartile of the African group is at best comparable to the lowest income quartile of the White and the Indian population. Furthermore, over 50 percent of those in the African and Coloured population groups earn an income that is less than the overall mean household income (the mean is indicated by the straight line that passes through all the boxes). In contrast the other two groups are way beyond this mark. Overall, the richest ten percent have a share of the income which is about 18 times that of the poorest 10 percent.

To illuminate the inequality further, the head-count ratio[4] is computed using the same IES 95 data and using a poverty line of half and two-thirds of the median household income. The results indicate that overall, 25.1 and 35.9 percent of the South African population is below poverty lines set at 1/2 and 2/3 of the median respectively. When disaggregated by population group, 33.8 and 19.8 percent of the African and Coloured population groups respectively fall below a poverty line set at 1/2 of the median household income. The corresponding figures for Indians and Whites are only 2.1 and 2.4 percent respectively. Raising the poverty line to 2/3 of the median income puts the head-count ratios for the African, Coloured, Indian and White population groups respectively at 47.1, 31.6, 4.9 and 5.1 percent. The head-count ratio for the African group always stands higher than the population mean. The Reconstruction and Development Programme (RDP 1995) states that about 95 percent of the people referred to as poor are African. Further elucidation of the highly skewed income distribution is provided in Figure 2.3.

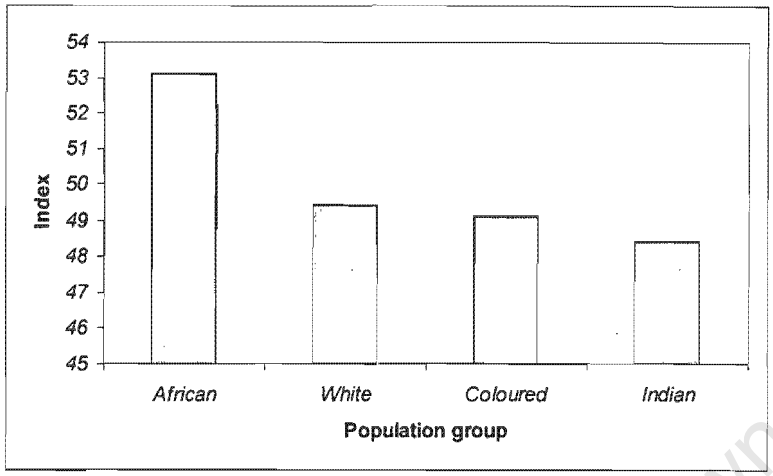
Figure 2.3
Income quintiles by population group (1995)



The Gini coefficient[5] for SA calculated from the income and expenditure data is 0.53. This is 6 percentage points less than the Gini coefficient of 0.59 estimated using the data from the Living Standards and Development Study (LSDS) 1993. It is one of the highest measured levels of income inequality in the World. A closer look at the decomposition of the 1995 Gini coefficient by population group reveals that the African group has the highest index of 0.53. This is a point that has to be considered seriously, as it implies that the group that has the lowest mean household income and highest levels of poverty does not have internal homogeneity. Poverty alleviation measures targeted at this group need to be able to use criteria that can potentially minimise leakage of the benefits to those who are relatively affluent. Figure 2.4 presents the Gini coefficients of the various population groups.

[4] The head-count ratio refers to the fraction of the population below a poverty line
[5] The Gini coefficient measures the extent to which the distribution of income among individuals or households deviates from a perfectly equal distribution. A gini coefficient of zero means perfect equality, while a coefficient of 1 implies perfect inequality. Countries with a gini coefficient of 0.5 and above are considered to have high levels of income inequality.

Figure 2.4
Gini index by population group (1995)



2.3. MACRO-ECONOMIC ENVIRONMENT

After coming to power in 1994, the ANC (African National Congress) led government in its economic and social policy, the RDP (Reconstruction and Development Programme), charted the road ahead to redressing the inequalities inherited from its predecessor apartheid government. The government put the fight against poverty and deprivation high on its agenda.

While the RDP on the one hand espoused a tight fiscal policy with the objective of reducing the fiscal deficit, on the other hand the social policies of the various sectors seemed less cognizant of the tight fiscal policy stance. Social policies were committed to providing social amenities such as health care, housing, clean water and electricity as a right (Pillay and Bond 1995).

The RDP set the broad parameters for the government's economic policy. However, a subsequent macro-economic policy, GEAR (Growth, Employment and Redistribution), which was announced in 1996 (Department of finance 1996) had more significant impact on social

sector policies. GEAR set explicit and relatively ambitious budget deficit reduction targets. It stipulated a decrease of the fiscal deficit from 4.5 percent in 1996/97 to 3 percent in the fiscal year 1999/2000. This policy stance is likely to strain resources for social services including health, and thus, limit access to health services of acceptable quality.

Economic growth, however, has not been in keeping with the projections of GEAR. It fell short of the 3 percent growth rate envisaged which has further constrained social spending levels (Gilson and McIntyre 2001). Hence, it is not difficult to see how arduous it may be to tackle health inequalities within such a scenario of increasing resource constraints, on the one hand, and increasing demand for health resources that is being aggravated by the devastating effects of emerging and resurgent diseases and changing disease pattern. This advances the case for considering technical efficiency and productivity in the provision of services.

2.4. THE HEALTH SECTOR

2.4.1. EPIDEMIOLOGICAL PROFILE

South Africa's epidemiological profile follows the social stratification of its population. Causes of morbidity and mortality manifest marked differences among the various population groups. Infectious and parasitic diseases account for about 14 percent of deaths amongst Africans, while for only two percent amongst whites. On the other hand, about 40 percent of deaths amongst whites are attributed to cardio-vascular diseases. The corresponding figure for Africans is 12 percent (African National Congress 1994b). These data suggest that the causes of ill-health and death in SA include both diseases of poverty that are amenable to low-cost health interventions and chronic degenerative diseases typical of the developed economies.

Tuberculosis is regarded as one of the major health problems in the country. The number of cases reported in 1998 was 169 per 100,000, with the highest rate (464) reported from the

Western Cape province (Health Systems Trust 1999). Its resource consumption is thus tremendous and increasing partly because of its association with HIV/AIDS. Its burden to the health sector was estimated at R200 million in 1992 (van Rensburg, *et al* 1992). The HIV/AIDS epidemic is also assuming alarming proportions. Estimates indicate that at the end of 2001, there were about 5 million HIV infected individuals in South Africa of which about 300,000 were children (UNAIDS 2002). The prevalence rate in the adult population (age group 15-49 years) is 20.1 percent – a rate which is more than two times the average prevalence for sub-Saharan Africa, and about ten times the global average (*ibid*).

Major causes of potential years of life lost (PYLL) include perinatal conditions, acute infectious diseases as well as chronic and degenerative diseases. This is a picture of a country with a dual epidemiological profile.

2.4.2. HEALTH POLICY

The country's national health policy, *the White Paper for the transformation of the health systems in South Africa* was released in 1997. The policy is founded on the tenets of the comprehensive Primary Health Care (PHC) approach in consonance with the health objectives of the country's Reconstruction and Development Programme (RDP).

The issuance of the White Paper has created the necessary framework and vision for the transformation of the country's health care system (Barron *et al* 1997). The policy clearly acknowledges the organic link between socio-economic development and health, and stipulates that improvements in health need a concerted and multi-faceted effort by all sectors of the economy.

Cognisant of the inappropriateness of the apartheid era health services in addressing the health care needs of the vast majority, it is poised to effect a total transformation of the system. The policy envisages that there will be 100 percent population coverage by an integrated package of essential PHC services at the first point of contact.

The objective of universal access to basic health care services is anticipated to be achieved within ten years, provided that the following two conditions hold true:

i. re-allocation of public sector health care resources; and
ii. provision of additional health care resources over and above the budget allocations from the government. These extra resources are expected to be obtained through the development of social health insurance and the retention of user fees collected by public sector hospitals.

The principles of equity and efficiency form a central part of the national health policy. With respect to equity, some of the guiding principles include:

- increasing access [physical] to integrated health services to all people, predominantly focusing on the poor and other vulnerable groups;
- establishment of health care financing policies that are promotive of equity; and
- distributing human resources equitably.

Although there is no clear statement of what is implied by *equity*, in the glossary of the White Paper it is defined as '*the universal provision of services on the basis of need rather than any other criterion'* (South Africa 1997: 224).

Efficiency is also stated as one of the objectives nurtured by the policy. The following statement in the policy describes the *status quo* in terms of efficiency and equity:

> "In 1992/93, South Africa spent approximately 8.5 percent of GDP on health services, both public and private. This represents a very high level of spending for a country at South Africa's level of development. However, the distribution of resources is inequitable and wasteful" (South Africa 1997:40).

This point brings to the fore that equity and efficiency are seriously compromised, and thus, the policy emphasises the need for intensified efforts to rectify these shortcomings if the targets set in the health policy are to be achieved timeously.

2.4.3. HEALTH FINANCING

South Africa's expenditure on health care as a proportion of the GDP is relatively high compared to the average of its counterparts classified as upper middle income countries. In the period 1990-95, total expenditure on health was 7.9 percent of the GDP. This translates into a per capita expenditure of $ 257. The breakdown of the expenditure shows that private sector expenditure amounts to 4.3 percent of the GDP as opposed to 3.6 percent in the public sector (World Bank 1998b)

The source of funding for the public sector health care system is predominantly tax revenue (McIntyre *et al* 1998a). Thus, the amount available in large part depends on the tax base, which in turn depends on economic growth, efficiency of tax collection and the overall fiscal policy of the government. Given the objective of narrowing the budget-deficit, which is stated in the government's economic policy, Growth, Employment and Redistribution (GEAR), expenditure on health care may be seriously affected along with other basic-need spending (van Rensburg *et al* 1998).

According to the medium-term expenditure framework of the Department of Health, which is part of the Ministry of Finance's overall expenditure framework until the year 2000, public-sector health care expenditure is assumed to grow by 3.6 percent in real terms (Sidiropoulos *et al* 1998). However, evidence suggests that real per capita spending has declined. Per capita real expenditure decreased by about 5 percent in 1998/99 as compared to that in the previous year. It is projected that real per capita expenditure will decline to 512 Rand in the year 2000/2001 from its 1995/6 value, which was 516 Rand (McIntyre *et al* 1998a).

The geographical distribution of public sector expenditure is skewed, with the Western Cape and Gauteng provinces well above the national average. In 1998/99, real per capita expenditure in the Western Cape and Gauteng provinces respectively was 47 percent and 81 percent above the national average. Six out of the nine provinces had a level, which was below the national mean of 499 Rand (McIntyre *et al* 1998a).

A closer look at the trend indicates that there was an increase in some of the provinces over the years 1995/96-1998/99, even though no discernible pattern exists. For example, Gauteng, which had a level well above the national average in 1995/96, registered an increase, whereas the North West with a per capita expenditure much lower than the national average showed a decrease.

It should be noted at this juncture that equalisation of per capita expenditures across the provinces may not be realisable in the short run, nor is it to be taken as an ultimate objective for various reasons including:

i. Current expenditures follow the capital outlays, which were incurred a long time ago. For example, the large academic and non-academic hospitals practicing high-technology medicine are concentrated in the metropolitan areas of a few provinces.

Keeping these hospitals functional creates an expenditure gradient that favours some of the provinces.

ii. Drastic cuts in those provinces are likely to be encountered with firm opposition, especially from the medical profession, which is a highly influential group in the sector.

iii. Highly sophisticated hospitals are meant to serve as referral centres for areas beyond their respective provincial catchment population. Considerations of efficiency, especially economies of scale, dictate that such hospitals be based in few selected areas.

iv. At face value, a move towards equalisation of per capita expenditure may not be taken as a move towards equity, as equitable distribution could also be achieved in the face of unequal per capita expenditures. Issues such as *need* and geographical input-price differentials can impact on whether a given allocation is equitable or not.

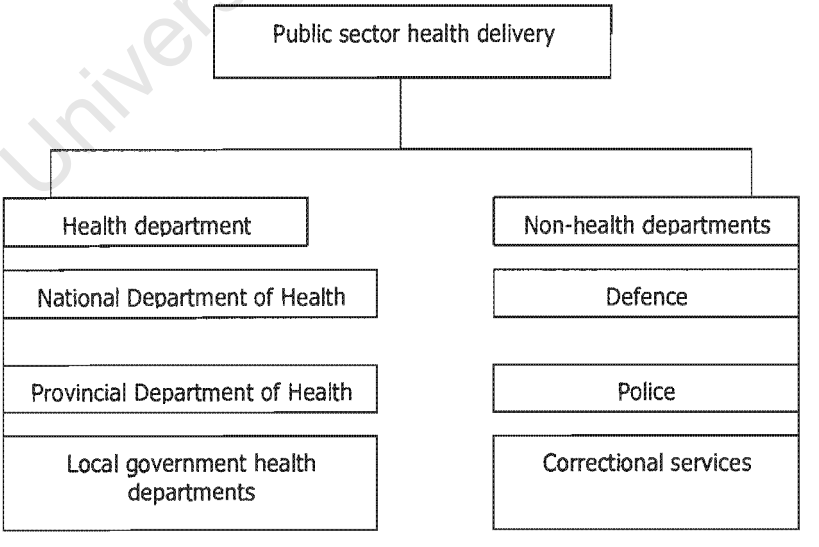
There are also private health care financing sources and intermediaries, the most important of which are medical aid schemes and direct out-of-pocket payments by households (McIntyre *et al* 1998b). In 1995, the number of medical aid schemes' beneficiaries was about 6.8 million (Sidiropoulos *et al* 1998), which is about 17 percent of the total population. The second most important contributor to private health care finance, direct out-of-pocket payments, includes *schemes gap payments* (which refers to the difference between the fees charged by private health care providers and the amount reimbursed by medical aid schemes), *co-payments, user fees* at public sector hospitals, payments for consultation and purchase of prescribed drugs by non-scheme members and purchase of over-the-counter drugs (*ibid*). User fees in public sector hospitals recouped about 9 percent of their total expenditure in 1992/93 (McIntyre *et al* 1995). However, this level of cost recovery has declined since then. Commercial health insurance, which is directed at individuals, is small but a growing component of private health care financing (McIntyre *et al* 1998b).

It is important to mention here that the HIV/AIDS epidemic is sapping finances of the health sector, thus undermining the government's initiatives to expand access to good quality health care to its population. In his 2002 Budget Speech, the Minister of Finance has indicated that about R4 billion is currently spent by provincial health departments on HIV/AIDS-related illnesses. More finances are also required to fund prevention programmes in schools and communities (Manuel 2002).

2.4.4. ORGANIZATION AND DISTRIBUTION OF SERVICES AND FACILITIES

Both private and public sectors are involved in service provision. The private sector includes both for-profit hospitals and not-for-profit hospitals owned by welfare organisations or by industries. Service delivery in the public sector is organised at the three spheres of government and an additional category including provision by departments outside the health sector as depicted in Figure 2.5 below.

Figure 2.5
Organization of services



The central department of health, in addition to policy formulation and stewardship of the health sector, provides services whose provision by the other levels is deemed cost-ineffective (e.g. specialised laboratory services). Within the jurisdiction of the provincial health departments are provided most of the hospital and curative non-hospital services. The local government health departments focus on providing preventive primary care services with an emphasis on communicable diseases and environmental health. The departments outside the health sector provide health care mainly to their employees and their dependants.

The hierarchy of facilities in the system assumes the following structure in ascending order:

Box 2.1
Organisation of public sector hospitals

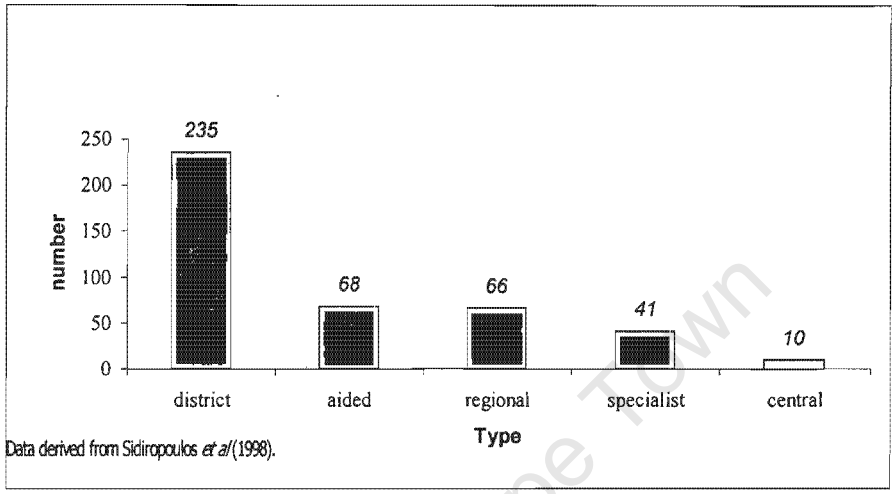
Hospital type	Characteristics
District hospital	• Have at least 30 beds • Provide first-level care • Include casualty department, family medicine, paediatric rehydration units, polyclinic and sleep-over beds
Regional hospital	• Provide secondary-level care • Include the following departments: anaesthetics, general medicine and surgery, obstetrics and gynaecology, orthopaedics, paediatrics, psychiatry, radiology
Central hospital	Provide services requiring expertise of sub-specialists and rare specialities such as cardio-thoracic surgery, neurosurgery, plastic surgery and urology
Specialist hospitals	Provide care only for specific types of patients such as tuberculosis and chronic psychiatric patients

Adapted from Sidiropoulos *et al* (1998)

Facilities with less than 30 beds are referred to as community health centres or clinics. All levels of hospitals are expected to carry out the teaching function. In 1997, there were a total

of 420 public sector hospitals whose distribution between type of hospital is as shown in Figure 2.6.

Figure 2.6
Hospitals by type



The category described in the above figure as *aided* does not, in a strict sense, denote the level of a hospital. It rather indicates type of ownership and financier. *Aided hospitals* are those that are owned privately but get funding from the government. The total bed-capacity of public sector hospitals in 1996 was estimated at 111,000. The breakdown of the bed-density by province is depicted in Table 2.3.

Table 2.3
Bed density by province

Province	Public hospital beds per 1,000 population (1995)
Eastern Cape	3.0
Free State	2.7
Gauteng	2.3
Kwazulu-Natal	2.7
Mpumalanga	1.6
Northern Cape	3.0
Northern Province	3.1
North-West	2.3
Western Cape	3.3
Total	**2.7**

calculated by taking bed-size from Sidiropoulos *et al* (1998) using population figures from HST (1998).

There were 304 private hospitals in 1996, the majority of which (41 percent) were located in Gauteng province (Sidiropoulos *et* al 1998) followed by the Western Cape province (14.5 percent) and Kwazulu Natal (11 percent).

The number of clinics in 1997 was 2,678 of which only 74 were in the private sector. It can be seen that while private hospitals comprise 42 percent of all hospitals, the share of the private sector in clinic ownership is only 2.8 percent of the total number of clinics. This focus of the private sector on hospital services may possibly be the result of a strong profit motive, as the hospital sector is a relatively more lucrative business.

The focus of the private sector on hospital services has some important implications for policy and planning. First, an unchecked proliferation of hospitals whose production process is high technology oriented may lead to spiralling of costs with serious repercussions for the health care system. Second, government has to take heed of the phenomenon of crowding-out of investments, which implies the likely duplication of services that the private sector is highly motivated to provide. This might result in misallocation of resources, which could be deployed for running community health centres, to which the private sector does not seem to have a strong motivation. Third, an effective quality assurance and control mechanism needs to be in place to regulate the activities of the private hospitals to avoid adverse effects on consumers that may emanate from asymmetry of information and a strong profit motive. This entails the establishment of an effective organ of quality assurance in the Department of Health's structures, as hospital production is a complicated process compared to that in clinics.

2.4.5. Human resources

The statistics on health care human resources indicate that South Africa has physician-to-population ratios that are considerably lower than those of countries referred to as *upper middle income*. The number of physicians per 10,000 people for this group of countries was 16 in 1994 (World Bank 1998b), as compared to South Africa's figure of 4.4[6] for 1997 (see Table 2.4). Disaggregating this figure by province indicates that even the most resource-endowed province does not have a population-to-physician ratio approaching that of the upper middle-income countries. National and provincial personnel-to-population ratios for some of the major categories of health workers in South Africa are provided in Table 2.4.

Table 2.4
Personnel distribution in SA provinces

Province	Doctors per 10,000 people 1997*	Nurses per 10,000 people* (1997)	Pharmacists per 100,000 people (1998)**	Dentists per 100,000 people (1998)**
Eastern Cape	2.3	29.2	2.0	0.5
Free State	3.4	34.4	2.4	0.6
Gauteng	9.3	40.2	3.7	1.7
Kwazulu-Natal	4.5	41.2	3.3	0.4
Mpumalanga	2.1	22.3	2.2	0.7
Northern Cape	7.4	28.3	2.0	0.5
Northern Province	1.0	30.6	1.9	0.3
North-West	1.8	21.2	1.7	0.6
Western Cape	7.7	47.6	4.6	1.3
Total	4.4	42.5	2.8	0.8

Sources: * Sidiropoulos *et al* (1998); figures include both public and private sectors
** Health Systems Trust (1999); only those in the public sector

The above table indicates the disparities in the distribution of personnel in the various provinces. The discrepancies are more pronounced in the physician-input distribution with Gauteng province having a ratio which is more than nine times that of the Northern Province. The provincial distribution of pharmacists and dentists might also be expected to follow the same pattern as that of the physicians if those working in the private sector were included. Differences in the distribution of nurses are relatively smaller than the other two categories.

[6] Does not include specialist physicians

It should again be noted that by merely inspecting the spatial differences in personnel distribution, a conclusion of inequity in health human resources allocation is unjustified. Input-ratios tell only part of the story, and on their own, are not sufficient to enable one to make judgements about the presence of inequity.

Besides the inter-provincial variations in health human resources, studies have also revealed inequalities in intra-provincial allocation. A case study of the Western Cape province has shown wide discrepancies in personnel-to-population ratios in the various administrative regions of the province (Makan 1998). The implication of this finding has an important bearing on micro-level planning of resources. Reliance on provincial averages alone may be flawed, as it does not illuminate the actual scenario on the ground. Hence there is a need for a careful scrutiny of data at the level of a district or a micro-geographic area if initiatives to rectify inequitably skewed resource-allocations are to be consequential.

2.4.6. Access and utilisation of services

In 1994, about 62.2 percent of South African households lived at a distance less than five kilometres from a health care facility (Statistics South Africa 1995). Given the skewed allocation of resources it is not unexpected to find stark contrasts among the provinces. For example, the Western Cape Province had the highest proportion of households (72.9 percent) within a five-kilometre radius of a health facility as compared to the Northern Province, which had the lowest figure (46.4 percent). Besides differences in access, differences in population densities among the various provinces may also account for differences in population coverage.

Many low and middle-income countries surpass the above figures for South Africa. For example, in 1993 the percentage of population with access to health care in Botswana, Gabon, Malawi and Madagascar respectively were 86, 87, 80 and 65 (World Bank 1998b). Access in this case is defined as the share of the population that can expect treatment for common diseases and injuries, including essential drugs on the national list, within one hour's walk or travel. Both definitions hinge on distance to the facility as a measure of access, but differ in their measurements of distance: time versus physical distance. This may not introduce striking differences, as the cut-off points given in terms of time of travel to the facility and physical distance are more or less equivalent. A caution that has to be exercised in all cases is the fact that access to health care is a multi-dimensional concept and that the above figures should not be misconstrued for effective access, where people are able to get treatment whenever the need arises. Financial constraints, for instance, may reduce the effective access to services.

In 1992/93, the per capita number of visits to public sector hospitals was 1.8 (McIntyre *et al* 1995). Although this is low by accepted norms for developing countries, it might be in line with the expected if visits to the private sector are included. Utilisation rates for selected services are presented in Table 2.5.

Table 2.5
Utilisation of selected preventive services, 1998

Service	Percentage
Immunisation of children aged 12-23 months	63.4
Pregnant women who received antenatal care	94.2
Pregnant women who received tetanus toxoid vaccine	58.8
Contraceptive prevalence rate, women aged 15-45	62.1
Deliveries attended by trained personnel in health institutions	84.4

Source: HST (1999)

It can be seen that the achievements in preventive maternal and child health services are equivocal. Whereas there is almost complete antenatal care coverage, immunisation coverage of children aged 12-23 months is relatively low. Furthermore, despite a high attendance in the antenatal care programme, the proportion of mothers who received tetanus toxoid is only 58.8 percent.

Spatial differences in the utilisation of services indicate differences, which in some instances are counter-intuitive. For example, the proportion of women who received tetanus toxoid in the Western Cape, one of the best-resourced provinces, is extremely low (17.8 percent).

2.5. Summary

In this chapter an attempt has been made to develop a profile of the study country with a view to facilitating forthcoming discussions of equity and efficiency of the health care system. It can be discerned from the discussion that the system has a large backlog of health inequities to be redressed in the face of resources that are dwindling as a result of contractionary fiscal policies and escalating needs.

The exposition also highlights the situation of health care and health status and the various socio-economic and demographic factors that are in play. An understanding of the interplay among the multitude of factors is essential in making informed resource allocation decisions with an objective of upholding the principles of equity and efficiency.

CHAPTER 3
EQUITY IN HEALTH AND HEALTH CARE: CONCEPT AND MEASUREMENT

It is when equals have or are assigned unequal shares, or people who are not equal, equal shares, that quarrels and complaints break out.
Aristotle

3.1. INTRODUCTION

This chapter aims at giving a brief description of the concept of equity as it applies to health and health care with a view to creating an understanding of the various perspectives regarding its definition. Furthermore, the chapter describes some of the issues surrounding the measurement of equity, such as classification of social position and quantification of inequalities. The chapter is intended to lay the necessary framework for the three empirical chapters on equity that follow.

Resource allocation decisions in any society entail two criteria (Magill 1997):

i. A scientific, efficiency oriented component; and

ii. A normative or ethically oriented component, that is equity.

Equity and efficiency are the twin objectives of any social policy that, in principle, are considered in any resource allocation decisions.

There is considerable consensus among policy-makers in developing countries that equity should be accorded a prominent place in health policy decisions (Gilson 1988). The health policies of most countries have explicit, albeit vague, statements on equity. Equity is also enshrined in the World Health Organisation's (WHO) Alma Ata declaration as one of the pillars of the Primary Health Care (PHC) strategy (WHO 1978).

However, despite the consensus that equity be accorded weight in allocating scarce resources, interpretations of its meaning abound. Endeavours to formulate definitions, which are simultaneously sufficiently general to command a broad consensus, and sufficiently specific for the purposes of practical application have proved difficult (Le Grand 1984).

A sound conception of equity needs to meet certain criteria, the most important of which are (Le Grand 1984, Pereira 1993):

i. A limited information requirement, so that its usability will not be limited;

ii. Easy comprehensibility, so that it allows for an informed discussion among all stakeholders and leads to a clear policy direction;

iii. It should be possible to find allocation of resources that is both equitable and pareto-efficient; and most importantly

iv. It must enjoy intuitive acceptability.

Inequities in health among different population groups in society are of great concern to policy-makers and society at large. Therefore, to design and implement sound social/health policies that would improve the quality of life of the people, a clear understanding of the magnitude, nature and causes of inequities in health is of immense significance (Lairson *et al* 1995). To this end, issues such as the following need to be addressed (Whitehead 1988):

- Are certain groups of the population deprived of the opportunity to achieve an acceptable level of health because of their socio-economic circumstances (e.g. income, race, gender, ethnic group, employment status)?
- Does the distribution of health show spatial differences, where some areas have utterly unfavourable levels of health status and others enjoy good health?
- Is the geographical distribution of resources under the disposal of the health care system fair?

These are, but some of the major issues that we need to address in order to redress existing inequities.

3.2. EQUITY AS AN OBJECTIVE OF HEALTH POLICY

3.2.1. INTRODUCTION

The reduction of inequities in health and health care remains one of the challenges of health policies of all countries, both developed and developing. Equity is in fact one of the major objectives and priorities of health policy for economic, social and moral reasons (Dahlgren and Whitehead 1992).

Inequality in health is viewed seriously as it hampers people from flourishing and realizing their full potential in life by conditions that are potentially avoidable. Likewise, inequality of health care distributions merits close attention because health care services play an instrumental role in promoting people's health, which in turn is regarded as necessary for a good life. Health care is different from other goods and services in that it affects life and death directly.

The principle that health should be distributed according to *need* is frequently encountered in health policy statements (Culyer and Wagstaff 1993, Newbold *et al* 1995, Le Grand 1996). Given the pervasiveness of the scarcity of health care resources, and given that the allocation of resources should not be based solely on the market mechanism, it is imperative to have a thorough understanding of the core concept, *need*. This is attempted in the section that follows.

3.2.2. HEALTH CARE NEEDS

The following two definitions of need are most frequently cited in the literature (Culyer and Wagstaff 1993, Le Grand 1996, Olsen 1997):

i. Need as initial health

Need in this case is equated with ill-health (severity). People who are more ill than others are regarded as having a greater need (Gillon 1985). In this notion of need, it is implied that people with similar health status have equal needs and those with different health deficits have different needs. The flaw of this notion is that the need for health care exists regardless of whether or not the persons can avail themselves of a health care intervention.

An extreme variant of this version of need is the *rescue principle* which prescribes the use of whatever resources we have at our disposal to save life (Dworkin 1994). It follows that a terminally ill person should receive the maximum health care even if the outcome is not promising. Hence much of the scarce health care resources will be devoted to patients with a very small probability of recovery. This principle of allocation is inequitable with respect to those whose degree of health deficit is smaller, and is inefficient.

ii. Need as capacity to benefit

This approach regards need as instrumental to the attainment of a certain ultimate goal. Daniels (1985:23) in a similar tone has said, "without abuse of language, we refer to the means necessary to reach any of our goals as needs". In this consequentialist view of need, two conditions are assumed to hold true (Culyer and Wagstaff 1993):

a. The entity should indeed be instrumental to attain the goal under consideration; and
b. The goal itself should be deserving enough to justify the use of the term "need".

This interpretation of need, however, leaves the meaning of "capacity to benefit" vague. A more refined version relates "capacity to benefit" to "health gain" — reduction in the patient's health deficit that results from receiving the treatment (Le Grand 1996). This method has some practical problems that relate to the measurement of health gain and the ascription of a health gain to a specific treatment (*ibid*).

Despite the above noted limitations of the two interpretations of need, incorporating both into the ethics of resource allocation might be preferred, as the two are likely to complement each other. This, for example, fits well into the principles of priority setting in health care in Norway (See Box 3.1 in later parts of the chapter) (Olsen 1997).

Needs may be identified subjectively by the individual's own assessment or using objective criteria independent of the individual's own assessment. In most instances, however, it is the independent assessor's opinion that is given precedence. In health care, although the patient's opinion counts, the greatest weight is attached to the clinician's objective assessment. It follows that needs are objectively ascribable, meaning that we can ascribe them to someone even if (s)he fails to realise their presence because (s)he has preferences conflicting with the needs.

Needs are classified into two (Daniels 1985), so as to aid distinguishing relevant needs from the class of all needs that we come across. These are:

i. Course-of-life needs

These are needs that we all have throughout the course of life or at some stage of it. Examples under this category are, food, shelter and clothing. Lack of this category of needs jeopardises the normal functioning of the individual as a member of a natural species. Such needs are objectively ascribable and important.

ii. Adventitious needs

These are the things we need because of particular contingent projects on which we embark. Deficiency of these needs does not impair the species' typical functioning.

From the foregoing discussion the following observations are made:

i. The need for health care is a derived need, which is sought for being instrumental to the attainment of good health, which in turn is essential to flourishing and happiness. It then follows that if health care does not assist in the attainment of its intended goal (reducing suffering and improving health in quantity and quality), then it will not be needed (Culyer 2001). This brings into the picture, the issue of the effectiveness of interventions. The need for ineffective health care is automatically ruled out.

ii. The notion of need equated to the severity of illness does not take account of whether the individual can benefit from the intervention or not. It does not discriminate against ineffective health care, which may be wasteful of the scarce resources of society.

iii. Government's focus should be laid primarily on those categories of need, which are essential for the normal functioning of the species. This is obvious for the reason that there is a critical shortage of resources, especially within the context of developing countries.

iv. The fact that the objective assessment of need is given precedence over the individual's subjective assessment implies that not all needs that may be verifiable objectively will translate into utilisation of health care, even if material deterrents to access do not exist. This signifies that health care distributive justice ought to take account of non-material or non-financial deterrents to access, such as cultural issues.

v. The two notions of need — need as health status and need as capacity to benefit have to be used as complementary to each other, so as to account for efficiency concerns and be equitable in a sustainable way.

3.3. THE MEANING OF EQUITY IN HEALTH AND HEALTH CARE

As discussed earlier, the concept of equity can be interpreted differently by different people at different times (Whitehead 1993). In emphasising this it has also been said that, "equity like beauty is in the mind of the beholder" (McLachlan and Maynard, 1982:p 520). (For further discussion on the philosophical foundations of equity see Appendix 1).

However, despite the multitude of definitions, there is a view unifying all of them. All regard equity as being about fairness of the distribution of something or another (Mooney 1983). In line with this, Wagstaff and van Doorslaer (1993) found a broad agreement among policy-makers and researchers in Europe and North America on the principle that health care should be distributed according to need and financed according to ability to pay. For further understanding, Box 3.1 below presents health equity policy statements of some countries.

Box 3.1

Health equity policies in various countries

Country	Definition	Source
Australia	Equal access for equal need (equity in delivery)	Mooney (1996)
Sweden	The objective of the health care system is to promote good health and on equal terms. Geographical differences should be reduced and that the possibilities of receiving care must not be influenced by circumstances such as ability to pay, nor should it be influenced by the type and duration of illness	Gerdtham (1997)
Canada	Resources committed to the delivery of health care should be allocated in accordance with medical necessity (the ability to benefit from health care) as opposed to willingness or ability to pay for services	Newbold *et al* (1995)
Norway	A group of patients which suffers more than another group, should be given priority if both groups can be helped equally much for the same costs. When choosing between two programmes for patients who are equally severely ill and who can be helped to the same costs, the group of patients who have the most benefit from programmes should be given priority	Olsen (1997)
New Zealand	Equal access for equal need	Peacock *et al* (1999)
The Netherlands (Curaçao)	Removal of financial barriers to access	Alberts *et al* (1997)
South Africa	The universal provision of services on the basis of need rather than any other criterion	South Africa (1997)

As can be discerned from the above statements of health policy, the principle of equity seems to have at its centre the notions of *access* and *need*. The preferred definitions seems to be *equality of access for equal need*, which is mainly horizontal equity oriented.

On the basis of what has been said so far, in this study the following definitions are adopted:

i. Equity in health

This study uses Whitehead's definition of equity in health (Whitehead 1993). Health differences that result from avoidable factors should not be allowed to exist between individuals or groups. Individuals should have an equal choice set to attain their optimum health. Thus inequalities in health/ill-health that are systematically related to the individual's socio-economic status, race, gender, education *etc* constitute inequities. The focus in this study is on inequalities that are related to one's socio-economic status and not health inequalities per se. This implies that the emphasis is on those health inequalities that health and social policies can address.

ii. Equity in health care

This implies equal access for equal need, where access and need are defined as follows:

- need is assessed by self-reported acute or chronic illness or disability in the adult population (those aged 18 years and above).
- access refers to the absence of barriers to utilisation of health services that are attributable to the individual's socio-economic status when there is an expressed need for it.

Thus, this study is concerned in assessing socio-economic inequalities in the utilization of health services in those who have an expressed need for it. The major focus is the assessment of horizontal equity, that is, whether there is equal treatment for equal need. It should be noted that there is a limitation in measuring need by self-reported illness, as the severity of the illness, particularly in the short-lived (acute) illnesses can not be discerned from the type of questions (which demand a "yes" or "no" answer) asked in the household surveys (for other limitations of self-reported adult illness see discussion on Chapter 6).

The measurement of equity involves the classification of people by different socio-economic and demographic variables. Furthermore, it entails the quantification of the degree of inequality. These issues will be briefly addressed in the following discussion.

3.4. MEASUREMENT ISSUES

3.4.1. INTRODUCTION

The measurement of equity in health and health care is a vital step in government initiatives to redress inequities in an informed manner. This undertaking requires information on the classification of socio-economic status (SES) and quantification of health/health-care inequities.

As is true with the concept of equity, the meaning and measurement of health inequalities and social group differences is also highly debatable and controversial. Definitions of socio-economic position vary between countries. Likewise, the measurement of illness and health inequalities varies between countries according to the design of the data collection instruments used to obtain morbidity and mortality data (Lahelma *et al* 1994). This is a serious limitation, especially in making inter-country comparisons. The availability of standard definitions and measurement techniques is essential even in making spatio-temporal comparisons of equity within and among different provinces or health districts of a country. The following sections discuss issues surrounding the definition and measurement of health inequalities and SES.

3.4.2. CLASSIFICATION OF SOCIAL POSITION

The measurement of SES is one important step in the assessment of equity. The existence of significant relationships between socio-economic indicators and inequalities in health is a well-established fact (Pereira 1990). Existing evidence suggest that people with lower SES suffer a disproportionately heavier burden of illness and have higher mortality rates than those on the higher end of the SES scale (Whitehead 1989).

For the purpose of comparison, a number of ways are used in categorizing populations into relatively homogenous groups. SES is commonly proxied by three interrelated indicators (Alberts *et al* 1997): material, knowledge and socio-cultural components, which may be represented by income, education and occupation respectively.

Studies of inequalities in health conducted in the United Kingdom seem to favour the use of "occupation" as a basis of classification (Carr-Hill 1990, Manor *et al* 1997). Researchers in other parts of continental Europe frequently use "education" either independently or in conjunction with "occupation" (e.g. Rahkonen and Lahelma 1992, Lahelma *et al* 1994). Some studies in the United States use racial groups as SES categories (Murray *et al* 1999), and others use income and education in combination (Manor *et al* 1997). In South Africa researchers have, in the main, used classification based on predominantly race and income (e.g. Gilson and McIntyre 2001). Gilson and McIntyre have also tried to proxy SES using various characteristics, such as location and an environmental health index.(*ibid*)[1]

Each of the above measures of SES has its own drawbacks. For example, `occupation' does not take account of job mobility and occupational classes are not exhaustive, thus leading to

[1] A composite index made up of three variables: access to piped water, access to sanitation facilities and connection to electricity. All are given equal weights and the total score ranges between 0 (least favourable) to 3 (most favourable)

forcing occupational titles into the nearest occupational category (Carr-Hill 1990). `Education' has a skewed distribution, which detracts from its appeal (Lahelma *et al* 1994). All measures are not likely to capture the dynamic nature of socio-economic inequalities in health, and a proxy measure, which at one point in time was sensitive, may with change of circumstances be unable to capture the new reality. For example, in the South African situation, although race can be a sensitive measure of socio-economic status at the moment given the country's apartheid history, over time (however short or long it may be) classification on the basis of race may not be the best proxy for SES. This is because socio-economic differences within race groups are likely to increase, while those between race groups may decline. This entails the need for complementing race with other socio-economic measures in the long run. Thus, there is a need to re-assess the measures occasionally so as to avoid unintended consequences.

Geographical criteria for classifying SES, known as micro-geographic markers, become important if the localities being considered are small and relatively homogenous (Murray *et al* 1999). Micro-geographic markers take account of the characteristics of an area, which adds to their strength. However, the limitation of these criteria is a possible misclassification of people, as a high degree of internal homogeneity of a locality may not be guaranteed.

Some studies have also used composite asset and deprivation indices constructed from multiple variables using the method of principal components analysis (Gwatkin *et al* 2000, McIntyre *et al* 2001). This methodology is especially important in data sets which do not contain adequate income or occupational data.

The current study uses the following as measures of social position: income status, population group and residential location (rural/urban).

3.4.3. HEALTH STATUS MEASUREMENT

In assessing health inequalities, the selection of an appropriate measure of health status is of paramount importance. A multitude of measures have been devised and used to this end. Measures of (ill) health have mainly focused on negative parameters whose focus is geared towards quantifying the prevalence/incidence rates of specific or general causes of morbidity and mortality (e.g. prevalence of disability, infant mortality rate, *etc*). The infant and under-five mortality rates are used frequently, as they affect life expectancy at birth significantly and are regarded as key health indicators on their own (Wagstaff 2000).

The relative advantage of mortality rates over indicators of morbidity is that, they are non-recurring incidents, and therefore easy to remember. Moreover, they avoid bias that may ensue from differences in perceiving and reporting an illness episode by people with different socio-economic and cultural backgrounds which may result in counter-intuitive and misleading findings. For example, Gilson and McIntyre (2001) reported that the prevalence of self-reported morbidity in South Africa did not conform to expectation: people in the lowest SES had illness prevalence rates which were lower than those of the highest SES.

Despite some of the above stated limitations, many studies of inequalities in health have used illness as a measure of need. Illness may be defined in terms of directly observable objective signs or self-reported subjective symptoms (Twaddle 1979), which depend upon recall and awareness of the respondent.

As it is difficult to capture the multi-dimensional nature of the concept of morbidity in a single indicator (Lairson *et al* 1995), many studies use a combination of self-reported measures. These include:

i. presence of chronic illness (lasting or expected to last for more than six months), which may or may not limit the individual's capacity to do the activities of daily living;

ii. presence of short-lived illness (Lasting for less than a month); and

iii. self-rated general health status.

Although simple in design, self-rated general health status, which is the response to a question such as "how is your health in general", has been found to be a valid measure of health status (Lairson *et al* 1995). It has been observed to perform as a reliable predictor of future mortality, and thus a good indicator of the need for health care at the population level (Newbold *et al* 1995). In South Africa, the "October household survey" series questionnaire use self-reported illness episodes in a one-month recall span (CSS 1997).

In this study health status is measured by infant and under-five mortality, under-five child malnutrition and self-reported illness in adults, which in this case consists of those aged 18 years and above.

3.4.4. QUANTITATIVE METHODS FOR HEALTH INEQUALITY MEASUREMENT

Having decided on the attribute to be compared among individuals/groups, the next logical step is to find an appropriate technique to quantify the degree of the existing inequality. Several methods have been in use to date. Some have their origin in research on income inequality (e.g. Lorenz curve and the associated Gini coefficient) (Atkinson 1970, Vagero and Lundberg 1989) or from modifications of these (e.g. concentration index) (Wagstaff *et al* 1989). Other methods are based on measures of association (index of dissimilarity, slope index of inequality) (Manor *et al* 1997).

The following sections describe those measures of inequality that are used in this study (for a review of other measures of inequality see Appendix 2).

i. RATE RATIO

This involves comparing the rate of (morbidity/mortality) of the lowest socio-economic group to that of the highest. Wagstaff *et al* (1991) call this measure the `range', because the extent of inequality is judged primarily on the basis of the distance between the top and bottom socio-economic groups. While simple to calculate and understand, the disadvantage of this method is that it does not accommodate the intermediate classes and does not consider health differences within each group being compared. To compensate for this weakness, sometimes socio-economic groups are re-categorised. For example, if socio-economic categorisation is done on the basis of income quintile, then the bottom two quintiles (quintiles one and two) may be lumped together and compared with the top two (i.e. quintiles four and five). This increases the proportion of the population included.

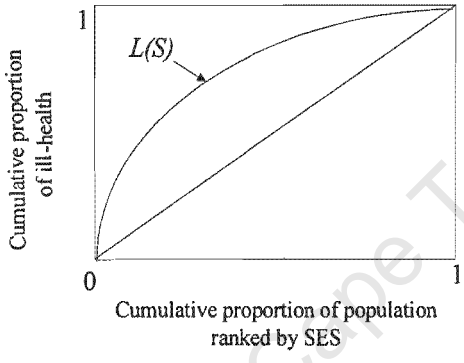
ii. CONCENTRATION INDEX

The concentration index (C) plots the cumulative proportions of the population ranked by their socio-economic status against the cumulative proportions of (ill)health. It gives the global extent of socio-economic inequality in health. It has to be noted that in contrast to the Gini coefficient, the concentration index ranks people not by their health status, but by their socio-economic status beginning with the least advantaged. The illness concentration index is presented in Figure 3.1.

If the illness concentration curve lies above the diagonal as shown above, it implies that there is a disproportionately higher burden of illness among the poorest. In other words, there is a

pro-rich inequality in illness. In this case the concentration curve assumes a negative value. On the other hand, when the illness concentration curve lies below the diagonal, it implies a pro-poor inequality in illness, and the concentration index takes positive values. If illness is distributed equally, the concentration curve overlaps with the diagonal.

Figure 3.1
illness concentration curve



The concentration index takes values ranging from +1 to −1 (a value of zero signifying absence of inequality), and is twice the area between the concentration curve and the diagonal with (Kakwani *et al* 1997):

$$C = 1 - 2\int_0^1 L(s)ds \tag{3.1}$$

From individual level data the concentration index can be computed as follows:

$$C = \frac{2}{n\mu}\sum_{i=1}^{n} x_i R_i - 1 \tag{3.2}$$

where,

x_i $(i = 1,..,n)$ is the ill-health score of the i^{th} individual;

$\mu = (\frac{1}{n})\sum_{i=1}^{n} x_i$ is the mean level of ill-health; and

R_i represents the relative rank of the i^{th} person.

In the above calculation, individuals are ranked according to their socio-economic status beginning with the worst off.

The above computation, however, does not enable statistical inference, i.e. it is not possible to know whether or not the calculated concentration index has a statistical significance. To this end, a standard error for C can be calculated using a convenient regression as follows (Kakwani *et al* 1997):

$$2\sigma_R^2(\frac{x_i}{\mu}) = \beta_0 + \beta_1 R_i + u_i \qquad (3.3.)$$

In the above equation β_1 is equal to the concentration index and the standard error of β_1 represents the standard error of the concentration index.

The concentration index meets three important characteristics that a good measure of inequality is expected to fulfil (Wagstaff *et al* 1991):

i. it takes account of the socio-economic dimension of inequality in health;

ii. it reflects the experience of the entire population rather than two extreme groups on the SES scale (e.g. income quintile 5 versus income quintile 1) as is the case in range measures (e.g. rate-ratios); and

iii. it is sensitive to shifts in the population across socio-economic groups.

The CI is free of the flaws experienced by ratios where top and bottom comparisons are done. Furthermore, unlike the Gini coefficient, it ranks people by SES, thus taking care of the socio-economic dimension of health inequalities, and is sensitive to changes in the distribution of the

population across socio-economic groups (see Appendix 2 for further explanation of the health gini coefficient).

3.5. SUMMARY

In any study of equity in health and health care, three issues need to be considered seriously if the outcome of the assessment is to be informative. These are the issues of classification of social position, measurement of health status and quantification of the degree of inequality. As the purpose of assessing the magnitude of inequalities is to inform policy and planning, one needs to take into account both simplicity in computation and comprehension and precision of measurement. There is a likely trade-off between computational simplicity on the one hand, and precision and robustness of the measurement techniques on the other. Although the computation of the concentration index may not be straightforward , it has an appeal in the sense that it gives a composite and global measure of socio-economic inequalities in health or health care utilization, which is easy for comparison over a period of time. Unlike other related measures (e.g. the gini coefficient), it measures inequalities in health and health care that are related to the socio-economic status of the individual. This is an important attribute, as the concern of social policies is on those inequalities that are attributable to socio-economic status. The concentration index also has a visual appeal. The corresponding concentration curves from which concentration indices are computed are easy to understand, as one has to simply see for deviations from a diagonal line, which is the line of equality. Furthermore, by inspecting concentration curves for different time periods, it is possible to easily assess trends in socio-economic inequalities in health and health care (this presupposes that the curves do not overlap). Hence, the concentration curve and index will be the main techniques to be used in equity analysis in this study.

In this study, health status is measured by infant and under-five mortality, under-five child malnutrition and self-reported illness in the population aged 18 years and above. All systematic inequalities in health across various income quintiles, population groups and residential locations (urban/rural, province), are regarded as inequities as they are amenable to policy interventions. Thus, in the empirical chapters (Chapters 4 and 5) on equity in under-five mortality and malnutrition, inequalities that are systematically related to the household's income status, population group and residential location are considered as inequities, because of the fact that these can be avoided with appropriate social policies and interventions. Similarly, with respect to reported adult illness and service utilization, inequalities related to the individual's/household's income status are regarded as inequities. Furthermore, inequities in access are said to exist if there are income-related inequalities in the use of health services among those who have reported sickness.

Having this background, the next three chapters will address the issue of equity in health and health care utilization at the various stages of the life course using different measures of ill-health. Chapter 4 will be devoted to the assessment of child health with special reference to infant and under-five mortality. Chapter five will examine equity in child health from a different perspective. This will focus on one of the sensitive measures of child health – under-five child malnutrition. The purpose of including this is to have a comprehensive view, as mortality alone may not capture the whole story. Chapter 6 will complete the discussion on equity in health by examining equity in adult illness and utilization of services.

CHAPTER 4
EQUITY IN CHILD SURVIVAL

4.1. INTRODUCTION

The interest in socio-economic inequalities in health has been on the upsurge with the renewed commitment of governments and international organisations to improve the health of the poor (Gwatkin 2000, Wagstaff 2000). Socio-economic inequalities in health have been high on the policy-making agenda since the 1980s, particularly, after the publication of the Black Report in the United Kingdom (Pereira 1990, Carr-Hill 1987). The performance of a country's health system is no longer assessed only in terms of its average achievement on health indicators. The distribution of health among different socio-economic groups is also regarded as a key assessment criterion.

Socio-economic inequalities in health are pervasive in the developed world (Kunst *et al* 1995). Morbidity and mortality rates have been found to be higher in people with lower SES as measured by proxy variables such as education, occupational status, and income (Eachus *et al* 1999, Kunst and Mackenbach 1994, Lahelma *et.al* 1994, Mastilica 1990, Regidor *et al* 1999). The situation in developing countries is no better than this. The few studies available have indicated the existence of morbidity and mortality gradients between people with different socio-economic positions and ethnicity/race (Gilson and McIntyre 2001, Brockerhoff and Hewett 2000).

The decline in mortality witnessed in the second half of the last century is regarded as one of the most highly esteemed achievements of mankind since the advent of the industrial revolution (Schultz 1993, Feachem 2000). The rates of infant and child mortality have been registering remarkable decreases for several decades (Stanton 1994, Hojman 1996). However,

in Sub-Saharan Africa, the rapid fall seen over the period 1960-1980 was followed by a major slow down in the 1980s and 1990s (Cornia and Mwabu, 1997).

Nevertheless, the overall success achieved to date is threatened by the prevalence of widespread systematic unjust and avoidable inequalities, both within and between countries. Globally, the poorest and socially disadvantaged groups bear the greatest burden of ill-health in terms of morbidity and mortality. A snapshot of the rate of under-five mortality (U5MR) between developed and developing countries as well as between different socio-economic groups within countries attests to this. Under-five mortality in the low-income countries of Sub-Saharan Africa stands at a staggering rate of 173 per 1000, as compared to only 6 per 1000 in the industrialised countries (UNICEF 2000). The infant mortality rate (IMR) among the Black population in South Africa is more than five times that of their White counterparts (Gilson and McIntyre 2001).

The IMR and U5MR affect life expectancy at birth and, on their own, are regarded as key indicators of a population's health (Whitehead and Drever 1999, Wagstaff 2000). Consequently, socio-economic gradients in these important measures are considered as unacceptable, as they are largely avoidable and call for urgent action. Studies have indicated the presence of high mortality differentials in populations where the distribution of income is highly unequal regardless of the average level (e.g Wolfson *et al.* 1999, Chiang, 1999 Kennedy *et al* 1998, Waldman 1992). Countries with highly unequal income distributions have higher rates of infant mortality compared to countries with similar levels of per capita GDP but more equal income distributions (Flegg 1982, Rogers 1979). A cursory glance at the case of two countries will suffice to illustrate this. In South Africa where the ratio of the income share of the richest 20 percent of the population to that of the poorest 40 percent is about seven times, the IMR and U5MR are respectively in the order of 60 and 83 per 1000. In contrast in

Tunisia where the income share of the highest 20 percent is less than three times that of the poorest 40 percent, IMR and U5MR stand at 25 and 32 respectively (UNICEF 2000). It should also be noted that although both countries are classified as upper middle-income countries, the GNP per capita of South Africa is higher than that of Tunisia.

The inverse relationship between socio-economic status and morbidity and mortality at all stages of life has been well established. Under-five mortality is concentrated among the poor and socio-economically disadvantaged groups of society regardless of the overall wealth or poverty of the country (Waldman 1992, Stanton 1994, Wagstaff 2000, Gilson and McIntyre 2001). The case of South Africa is no exception. A study in the Western Cape Province (one of the "richest" provinces) has, consistent with expectation, documented inequalities in infant mortality related to race and place of residence. The IMR for Black (interchangeably used with African) South Africans in this province, which was of the order of 33 per 1000 live births is about three times that of their White counterparts (Bachmann *et al* 1996)

South Africa faces a great challenge to redress injustices perpetrated during the previous political system. The gap in health and development indicators between the haves and have-not's across the racial and socio-economic spectrum is immense and has been of grave concern to the new government in its initiatives to bring about social justice. Cognisant of the prevailing socio-economic disparities in child survival, the government's White Paper for the transformation of the health system (South Africa 1997) emphasises not only reducing the average values of infant and child mortality, but also bridging the inequalities in mortality. Policies that are designed to redress avoidable and unacceptable inequalities need to be based on a solid information base. To this end there is a need for a continuous assessment of the magnitude and dynamics of the state of equity in health, so as to be able to gauge the impact of public policies and interventions targeted at reversing the trend.

In the post-1994 period, South Africa has designed and implemented a number of comprehensive and sector-specific policies to uplift the welfare of its people including health, and rectify the deeply rooted inequities that were inherited from the previous government. An example of the latter is the policy of free health care to children under six years of age. It is therefore, of paramount importance to assess the dynamics of inequity in under-five mortality to scrutinise the effectiveness of the relevant policies that have been put in place to date.

This chapter aims at empirically assessing the degree of inequity in health status with particular reference to infant and child mortality, its correlates and temporal changes with a view to identifying relevant policy instruments. The specific objectives of this chapter are to:

i. assess and quantify the magnitude of inequalities in infant and child mortality that are systematically related to socio-economic status;
ii. identify the various socio-economic and demographic factors that influence U5MR and IMR; and
iii. evaluate temporal changes in socio-economic inequalities in IMR and U5MR.

4.2. METHODS

4.2.1. SOURCE OF DATA

The analysis makes use of data from the October Household Survey (OHS) of 1998. The OHS is an annual sample survey of South African households conducted by Statistics South Africa. It is based on a multi-topic questionnaire that includes questions related to household welfare. These include aspects such as, demography, socio-economic conditions, access to services and amenities, health and perceived quality of life. The OHS-98 was based on a sample of 20,000 households, which comprised about 100,000 individuals (Statistics South Africa 2000). Comparisons are then made with findings of a study based on the Living Standards and Development Survey of 1993 (Wagstaff 2000). Although the LSDS and OHS 98 data may have

had differences in instruments used for data collection, subtle differences in the phrasing of questions wouldn't preclude conducting analysis of changes over two time periods sufficiently apart from each other. Comparison of data in the OHS series may be the ideal one. However, comparing data from OHS 95 and OHS 98 may not give a clear picture on the effectiveness of policies put in place, as the time interval between the two is too short to enable us to notice changes in under-five mortality brought about as a result of government efforts. Hence with all the precautions, this chapter tries to gauge trends in inequities in child survival based on findings from LSDS 1993 and OHS 1998 data.

4.2.2. MEASUREMENT OF INFANT AND UNDERFIVE MORTALITY

Infant and under five mortality rates were calculated using the direct method by taking the numbers of live births and deaths of children under-one and under-five years of age for the period covered by the survey. The direct method of measurement is based either on vital registration or on dated vital events from retrospective birth and death histories. The indirect method uses the number of children ever born and the proportion dead, classified by five-year age groups of mothers (Adetunji, N.D.).

In South Africa, as is the case in many developing countries, the absence of an effective vital registration system (Nannan *et* al 1998) leads to heavy reliance on surveys such as the Demographic and Health Survey, the October Household Survey and others. The quality of mortality estimates computed from retrospective birth histories collected through surveys, however, depends on the completeness with which births and deaths are reported. The most serious data quality problem is omission of reporting of births and deaths by some respondents in these surveys. Other problems include misreporting of birth date as well as age at death (DoH 1998). These problems may result in over- or underestimation of mortality, and thus distort the real picture. Nannan *et al* (1998) argue that because of these problems with

the vital registration system and surveys, the level of infant mortality rate in South Africa is not known. Thus, it would only be possible to make better of conclusions from relatively accurate data obtained by improving the design of surveys and vital registration systems.

Ideally, mortality rates computed using the two different methods (direct and indirect) are not expected to be significantly different from each other. However, studies in Africa indicate that the indirect methods produce significantly higher mortality rates than the direct method (Adetunji, N.D.). This implies that analysis of trends that combines results from the two methods may lead to erroneous conclusions and policy recommendations, and thus, caution needs to be exercised. The analysis of LSDS data by Wagstaff (2000) that is used for comparing results from the OHS 98 employs the indirect method in computing mortality rates. This is a drawback of the analysis presented later.

4.2.3. THE MEASUREMENT OF INEQUITIES

In measuring inequities the concentration curve/index and rate ratios as described in Chapter 3 will be employed. The mortality concentration curve depicts the cumulative proportion of infant/under-five child deaths on the vertical axis against the cumulative proportion of children (births) at risk on the horizontal axis, ranked according to their households' income and beginning with the most disadvantaged child. The concentration index is derived from the concentration curve.

Besides, rate-ratios are used to assess inequalities in IMR/U5MR related to attributes other than the household's economic position (e.g. race, residential location *etc*). These compare the rates of the worse off group with those of the relatively well off in the given attribute. The main measure of socio-economic status in this case is total household income. Although

household expenditure might have been preferable to income as a measure of socio-economic position, it is not available in the OHS data.

In measuring welfare in the context of developing countries, there is a strong case in favour of using measures based on consumption (expenditure) rather than those based on income (Deaton 1997). Having a satisfactory estimate of living standards requires measuring the household's annual income. This entails multiple visits to the household or the use of recall data. In the case of the OHS, it is the later that is used. In addition to errors that may be introduced due to poor recall, measuring income, especially in rural households whose incomes are derived mainly from self-employment in agriculture is difficult. In contrast, a consumption/expenditure measure can rely on consumption over the previous few weeks, thus minimising errors due to poor recall (*ibid*). Moreover, household expenditure on basic needs of life reflects the household's resource endowment, and therefore is regarded as a good measure of the health status of the household members (Glewwe 1991, Alderman 1993). Thus, this limitation of measuring socio-economic status using income has to be acknowledged, as it is likely to bias the results.

It should be noted that in OHS 98, the question on household income is closed-ended and contains income bands that are wide and get wider as income increases. For example, in the income bracket R4,501 – R6,000, a person whose earning is at the lower limit (R4,501) and another one who is .on the higher limit (R6,000) will all be regarded as equal with respect to their income. This is likely to affect the classification of households into the various income quintiles, and thus result in misclassification and a consequent blurring of the picture of mortality differentials related to household income. This limitation calls for caution when interpreting the results of the study.

The assessment of trends in inequities in childhood mortality in this study involves comparison of findings from OHS 1998 data to those of LSDS 1993 as reported by Wagstaff (2000). However, as there are methodological differences both in the study designs, computational methods and measures of socio-economic status, it is very important to keep in mind the flaws of the assessment and scrutinise the results of the comparison from various angles before proposing policy recommendations.

4.2.4. THE ECONOMETRIC MODEL

A probit model is estimated to further elicit mortality differentials related to some attributes of the individual and the household. The results will highlight effects of discrete changes of the independent variables on the probabilities of infant and child mortality, that is, as one moves from one category to the other (e.g. from quintile 1 to the subsequent quintiles, or from the African population group to the others, etc). The dependent variables, infant mortality and child mortality are dichotomous dummy variables. In the presence of infant/child death, they assume the value of 1; otherwise they are equal to zero.

The ordinary least squares (OLS) method is a commonly applied statistical technique. However, when the dependent variable is categorical rather than continuous, the OLS method becomes an inefficient technique and the underlying linear probability model (LPM) that is being estimated represents a poor *a priori* choice of model specification (Aldrich and Nelson 1984). Furthermore, the LPM does not constrain the probability of an event between the values of 0 and 1 and assumes that the effect of the independent variable is constant across different predicted values of the dependent variable.

The above problems of the OLS are remedied by categorical data models (also called discrete choice models). The probit model is one of those models and is based on the normal

cumulative density function. Probit coefficients represent effects of one unit change in the independent variable on cumulative normal probability of the dependent variable, that is, they are effects on Z (standard) scores. However, since the effect of a unit change in the independent variable on the value of the dependent variable depends on the level of the independent variable, the standard approach is to examine the effects of the independent variable on the probability of the dependent variable when all other independent variables are held at their mean values.

For ease of interpretation, it is preferred to present probabilities instead of coefficients. In probability form, the probit model may be defined by (Liao 1994):

$$\text{Prob}(y=1) = \Phi\left(\sum_{i=1}^{n} \beta_i x_i\right) \tag{4.2}$$

where

$\text{Prob}(y=1)$ is the probability of the event (infant/under-five mortality) occurring;

$\Phi(.)$ represents the standard normal cumulative distribution function;

x_i is a vector of explanatory variables; and

β_i is a vector of estimated parameters.

The marginal effect of each of the independent variables on the probability of infant/under-five mortality is given as:

$$\frac{\partial \text{prob}(y=1)}{\partial x_i} = \varphi\left(\sum_{i=1}^{n} \beta_i x_i\right)\beta_i \tag{4.3}.$$

This will imply for example, the change in the probability of infant or under-five child death as one moves from income quintile 1 to say, income quintile 2, or from an urban residential location to a rural one. Because the result of $\phi(\cdot)$ is a function of all the *x*'s, we can only

compute marginal effects by assigning certain values to the *x*'s (Liao 1994). This is done by keeping all variables (except the one under consideration) at their mean or modal values (if categorical variables).

The definition, measurement and expected sign of each of the independent variables is presented in Table 4.1.

Table 4.1
Explanatory variables: definition, measurement and expected sign

Independent variable	Definition and measurement	Base ccategory	Expected sign
Quintile	Socio-economic status measured by household income quintile: Q_2 = 1 if household belongs to expenditure quintile 2 = 0 otherwise Q_3 = 1 if expenditure quintile 3 = 0 otherwise Q_4 = 1 if expenditure quintile 4 = 0 otherwise Q_5 = 1 if expenditure quintile 5 = 0 otherwise	Q_1	-
COLOURED	= 1 if the child belongs to the population group "Coloured" = 0 otherwise	African	-
WHITE	= 1 if White = 0 otherwise	African	-
Location	Dummy for urban-rural status: = 1 if the household is located in rural areas = 0 if URBAN	Urban	+
House ownership	= 1 if owning a house = 0 if not owning	Not house owned	-
Aggregate Poverty level in the province[1]	Above-average= 1 = 0 otherwise	Below average poverty level	+
Gender	GIRL = 1 = 0 otherwise	Male child	+
Distance from health facility	= 1 if distance to health facility >=5 Km = 0 otherwise	Distance <5 Km	+

In the above table, the continuous variable, *income,* is transformed into a discrete variable to allow for a highly non-linear effect. Furthermore, *aggregate poverty level in the province* and

distance to a health facility are transformed into discrete ones for ease of interpretation and to take account of some targets. For example, distance to health facility is dichotomised using the target distance of five Kilometres as a cut-off point for an operational definition of an acceptable physical access to a health facility (Bradshaw 1998). Thus, instead of assessing the effect of each kilometre distance from a health facility on the probability of infant/under-five child mortality, we will see in aggregate what the effect of having or not having reasonably good physical access to health care would be on the probability of infant or under-five mortality. Similarly, with respect to the levels of poverty, by classifying the provinces into poor and non-poor, it is hoped that the analysis will make it simpler to understand by quantifying the change in the probability of mortality as one moves from those designated as poor to those that are relatively non-poor.

Although the mechanisms through which household income status influences under-five mortality are intricate, evidence suggest that a household's economic position is negatively related to infant/under-five mortality. For example, the tremendous improvements in infant and child mortality that were observed in sub-Saharan Africa did not maintain momentum after the 1980s when the region started experiencing a slackening of economic performance (Cornia and Mwabu 1997). Other studies have found that income does not have a linearly negative effect on childhood mortality, and that the link is not so strong and subject to diminishing returns (Hojman 1996). Furthermore, countries with marked income inequality have higher rates of childhood mortality than countries with the same per capita income but lower income inequality. Given the above evidence, it is hypothesized that a movement from the poorest household income quintile to the subsequent quintiles will lead to a drop in infant

[1] To formulate this variable, the head-count ratio was computed for each province. A dichotomous variable was then created from the results, where those provinces with a head-count ratio above the calculated national average were classified as relatively poor and those below the national average, otherwise.

and child mortality. Hence a negative sign is expected for the coefficients of the variables quintile 2 through quintile 5.

The population group to which the child's household belongs is an important correlate of the child's health status. Brockerhoff and Hewett (2000) have revealed widespread ethnic inequalities in under-five mortality in sub-Saharan Africa The findings of a study in the Western Cape province of South Africa also demonstrate a greater burden of infant mortality in population groups regarded as disadvantaged (Bachman *et al* 1996). Hence it is expected that compared to the base category, the African population, infant and under-five mortality rates will decline in Coloured and White population groups. Therefore the signs of the coefficients of the variables Coloured and White are expected to assume negative signs. It should be noted that the Indian population group is excluded from the analysis because of the smallness of the sub-sample.

With respect to the effects of residential location, there is conflicting evidence. On the one hand, due to income and educational differentials that favour the urban areas (Garrett and Ruel 1999), one may expect a negative relationship between childhood mortality and degree of urbanization of area of residence. However, on the other hand, as a result of massive rural-urban migration that is typically observed in developing countries, there is an increase in urban poverty with a resultant negative effect on child health. Even though the influx of the rural population to urban areas may also be a problem in South Africa as much as it is in other developing countries, poverty is still concentrated in rural areas. The 1998 *Poverty and Inequality Report* (May 1998) indicates that 71 percent of the rural population is poor compared with only 29 percent of the urban population. It is, thus expected that probabilities of infant and under-five mortality will increase as one moves from the base category (urban

area) to rural settings. Therefore, a positive sign is expected of the coefficient of the variable location.

Although house ownership may be correlated with household income, the question asked does not indicate the type of house owned. It is not possible to distinguish between ownership of a house with poor quality or one with high quality. Thus, *a priori*, no discernible correlation would be expected between household income and ownership of a house. Nevertheless, in a relative sense, one may expect those poor people who own a house (even if a shack) to be better off than their counterparts who don't own a similar one. It is therefore expected that the probabilities of infant and child survival will decrease in households who own a house compared to those who don't. A negative sign is therefore expected of the coefficient of the variable representing house ownership (base category is non-ownership of a house).

The aggregate level of poverty in a province as measured by the head-count ratio, may among other things, proxy community-level physical and social environmental factors that have a bearing on health. Morbidity and mortality are higher in communities that are poverty-stricken and without amenities that are necessary for health promotion. It is hypothesized that infant and child mortality are highly concentrated in communities whose levels of poverty are above the national average. Hence a positive sign is expected of the variable representing the aggregate poverty level (reference category is below national head-count ratio).

In many developing country societies, preferential treatment is given to male children. In the Asian continent for example, female children experience higher mortality compared to their male counterparts, as a result of preferential feeding and greater access to health care for male children (Stanton 1994). We therefore expect gender differentials in infant and child

mortality favouring male children. In this case, a positive sign is expected of the coefficient of gender.

The remarkable global decline in under-five mortality observed up to the early 1990s, among other things, has been attributed to improved access to basic health services, development of vaccines and improved living conditions (Gelband and Stansfield 2001). The probability of infant and under-five mortality is expected to be less in areas which are in close proximity to a health facility compared to those which are not. Hence a positive coefficient is expected (base category being distance less than 5 Km).

4.3. RESULTS

4.3.1. GENERAL

The infant mortality rate for South Africa, as computed from the OHS 98 data (55 per 1000 live births) lies between the figures obtained from the 1998 South Africa Demographic and Health Survey (SADHS) (45 per 1000) (DOH *et al* 1998) and UNICEF's *The State of the World's Children 2000* (60 per 1000) (UNICEF 2000). Although the rate computed in this study from the OHS-98 is based on a one-year period (live births and deaths of under one year children in the last 12 months), its closeness to those of other studies and reports lends it credibility. All the above-mentioned figures are way below those computed from the Project on Living Standards and Development Survey (LSDS) of 1993 that puts the IMR at 74 per 1000 live births (Wagstaff 2000). Similarly, the U5MR computed from the OHS-98 data (70 per 1000) again lies between those estimates provided by the SADHS (59 per 1000) and UNICEF (83 per 1000). The U5MR computed from the LSDS data reveals a high level of the order of (113 per 1000).

Mortality figures from data sets obtained through different study methodologies or computational techniques are likely to show discrepancies. This makes analysis of trends very complicated, especially in developing countries where the vital registration system which can give relatively better figures is deficient. The following table presents some figures for IMR and U5MR in order to highlight the difficulties involved.

Table 4.2
IMR and U5MR figures from various studies

Author (year)	Year referred to	Method of calculation (where stated in the original study)	Data set	Value
IMR/1,000				
UNICEF (2000)	1960	Not stated	Not stated	89
Gilson and McIntyre (2001)	1993	Direct	LSDS 1993	61
Wagstaff (2000)	1993	Indirect	LSDS 1993	74
Nannan et al (1998)	1993	Not stated	LSDS 1993	81
DOH et al (1998)	1998	Not stated	SADHS 1998	45
Current study	1998	Direct	OHS 1998	55
UNICEF (2000)	1998	Not stated	Not stated	60
U5MR/1,000				
UNICEF (2000)	1960	Not stated	Not stated	130
Wagstaff (2001)	1993	Indirect	LSDS 1993	113
DOH et al (1998)	1998	Not stated	SADHS 1998	58
Current study	1998	Direct	OHS 1998	70

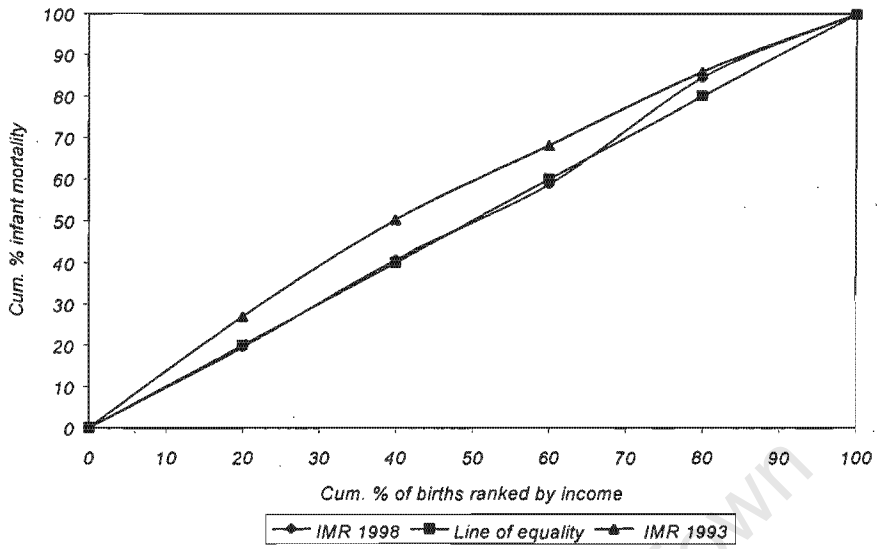
As can be seen from the above table, although the IMR and U5MR figures show differences, the overall trend seems to be that of decline, even though the rates of decline may not be consistent as will be presented in the discussion section that will follow.

4.3.2. INEQUITIES IN INFANT MORTALITY

The presence of inequities in infant mortality is assessed by comparing infant mortality concentration indices estimated from two different data sets for two different time periods. Temporal comparison of the concentration indices is of paramount importance as it may reveal changes that could possibly have taken place as a result of the various policies and interventions of the government

The infant mortality concentration index estimated from the 1993 data indicates inequalities that are against the poorest segments of the South African population. The concentration index is significantly different from zero (see Table 4.3), implying the presence of notable inequities in infant mortality related to the household's adverse socio-economic position. This pro-rich inequality, is however, not observed in the 1998 data. The concentration index, although apparently pro-poor, it is not significantly different from zero. This indicates the absence of income-related inequities in IMR. For ease of comparison, the concentration curves and indices for the two periods are presented respectively in Figure 4.1 and Table 4.3.

Figure 4.1
Infant mortality concentration curves



The above concentration curves clearly indicate the change in relative inequality in infant mortality rate that has occurred during the two time periods. While the curve for 1993 indicates a marked deviation from the line of equality, which is adverse to the poorest, the curve for 1998 seems to overlap with the diagonal line signifying the absence of noteworthy socio-economic inequalities in infant mortality. The information depicted by the concentration curves is presented in summary form by concentration indices in Table 4.3.

Table 4.3
Infant mortality rates and concentration indices- 1993 and 1998

		IMR 1993*	IMR 1998
Quintile 1		97.3	53
Quintile 2		83.7	54
Quintile 3		64.3	51
Quintile 4		64.0	69
Quintile 5		51.0	47
Overall average		74.1	55
	C	-0.125	0.003
	t-ratio C	-6.872	0.057

*data derived from Wagstaff (2000)

The concentration indices and their t-ratios above indicate that the pro-rich concentration index of 1993 is significantly different from zero, revealing the presence of a higher burden of infant mortality in the socio-economically disadvantaged groups. In contrast, the concentration

index for 1998 is not different from zero. Thus, it seems that the glaring inequality in infant mortality that existed previously seems to have been narrowed in the later years. As was noted previously, although the IMRs from OHS 1998 are calculated directly from the numbers of births and deaths in a period of 12 months prior to the survey date, their closeness to those of other reports (e.g. SADHS in 1998 and UNICEF's figures for 1998) add to the credibility of the figures. Inspection of the above table thus, reveals that the narrowing of inequalities was not a deterioration of the rates of those who enjoy favourable health conditions, but it was rather brought about by an improvement for those with the worst IMR figures. Although there are declines in IMR in most of the expenditure quintiles, the rate of decrease was more pronounced in the poorest. The rate of decrease in IMR in the poorest quintile is about 44 percent as compared to only about 8 percent in the richest quintile.

The racial disparity in IMR in 1998 is still wide. The Indian population group seems to have the highest rate (74 per 1000). However, this is likely to be due to a small sub-sample of this population group in the study. After omitting the Indian sub-sample, The African population group seems to be the worst-off in terms of infant mortality. Table 4.4 presents the IMRs, and rate-ratios with their corresponding 95 percent confidence intervals for different population groups and geographic areas.

Table 4.4
Rate-ratios and 95 percent confidence intervals for IMR-98

Attribute	IMR	Rate-ratio	95% confidence interval	
			Lower	Higher
Population group				
African*	58	1.0		
Coloured	41	1.4	0.9	2.3
White	11	5.3	1.8	25.7
Place of Residence				
Rural*	64	1.0		
Urban	44	1.45	1.2	1.8
Province				
Eastern Cape*	78	1.0		
Free State	65	1.2	0.7	2.0
Gauteng	41	1.9	1.2	3.2
Kwazulu Natal	72	1.1	0.8	1.5
Mpumalanga	67	1.2	0.8	1.6
North West	58	1.3	0.9	2.0
Northern Cape	49	1.6	0.9	3.0
Northern Province	25	3.1	2.0	5.0
Western Cape	13	6.2	2.7	17.4

* reference group

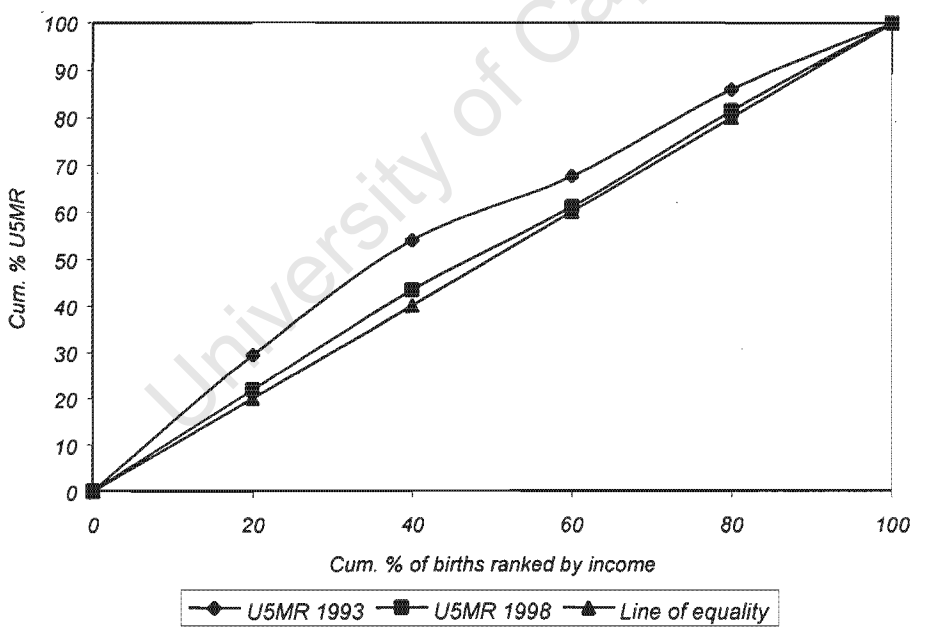
The figures in the above table indicate the existence of significant spatial and population-group disparities in IMR. The rate-ratios in the three provinces -Gauteng, Northern Province and the Western Cape- are statistically different from 1, implying the presence of a significantly higher IMR in the Eastern Cape Province compared to the three mentioned provinces. In the rest of the provinces, IMR's do not seem to be significantly different from those of the Eastern Cape Province. Furthermore, rural areas seem to bear a heavier burden of IMR compared to the urban areas. The data, however, need to be viewed with caution, as there seems to be a counter-intuitive result for the Northern Province – a relatively low IMR despite being one of the poorest provinces. This also highlights the shortcoming of computing infant and child mortality rates from routine household survey data such as the OHS, as there is likely to be under-reporting of child deaths especially by the poor (Wagstaff 2000).

4.3.3. INEQUITIES IN UNDER-FIVE MORTALITY

Inequities in under-five mortality were assessed on the basis of the two data sets as described in the case of infant mortality. Under-five mortality includes both infant (under-one) and child (1-4) mortality rates. Hence the impact of the IMR is also to be felt in this case. Whatever differences may be observed between IMR and U5MR are to be attributed to the second component of the U5MR, child mortality rate (CMR).

As observed in the case of inequalities in IMR, income-related pro-rich inequalities are more pronounced in 1993 than they are in 1998. A graphical illustration is presented in Figure 4.2 for a clearer understanding of the situation.

Figure 4.2
Under-five mortality concentration curves



In Figure 4.2, it is observed that the curve for 1993 deviates from the line of equality markedly as opposed to the one in 1998, which virtually overlaps with the line of equality. Thus, while Panel *A* reveals remarkable income-related inequalities that are to the detriment of the poor,

Panel *B* reveals the relative absence of such inequalities. The latter implies that there is virtually no concentration of under-five mortality among the poorest segments of the population, as was seen in 1993. The rates and corresponding U5MR concentration indices that are computed on the basis of the above U5MR concentration curves are depicted in Table 4.5.

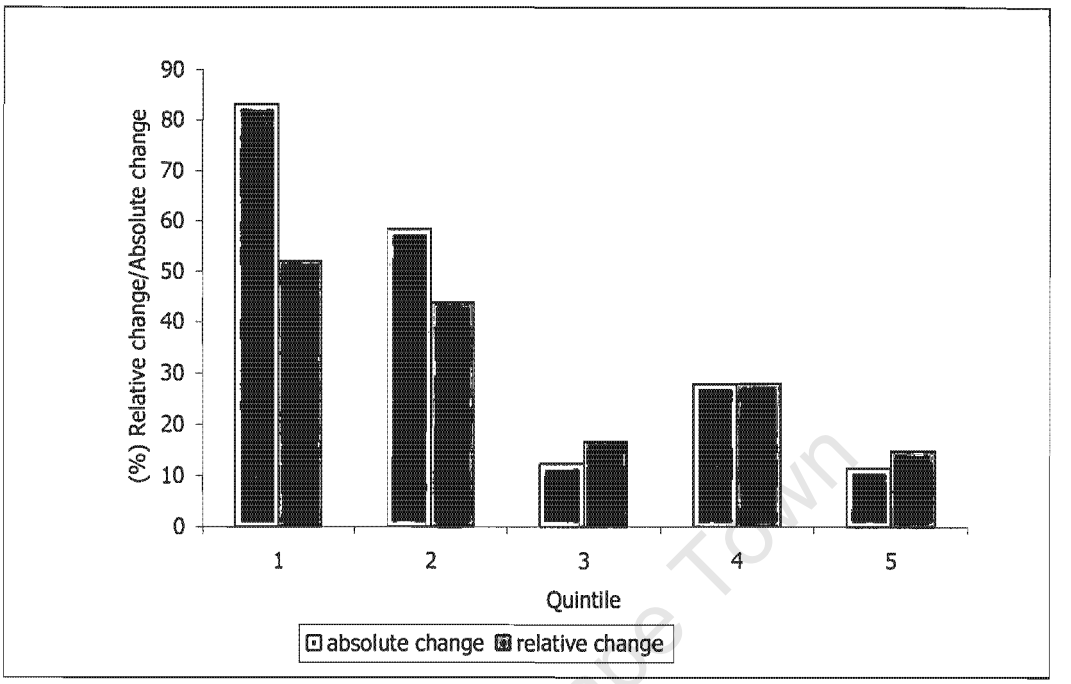
Table 4.5
Under-five mortality rates and concentration indices- 1993 and 1998

Quintile	U5MR 1993*	U5MR 1998
Quintile 1	159.7	76.5
Quintile 2	133.3	75.0
Quintile 3	74.5	62.2
Quintile 4	99.3	71.4
Quintile 5	76.7	65.2
Overall average	108.7	70.2
C	-0.147	-0.03
t-ratio *C*	-2.840	-1.562

*rates derived from Wagstaff (2000)

Although the concentration indices for both time periods indicate inequalities that favour the rich, the index for 1998 is not significantly different from zero. Hence the remarkable income-related pro-rich inequalities in U5MR that were witnessed in 1993 seem to be reduced in the later period. The above table further reveals that the narrowing down of the inequalities was a result of a reduction in the U5MR across all expenditure quintiles, with the highest relative and absolute decreases seen in the poorest segments of the population. This is further elucidated graphically in Figure 4.3, which depicts the absolute and relative decreases of U5MR in both time periods.

Figure 4.3
Decreases in U5MR, 1993/1998



As can be seen from the figure above, there was a reduction in U5MR across all socio-economic groups. However, the magnitude of the decline was very remarkable in the poorest 40 percent of the population, as compared to the richest.

The anomalous result in both infant and under-five mortality rates in income quintile 4 is a point of concern. There is either no decrease or an increase over the next lower income quintile. This is seen both in Wagstaff's analysis of the LSDS 1993 (Wagstaff 2000) and the current analysis using OHS 1998 data. This may be a genuine trend that needs to be explored further, or it may also indicate flaws in the data being used. It may possibly be attributed to the wide income bands used in the data sets that may result in misclassification of households into the various income quintiles. To identify other possible causes for this counter-intuitive result more probing will be required in future studies.

4.3.4. SOME FACTORS INFLUENCING INFANT AND UNDER-FIVE MORTALITY

Changes in probabilities of infant and under-five mortality were estimated using a probit regression model and the OHS 1998 data. As discussed earlier, probabilities are presented instead of coefficients for ease of comprehension and a better understanding of the magnitude of each independent variable on infant and under-five mortality.

The results reveal striking differences in the effects of the explanatory variables on the two dependent variables, infant mortality and under-five mortality. The estimation results are presented in Table 4.6.

Table 4.6
Probit estimation results

Variable	Infant mortality		Under-five mortality	
	$\frac{dF}{dX}$	P-value	$\frac{dF}{dX}$	P-value
Quintile 2	0.0011	0.894	0.0016	0.777
Quintile 3	0.0007	0.939	-0.0119	0.032
Quintile 4	0.0192	0.037	-0.0014	0.706
Quintile 5	0.0051	0.612	0.0003	0.842
Coloured	-0.0112	0.297	0.0124	0.149
Indian	0.0377	0.171	-0.0385	0.004
White	-0.0412	0.008	-0.0351	0.000
Rural	0.0171	0.005	-0.0027	0.355
Own house	-0.0131	0.094	-0.0115	0.006
poverty	-0.0021	0.740	0.0122	0.039
Girl	0.0025	0.649	-0.0102	0.006
Distance >5Km	-0.0081	0.801	-0.0122	0.576
$LR\ \chi^2_{15}$	35.20		82.29	
$P-value$	0.0014		0.0000	

In both models the goodness of fit statistics as measured by the likelihood ratio χ^2 test indicate that at least one of the coefficients is significantly different from zero, implying that the model is a good fit. The expression $\frac{dF}{dX}$ represents the change in the probability of an

infant/under-five death for a discrete change of the explanatory variables from 0 to 1, as all are dummy variables. The regression results for infant mortality indicate that rural location and being a member of the White population group have a statistically significant effect. The signs of the estimates of the two variables are consistent with *a priori* expectation and are in line with the findings of the bivariate measures (rate-ratios) presented previously (See Table 4.4). The estimates imply that compared to the African population group, the probability of an infant dying decreases by 41 per 1000 in Whites, when all the other variables are held at their mean values. Furthermore, the probability of infant death increases by 17 per 1000 in rural areas as compared to urban settings. As discussed previously, there is a counter-intuitive result with respect to the effect of income status (quintile), that is, the probability of infant mortality increases significantly in the income quintile 4 as compared to the poorest quintile. This defies a logical explanation and seems to be equivocal, as it does not show consistency. For example, no statistically significant change in probability is seen when one compares the poorest quintile with the richest (quintile 5).

In the case of child mortality, the place of residence (rural/urban) does not have a statistically significant effect. However, the number of variables that have turned to be statistically significant is greater. The probability of under-five mortality is seen decreasing significantly in the third expenditure quintile. As mentioned above, this seems equivocal, as it does not exhibit the same effect in the subsequent two richest quintiles. The probability of under-five mortality also decreases by a significant amount (35-39 per 1000) in Whites and Indians. Other variables with statistically significant effects are: house ownership, relatively high levels of poverty in the province of residence and gender.

Ownership of a house has an inverse relationship with under-five mortality. The probability of an under-five death decreases by an amount of 11 per 1000 in children from households who own a house. The results also suggest that under-five mortality is significantly lower in girls.

To assess the impact of provincial poverty levels on the probability of under-five mortality, a crude measure of poverty, the *head-count ratio,*[2] was computed for each province. The probit estimates indicate that under-five mortality significantly increases (an increase of about 12 per 1000) in those that are designated as relatively poorer provinces. These include, the Eastern Cape, Free State, North West, Mpumalanga and Northern Province.

4.4. DISCUSSION

This chapter examines inequities in health with reference to mortality of children under the age of five. It attempts to quantify the magnitude of avoidable, and unjust differentials in infant and under-five mortality that are related to the socio-economic status of the household as measured by total household income. Temporal comparison is also done to assess the changes witnessed over time. This has a special significance, as it traverses a period of structural break in the South African political system.

A significant reduction in socio-economic inequalities in IMR is witnessed over the period 1993-98. The rates for the two periods indicate not only bridging of the gap between the richest and the poorest, but also decreases in absolute IMR levels across all income groups. IMR has decreased by an average annual rate of about 5.8 percent. Different reports and studies give completely different figures of infant and child mortality rates depending on differences in study designs and computational methodologies, making temporal comparison of under-five mortality rates very difficult. Thus, in interpreting such results it is necessary to exercise

[2] Regarded in this case as the proportion of the population below R 250 monthly expenditure, which is half of the overall median of R 500.

caution and apply one's discretion about the facts on the ground. For example using IMR figures for 1993 (60 per 1000) from Gilson and McIntyre (2001), and the 1998 IMR figure of this study, the average annual rate of reduction in IMR would be only 1.7 percent. If the figure from Gilson and McIntyre is, however, compared with the figure from SADHS, an average annual rate of reduction of about 5.6 percent is obtained. Although Wagstaff's figures for IMR in 1993 (Wagstaff 2000) may be high, as the indirect method of computation was used, comparison of the IMR figure for 1998 from this study with those of Gilson and McIntyre who used the direct method for the LSDS 1993 data set also indicates a decrease in IMR, albeit marginally. Thus, the trend in the average figure appears to be that of a marginal decline in IMR over the stated period of time.

Differences in study design and methodology used to calculate mortality rates are likely to lead to different conclusions and policy recommendations. Mentioning some of the IMR estimates for South Africa will suffice to make the discrepancies clear. While an IMR of 11 per 1000 is calculated from OHS 1994, IMR from LSDS 1993 is relatively high (81 per 1000) (Nannan 1998). UNICEF gives IMR figures of 71 and 52 per 1000 for 1992 and 1994 respectively. According to the Institute of Futures Research (1996) the IMR for the period 1991-96 was 56 per 1000. The Development Bank of Southern Africa, on the other hand, puts the IMR for 1990-95 at 46 per 1000. While Gilson and McIntyre (2001) report an IMR of 61 per 1000 from the LSDS 1993 data, Wagstaff reports a rate of 74 per 1000 (compared to 81 above). Part of the reason for this discrepancy is that Gilson and McIntyre used the direct method and Wagstaff the indirect method of calculating IMRs. Checking the credibility of these figures is difficult because of the inadequacy and incompleteness of the vital registration system in South Africa. It is perhaps for this reason that Nannan *et al* (1996) have argued that the IMR for South Africa is not known. Thus, it appears that conclusive comparisons of trend data are difficult to make given differences in study designs as well as computational methods. It

should therefore be noted that the findings of this study have to be seen as crudely indicative and not conclusive.

Based on five-year retrospective estimates from surveys conducted by the Human Sciences Research Council and the 1998 SADHS, it is said that IMR and U5MR have shown improvements until 1991 (DOH *et al* N.D). However, if we are for example, to make comparisons between two time points, say 1980 and 1998, we will be able to see that there is a significant improvement in U5MR in 1998 compared to 1980. While the U5MR in 1980 was about 90 per 1000, the 1998 SADHS gives an U5MR of 58 per 1000 (*Ibid*).

The rate of decline, however, has to be examined by segmenting the various decades so as to factor in the possible effects of the HIV/AIDS epidemic. Figures obtained from UNICEF's estimates (UNICEF 2000, United Nations Statistics Division website) indicate that the U5MR showed an average annual rate of decrease of 1.8 percent during the two decades, 1960-1980 (from 130 in 1960 to 90 in 1980). During the decade of 1980-1990, the U5MR declined by a rate of 2 percent per annum (from 90 in 1980 to 60 in 1990), and during the period 1990-1998, an average annual decrease of 0.6 percent was observed (from 60 in 1990 to 58 in 1998). It can clearly be seen that although there is the decline in U5MR has continued, the higher rate that was experienced until the early 1990s has not kept momentum. A notable influencing factor that can be mentioned as a likely reason for the slackening of the rate of decline is the HIV/AIDS epidemic that has assumed alarming proportions in South Africa. In many countries in sub-Saharan Africa, the decline in child mortality has slowed, stopped or even reversed itself during the 1990s (Gelband and Stansfield 2001).

Similarly with IMR, although there are variations between the different estimates for specific years, there is an overall trend of improvement in the IMR. In 1960, IMR was estimated at 89

per 1,000 live births declining to between 61-81 per 1,000 in 1993 and between 45-60 per 1,000 (depending on the study – see Table 4.2).

Although mortality estimates derived from LSDS 1993 are high, and should be viewed with caution, analyses of IMR differentials related to socio-economic factors conform to expectation (Nanna 1996). Thus, even if changes in the absolute magnitude of childhood mortality over time need to be viewed with caution, relative inequalities related to socio-economic status may not be seriously affected to preclude the assessment of inequalities using concentration curves/indices. The concentration indices will only measure the magnitude of inequality related to income for each data set.

The 1998 finding (the absence of income-related inequalities in the mortality of the under-fives) appears anomalous. However, it is not unique to South Africa. Gwatkin *et al* (2000) in their analysis of socio-economic inequalities in health, nutrition and population have shown the absence of wealth-related inequalities in a number of countries. The case of neighbouring Namibia, which also shares the same apartheid history, may be cited here. The Namibian data show the absence of any significant wealth-related pro-rich inequalities in both infant and under-five mortality rates (*ibid*).

Inequalities seem to have diminished by declines in IMR across all expenditure quintiles, but more prominent in the poorest two quintiles. When setting targets to reduce inequities, there is a need to focus on *levelling-up*, and not *levelling down* (Whitehead *et al* 1998). This avoids the situation in which there is a notable narrowing in the health gap that is brought about as a result of the health of the well-off deteriorating, rather than an upward movement of those who are disadvantaged (*ibid*). The results of this study thus seem to suggest that there has been a levelling-up in South Africa, which is the desirable option.

The apparent absence of income-related inequalities in IMR in the 1998 data, however, raises some important issues. Although a modest decline in the Gini index seems to have been achieved, it is still high by international standards. Many studies have lent support to the conviction that relative inequality is a fertile ground for inequalities in health (e.g. Kennedy *et al* 1998, Stanton 1994, Waldmann 1992).

The absence of income-related inequalities in early childhood mortality in the latter year of analysis (1998), gives credence to the assertion of Ross *et al* (2000) that the link between mortality and income is not universal and that it is dependent upon the socio-political context. Furthermore, the psychosocial environment has a vital role to play in enhancing or extenuating the negative health effects of income inequality (Wilkinson 1996). This implies that the adverse health impact of inequality is mediated through one's relative position in the social hierarchy based on income.

The South African case may probably fit into this. The socio-political scenario that existed in 1998 is very different from that of 1993. The context that prevailed in the period pre-1994 was not health promoting for the overwhelming majority of the populace. It may perhaps be said that the low perception of their relative social position that the majority underprivileged people had during the apartheid era, might have been reversed in the latter period when the government that represents the majority has been in power for four years, thus mitigating the adverse effects of income inequality on mortality. This implies that political power in the hands of a government that represents the interests of the majority is likely to induce positive changes as people anticipate improvement of their lot.

The absence of income-related inequalities in childhood mortality, however, does not rule out the presence of inequalities related to other socio-economic attributes. Income is, but one of

the determinants of child survival. The results indicate that there is a gross inequality in IMR in relation to population group/race. Although there is no statistically significant difference in IMR between the African and the Coloured population groups, the difference between the Whites on the one hand, and Africans and Coloureds on the other hand, is substantial. Historically, the Africans and the Coloureds are generally regarded as the two most disadvantaged groups, albeit to differing degrees. Racial inequalities in infant and under-five mortality in South Africa are inequitable as they are related to discriminatory policies of access to resources which are unacceptable and avoidable.

The existence of racial differentials in IMR is in line with the findings of Bachmann *et al* (1996) whose study in the Western Cape Province showed that Black South Africans had relatively worse off IMR figures than Whites, and Coloureds. The current findings, however, do not reveal IMR gradients between the African and Coloured population groups. This may mean that inequalities, which are imperceptible when using aggregated macro-data, may show up when the data are disaggregated. Hence, whenever circumstances permit, there is a need for micro-level analysis in order to identify inequalities that may not be detected using aggregated data. Gilson and McIntyre (2001) using the 1993 LSDS data also indicated that the IMR in Black South Africans is more than five times that of their White counterparts. In a similar vein, Brockerhoff and Hewett (2000) in their study on inequality of child mortality among ethnic groups in Sub-Saharan Africa showed the presence of a significant mortality differential among different ethnic groups and suggested that the notion of ethnicity should be placed at the forefront of theories and analyses of under-five mortality in Africa. In line with this, it seems that in the South African context, the measurement of socio-economic status by population group seems a plausible option in studies involving mortality in early life.

The higher IMR in rural areas and some of the provinces compared to others is consistent with the findings of Bachmann *et al* (1996), Gilson and McIntyre (2001) and the Demographic and Health Survey of 1998. The present findings indicate that IMR in rural areas is 20 to 80 percent (mean=45 percent) more than those in the urban areas. This is also supported by the estimated probit model indicating a 1.7 percent decrease in IMR in urban areas compared to those in the rural. There is, however, an interesting contrast with U5MR, in which case the model indicates that there is no statistically significant difference in U5MR between rural and urban areas. Since the U5MR comprises IMR and child mortality rate (1-4 years of age), it follows that the rural-urban location has no effect on child mortality. The relationship between child survival and urbanization is complex (Stanton 1994). On the one hand, urbanization may have a positive influence on child survival. This is perhaps related to higher incomes and better access to health care and other health enhancing amenities among the urban population. On the other hand, urban population growth rate is positively related with the under-five mortality rate (*ibid*). This is partly attributed to the high levels of rural-urban migration with the attendant relocation of rural poverty to urban areas, which leads to the proliferation of squatter settlements with high levels of poverty and limited amenities.

Furthermore, the aggregate poverty level of the province of residence does not seem to have a significant association with infant mortality, while there is a tremendous decline in U5MR in provinces designated as relatively well off. The U5MR decreases by 39 per 1000 (3.9 percentage points) in the relatively well off provinces compared to those designated as relatively poor. These differences also highlight the fact that infant and child mortality are influenced by different factors (Hojman 1992).

Inequalities in under-five mortality exhibit the same trend as seen in IMR. The pro-rich inequality seen in the pre-1994 period is eliminated in the latter period. As it is true for the

IMR, this is achieved through a decrease in U5MR across all income quintiles, which is more pronounced in the poorest. The U5MR decreased by an average annual rate of about 13.7 percent in the poorest quintile, as opposed to a rate of only 3.2 percent in the richest quintile. This levelling-up approach to decreasing inequalities in U5MR is a move in the right direction that needs to be maintained. This trend may partly be attributed to the government's focus on improving access through to health care through its various strategies such as the clinic building campaign, the Cuban doctor programme and free care to children under six (Ntsaluba and Pillay1998).

Even though under-five mortality depends on a host of factors, many of which may be outside the health sector, the contribution of health care to improving child survival is not to be underestimated. An increase in the proportion of children receiving health care for the commonest ailments, that account for a greater proportion of the burden of disease, is found to be negatively associated with mortality (Rustein 2000, Stansfield and Gelband 2001).

The estimated probit model has revealed a significant racial gradient in U5MR. The U5MR decreases by more than 3.5 percentage points (35 per 1000) in the Indian and White population groups as compared to the reference category, the African group. Some variables that were found statistically insignificant in the case of IMR are seen to have a significant influence on U5MR. The ownership of a house leads to a decrease in U5MR of 1.1 percentage points (11 per 1000). Although, this may be regarded as a proxy for the household's economic position, a simple correlation analysis between the two variables does not show a significant association. This may also be expected, as the ownership question does not refer to ownership of houses of a specific standard. The mechanism of how ownership affects U5MR and leads to child mortality gradients thus seems complex and needs to be investigated thoroughly.

Contrary to the findings in the IMR, the aggregate provincial level of poverty increases U5MR levels significantly. Counter-intuitively, a significant U5MR gradient favouring girls is observed.

The study contributes to the already existing literature on inequalities in infant and under-five mortality in the following ways:

- Previous studies of infant and under-five mortality have only looked at one period, and thus, there have been no analyses of trends. By comparing inequalities in infant and under-five mortality in two time periods (pre- and post-1994), that is a period of structural break in the South African political history, the study demonstrates a change of the situation. The pro-rich income-related inequalities in IMR/U5MR that existed before the change of government are seen to disappear in the period after 1994. Furthermore the use of concentration curves and indices gives a better insight into the magnitude of the change in inequalities across income groups.

- It is also revealed that the measurement of socio-economic inequalities in IMR/U5MR in South Africa needs caution in the selection of an indicator of socio-economic status. It is seen that inequalities that have not manifested themselves when income is used as a proxy for socio-economic status, are prominently seen when "race/population group" is used as the proxy. Thus, reliance only on one measure might lead to erroneous conclusions. While previous studies have focused on only one or two indicators of socio-economic status, this study has used a range of indicators.

- The probit regression analysis of some of the factors influencing IMR/U5MR has also shown that within the context of South Africa, factors that influence infant and under-five mortality are not the same. For example, while area of residence (rural/urban) has a significant effect on IMR (IMR increases tremendously in rural settings), it has no

statistically significant effect on U5MR. On the other hand, factors which do not have a significant effect on IMR (house ownership, high aggregate levels of poverty and gender, that is being female) exert a statistically significant effect on U5MR. This implies that in identifying policy instruments to manoeuvre, it is necessary to identify factors that affect mortality in the different age-groups in the under-five population even though it is for most purposes regarded as a homogeneous group.

This chapter has empirically assessed inequities in child survival. It has been shown that despite the absence of income-related inequalities in infant and under-five mortality in the 1998 OHS data, there are glaring inequalities related to population group and area of residence (rural/urban, province). These inequalities constitute inequities, as they are avoidable and unacceptable inequalities.

The picture of inequities in child life wouldn't be complete without complementing the assessment of inequities in child survival with an assessment of inequities in some sensitive indicators of child health. Systematic inequalities in mortality may possibly not be observed. However, this does not necessarily mean that children from the poorest households lead a good quality of health. Thus, to remedy this flaw, the next chapter will assess inequities in child health as measured by under-five child malnutrition.

CHAPTER 5
EQUITY IN UNDER-FIVE CHILD MALNUTRITION

5.1. INTRODUCTION

As discussed in the last chapter, although children may escape mortality, that on its own does not guarantee that they will lead a healthy life, free of disability/sickness. Mortality is an extreme occurrence and end-point, and does not reflect quality of life (Braveman 1998). Thus, it is essential to complement analysis of inequalities based on mortality data by assessing some sensitive measures of child health. It is for this reason that this chapter attempts to analyse inequities in under-five child malnutrition. At the outset, it has to be noted that the lack of recent surveys of child malnutrition that contain detailed income and consumption data makes it difficult to undertake an inter-temporal comparison or assess the status of equity in child malnutrition at the current time. Only average values of the various forms of malnutrition from recent surveys are used to indicate the current status. Thus, caution should be exercised in interpreting the findings and relating them to trends in infant and under-five mortality rates discussed in Chapter 4.

Socio-economic inequalities in health manifest in all age groups. Studies have revealed wide socio-economic differences in rates of morbidity and mortality among children (e.g. Wagstaff 2000, Brockerhoff and Hewett 2000, Gilson and McIntyre 2001). Avoidable inequalities in health during the early years of life deserve special attention, as they are likely to perpetuate inequality in the future adult population.

The nutritional status of under-five children is one of the indicators of household well-being and one of the determinants of child survival (Thomas *et al* 1990). Child malnutrition is one of the most important causes of infant and child mortality (Pelletier *et.al.* 1995, Svedberg 1987).

It is also a reflection of the macro-economic situation of a country. Malnutrition may adversely affect the child's intellectual development and consequently, health and productivity in later life (Cravioto and Arrieta 1986, WHO 1995). This is likely to perpetuate inequities and inequalities in health and other dimensions of household welfare. Child malnutrition is also one of the measures of health status that the WHO recommends for assessing equity in health (Braveman 1998).

South Africa's health policy document, *the White Paper for the Transformation of the Health System in South Africa*, states that "nutrition for all South Africans should be promoted as a basic human right ..." (South Africa 1997:85). This may be regarded as an indication of the government's concern for equity in access to nutrition-enhancing interventions, which in the final analysis, are regarded as essential for good nutritional status.

The aim of this chapter is to contribute to the efforts to quantify inequalities in health with reference to the nutritional status of under-five children in South Africa. It specifically attempts to:

i. assess and quantify the magnitude of inequalities in health that are ascribable to socio-economic status; and
ii. identify the socio-economic factors (proximate determinants) that have a bearing on health inequality and the intensity of their influence.

5.2. METHODS

5.2.1. SOURCE OF DATA

The data used in this study is derived from the project for statistics on Living Standards and Development Survey (LSDS). This survey was conducted jointly by the South African Labour and Development Research Unit (SALDRU) and the World Bank in 1993. It was based on a

sample of 8,848 households, which consisted of 40,284 individuals. The survey is designed to collect household data that can be used to assess multiple aspects of household welfare and behaviour and to evaluate the effect of various government policies on the living conditions of the population using a multi-topic questionnaire. The section on nutrition, besides questions related to child health, includes anthropometric measurements. For the purpose of this study, data on 3765 under-five children whose records were complete in the required individual and household level variables are included.

The LSDS provides the most recent data set, which includes both anthropometric measures and data on income and consumption, which can be used as indicators of socio-economic status. The more recent Demographic and Health Surveys do not provide income or consumption data, which limits their use in conducting detailed socio-economic inequality analysis in child malnutrition..

5.2.2 MEASUREMENT OF NUTRITIONAL STATUS

There are various ways of assessing the nutritional status of under-five children. It can be assessed using clinical signs, biochemical indicators or anthropometry (de Onis 2000). The anthropometric approach is the most commonly used tool (WHO Working Group 1986) and is more advantageous compared to the other two (de Onis 2000). While clinical signs and biochemical abnormalities may only be useful in advanced cases of malnutrition, the anthropometric indicators are sensitive even in incipient ones. Furthermore, anthropometric measures are less costly and easy to obtain compared to the other two techniques.

Anthropometric indicators are constructed using data on the children's age, height and weight. Three key anthropometric measures calculated from the age, height and weight data are weight-for-height, height-for-age and weight-for-age. These measures are expressed in the

form of Z-scores[1], which compare a child's weight and height with those of a similar child from a reference healthy population. The World Health Organization recommends the US National Center for Health Statistics (NCHS) population as a reference for international use (WHO Working Group 1986). This reference population, which has been in use since 1977 (WHO Working Group 1986), however, has been found to have some technical and biological drawbacks, thus driving the WHO to conduct a multi-country study geared towards developing new reference values (de Onis 2000).

Following conventional cut-off points, malnutrition in its various forms is operationally defined as follows:

i. *stunting*: height-for-age that is less than the international reference value by more than two standard deviations;

ii. *wasting*: weight-for-height less than the international reference value by more than 2 standard deviations; and

iii. *underweight*: weight-for-age that is more than two standard deviations below the international reference value.

Stunting is regarded as an indicator of long-standing dietary inadequacy. A high prevalence of stunting in the community is associated with poor socio-economic conditions (Skoufias 1998). The WHO recommends stunting as a reliable measure of overall social deprivation (WHO Working Group 1986). The height-for-age measure is less sensitive to temporary food shortages and thus, is the most reliable indicator of long-standing malnutrition in childhood (Svedberg 1987). Wasting on the other hand reflects acute malnutrition. It has the advantage

[1] Height-for-age of Z-score of child "*i*"is given as:

$$Z-score=\frac{H_i-H_r}{SD\ of\ the\ reference\ population}$$ Where H_i is the height of the child; H_r is the median height of the reference population; and SD is the standard deviation of height of the same reference population.

that it does not require an accurate knowledge of the child's age, (this is particularly important in the setting of developing countries, where it may be difficult to get the exact age of the child). Wasting is also useful in evaluating the benefits of nutrition intervention programmes, as it is sensitive to short-term changes (unlike stunting which does not respond quickly). Low weight-for-age is difficult to interpret, as it cannot discriminate between temporary and permanent malnutrition. However, in populations were the rate of wasting is low, it can be interpreted in the same way as height-for-age (Skoufias 1998). Stunting and wasting are, thus the preferred measures of child nutritional status, since they can distinguish between long-standing and short-run malnutrition (WHO Working Group 1986).

In this analysis, outliers were removed in line with the exclusion ranges recommended by WHO (WHO 1995). Hence, weight-for-height Z-scores less than -4.0 and greater than +5.0, height-for-age Z-scores less than -5.0 and greater than +3 and weight-for-age Z-scores less than –5.0 and greater than +5.0 are excluded from the analysis.

5.2.3. MEASUREMENT OF SOCIO-ECONOMIC INEQUALITIES IN MALNUTRITION

Inequality in ill-health (malnutrition) is measured using an illness-concentration index (C) as discussed in Chapter 3. In this case C is derived from the distribution of malnutrition across income groups, where income is represented by per capita household expenditure.

5.2.4. THE ECONOMETRIC MODEL

To further examine the influence of a host of socio-economic factors on the child's nutritional status, a probit model is estimated and the marginal/discrete changes in the probability of the various forms of malnutrition calculated. The three forms of child malnutrition, stunting, underweight and wasting, are taken as the dependent variables separately. Each dependent variable is dichotomised as a zero-one variable. If stunting or underweight or wasting is

present, the variable takes a value of one, and if the child's height-for-age, weight-for-age or weight-for-height Z-score is normal, it takes the value of zero.

The definition, measurement and expected sign of each of the independent variables is presented in Table 5.1.

Table 5.1
Definition and measurement of variables

VARIABLE	DEFINITION AND MEASUREMENT	BASE CATEGORY (FOR CATEGORICAL VARIABLES	EXPECTED SIGN
AGE	Age of child in months		+
GENDER	Gender = 0 if female child = 1 otherwise	Female	-
COLOURED	= 1 if the child belongs to the population group "Coloured" = 0 otherwise	African	-
WHITE	= 1 if White = 0 otherwise	African	-
URBAN	= 1 if urban area = 0 otherwise	Rural	+
METRO	= 1 if metropolitan = 0 otherwise	Rural	+
QUINTILE2	= 1 if household belongs to per capita expenditure quintile 2 = 0 otherwise	Quintile 1	-
QUINTILE3	= 1 if per capita expenditure quintile 3 = 0 otherwise	Quintile 1	-
QUINTILE4	= 1 if per capita expenditure quintile 4 = 0 otherwise	Quintile 1	-
QUINTILE5	= 1 if per capita expenditure quintile 5 =0 otherwise	Quintile 1	-
NOT-PIPED	= 1 if water supply source not piped = 0 if piped water supply	Piped water supply	+
SHACK	= 1 if type of house is shack/hut = 0 if standard, non-shack houses	Standard houses	+
NON-FLUSH	=1 if non-flush toilet = 0 otherwise	No toilet	-
FLUSH	= 1 if flush toilet = 0 otherwise	No toilet	-
CLINICAL CARD	= 1 if the child has a clinical card = 0 if the child has no clinical card	No clinical card	-
PRIMARY EDUC	= 1 if mother has primary education = 0 otherwise	No education	-
SECONDARY EDUC	= 1 if secondary education = 0 otherwise	No education	-
TERTIARY EDUC	= 1 if mother has any form of post-secondary education = 0 otherwise.	No education	-

* The Indian population group omitted because of smallness in number

The variable gender has been found to be one of the factors that influence a child's nutritional status. Sex of the child may sometimes reflect the weight that a household attaches to different children in the intra-household allocation of resources. There is substantial empirical

evidence that in developing countries, households have a gender-bias that favours boys (Skoufias 1998). A study in a sample of households in rural India indicated that changes in food prices adversely affected the nutritional status of girls (Behrman and Deolalikar 1988). A study in Bangladesh (Chen *et al* 1981) also revealed that after the perinatal period, there was a higher mortality rate among females that was presumably caused by the preferential feeding and increased access to health care for males. In South Africa, there is no evidence of preferential feeding that favours boys. However, given the evidence from other developing countries, it is hypothesized that girls have higher levels of malnutrition compared to boys. Thus, given that the gender dummy in the model assumes the value of zero when the child is a girl, the sign of the coefficient is expected to be negative.

The population group to which the child's household belongs is an important correlate of the child's health status. Various studies on under-five mortality and morbidity indicate that children from ethnic groups/races that are regarded as disadvantaged bear a greater burden of mortality and morbidity (see for example Bachmann *et al* 1996, Gilson and McIntyre 2001, Brockerhoff and Hewett 2000, Stanton 1994). In South Africa, the African and Coloured population groups are regarded as the most disadvantaged groups in terms to access to resources. For example, the 1998 *Poverty and Inequality Report* (May 1998) classifies 61 percent of the Black (i.e. African population group), 38 percent of the Coloured, 5 percent of the Indian and 1 percent of the White population groups as poor. Thus, it is hypothesized that the probability of malnutrition will significantly drop as one moves from the base category, the African population group, to the Coloured and White categories. Hence the sign of the coefficient is expected to be negative. The Indian population group is excluded because of small size of the sub-sample that precludes a meaningful analysis.

There is conflicting evidence as regards the prevalence of malnutrition in relation to area of residence (urban vs. rural). Traditionally, urbanization is regarded as having a positive influence on health outcomes (Williams 1990). A recent study in Mozambique indicates that under-five malnutrition is much higher in rural than in urban areas (Garrett and Ruel 1999). This is mainly attributed to differences in rural-urban expenditure levels (proxy for income) and maternal education levels, which were found to be higher in urban areas. On the other hand, the rapid urbanization taking place in developing countries, has resulted in relocation of poverty and undernutrition to urban areas (Haddad *et al* 1999). Furthermore, it is found that urban-dwellers depend on purchases of food from the market (Ruel *at al* 1999). Consequently, the levels of food price would seriously affect the nutritional status of children in poor households. However, in this study it is hypothesized that the probability of average under-five malnutrition decreases progressively as one moves from the base category, rural area, to urban and metropolitan locations. Therefore, the signs of the coefficients for the variables "URBAN" and "METRO" are expected to be positive.

The link between poverty and child malnutrition has been well established. Household income is particularly strongly related with long-term malnutrition (Sahn 1994). Hence it is expected that as one moves from the reference category per capita expenditure quintile 1 to quintile 5, the probability of child malnutrition will progressively decrease. Thus the expected signs for the categories quintiles two to five will be negative.

The provision of public utility services such as sewerage and water supply has been suggested to have a significant effect on a child's nutritional status (Thomas and Straus 1992). A study in an urban ghetto in Lagos, Nigeria demonstrates that most children with malnutrition came from homes with inadequate water supply and poor refuse disposal methods (Abidoye and Ihebuzor 2001). A study in protein-energy malnutrition (PEM) in urban children in Ethiopia

indicates strong associations between the non-availability of latrine and poor housing condition and PEM (Getaneh *et al* 1998). It is thus hypothesized as follows:

- The probability of child malnutrition is higher in households with no piped water supply. Thus, since the base category is households with no piped water supply (*NOT PIPED*), it is expected that the sign of the coefficient will be negative. This implies that as one moves from households with no piped water supply to those with piped water, the probability of child malnutrition declines.
- Sub-standard houses are associated with higher levels of malnutrition. Hence a negative sign is expected of the coefficient of *SHACK*. In other words, as one moves from shacks to standard houses, the probability of malnutrition is expected to decrease.
- The non-availability of toilets in the house is assumed to have a negative influence on child nutritional status. Hence, moving from the base category (households with no toilet) to houses who have toilets (of different type), it is expected to see a decline in the probability of under-five child malnutrition.

The presence of child clinical card is taken as a proxy to the child's attendance in the under-fives clinic, where growth monitoring, immunization and other activities are conducted. Given the role that infections play in triggering child malnutrition, it is hypothesized that the presence of clinical card (proxy for attendance at child health clinics) is associated with a lower probability of malnutrition. Thus, the coefficient of the variable *CLINICAL CARD* is expected to be negative.

Parental education, especially that of women, is strongly associated with behaviours that promote health, even after controlling for income (Cebu Study Team 1991). A Côte d'Ivoire study (Sahn 1994) on child malnutrition indicates that education of women is an important

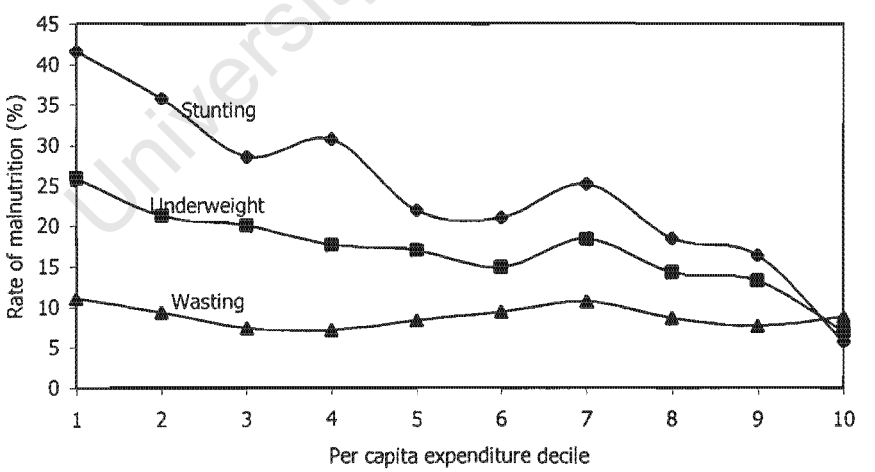
determinant of child nutritional status. The influence of education was found to be over and above the fact that better-educated women have higher earnings that tend to raise incomes and nutritional status. Thus, it is hypothesized that moving from no education to various levels of education will monotonically reduce the probability of under-five malnutrition. This implies that the coefficient of education will assume a negative sign (the base category being no education).

5.3. RESULTS

5.3.1 PREVALENCE AND SOCIO-ECONOMIC INEQUALITIES IN MALNUTRITION

Figure 5.1 indicates that the rate of stunting is the highest followed by low weight-for-age (underweight). Closer examination of the three states of child malnutrition reveals that while stunting and underweight are responsive to improvements in the socio-economic status of the household, wasting does not appear to be sensitive.

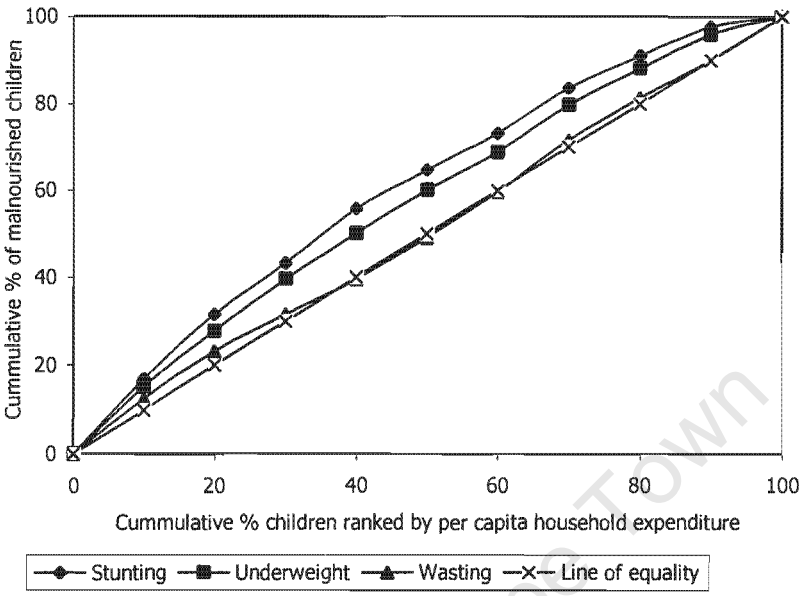
Figure 5.1
Child malnutrition by per capita expenditure decile



Children from the poorest 10 percent of households have rates of underweight and stunting, which are about three and eight times those of the richest 10 percent respectively. Furthermore, the rates of stunting and low weight-for-age are highest among the African population group. Wasting which is a manifestation of acute and short-lived malnutrition, however, does not exhibit significant socio-economic differentials. There are wide geographical variations in the rates of the three types of child malnutrition, with the rate difference between the provinces with the highest and the lowest prevalence being more than three-fold.

The overall concentration indices for stunting, underweight and wasting respectively are -0.215, -0.152 and -0.019. The figures for stunting and underweight indicate statistically significant inequalities, which are pro-rich, that is those in the lowest socio-economic strata bear a greater burden of malnutrition. However, this socio-economic gradient is not observed in wasting. This information is also presented using concentration curves in Figure 5.2.

Figure 5.2
Malnutrition concentration curves



It is noted that income-related inequalities are the strongest in stunting (the stunting concentration curve is the farthest from the line of equality), an indicator of chronic malnutrition that is often associated with socio-economic deprivation. As expected, no discernible socio-economic inequalities are observed in wasting (the wasting concentration curve almost overlaps with the line of equality), as income has little effect on the stochastic conditions (unforeseen environmental factors and diseases) which usually precipitate wasting. To gain further insight into the nature of the inequality, the concentration indices are disaggregated as shown in Table 5.2.

Table 5.2
Malnutrition concentration indices

Variable	Rate (%)	C	SE (C)	p-value	95% confidence interval	
					Lower	Higher
Stunting	24.5	-0.215	0.0160	0.000	-0.246	-0.184
Underweight	17.0	-0.152	0.0204	0.000	-0.193	-0.112
Wasting	8.9	-0.019	0.0298	0.520	-0.078	0.039
Stunting: African	26.9	-0.154	0.0165	0.000	-0.186	-0.121
Stunting: Coloured	18.8	-0.248	0.0716	0.000	-0.389	-0.107
Stunting: White	5.2	-0.284	0.1624	0.082	-0.604	0.036
Underweight: African	18.6	-0.083	0.0211	0.000	-0.125	-0.042
Underweight: coloured	12.2	-0.297	0.093	0.002	-0.480	-0.114
Underweight: White	3.5	-0.250	0.2014	0.217	-0.646	0.147
Wasting: African	9.5	0.033	0.0313	0.291	-0.028	0.094
Wasting: Coloured	5.2	-0.106	0.1508	0.483	-0.403	0.191
Wasting: White	4.8	-0.211	0.1706	0.218	-0.547	0.125
Stunting: Metro	18.0	-0.314	0.0445	0.000	-0.402	-0.227
Stunting: Urban	20.2	-0.247	0.0422	0.000	-0.330	-0.164
Stunting: Rural	28.0	-0.144	0.0189	0.000	-0.181	-0.107
Underweight: Metro	13.9	-0.281	0.0525	0.000	-0.385	-0.178
Underweight: Urban	16.4	-0.136	0.0489	0.005	-0.232	-0.040
Underweight: Rural	18.2	-0.109	0.0252	0.000	-0.158	-0.059
Wasting: metro	9.4	-0.099	0.6679	0.137	-0.2306	0.0316
Wasting: urban	9.5	-0.035	0.0670	0.604	-0.166	0.097
Wasting: rural	8.5	-0.031	0.0391	0.427	-0.108	0.046

Table 5.2 indicates that income-related inequalities in stunting and underweight increase monotonically with the increase in the degree of urbanisation of the household's area of residence. In other words, a dose-response pattern of relationship is observed. In all three areas of residence – rural, urban and metropolitan – the poorest bear the heaviest burden of stunting and underweight. However, inequalities in both stunting and underweight are lowest in rural settings and highest in metropolitan areas. It is further observed that although the rates of stunting and underweight are highest among African children, the pro-rich concentration indices are more pronounced for Coloured children (i.e. there are greater

disparities in nutritional status across income groups within the Coloured population). Stunting and underweight concentration indices for White children do not manifest statistically significant socio-economic inequalities. For all population groups and places of residence, wasting does not show any gradients linked to socio-economic status.

The findings also indicate that stunting has a remarkably high concentration among the poorest in all provinces (data for Northern Cape are not included due to the small numbers in this sub-sample). The pro-rich concentration indices for stunting have very high statistical significance in all provinces. Similarly, underweight concentration indices show statistically significant pro-rich inequalities in all but three provinces. In the Free State, Kwazulu-Natal and Mpumalanga provinces, underweight concentration indices do not exhibit income-related inequalities. Wasting does not show statistically significant pro-rich inequalities in all provinces except the Northern Province. Table 5.3A-C depicts this information.

Table 5.3A
Stunting concentration indices by province

Province	Rate (%)	C	SE (C)	p-value	95% confidence interval	
					Lower	Higher
Eastern Cape	31.5	-0.123	0.0318	0.000	-0.186	-0.061
Free State	25.2	-0.239	0.069	0.000	-0.415	-0.144
Gauteng	18.2	-0.270	0.0522	0.000	-0.371	-0.168
Kwazulu-Natal	24.6	-0.108	0.0357	0.003	-0.178	-0.038
Mpumalanga	19.5	-0.229	0.065	0.000	-0.357	-0.102
North West	23.4	-0.260	0.063	0.000	-0.384	-0.136
Northern Province	27.2	-0.255	0.037	0.000	-0.327	-0.182
Western Cape	16.5	-0.273	0.0.085	0.002	-0.440	-0.105

Table 5.3B
Underweight concentration indices by province

Province	Rate (%)	C	SE (C)	p-value	95% confidence interval	
					Lower	Higher
Eastern Cape	9.4	-0.199	0.0671	0.003	-0.331	-0.067
Free State	19.9	-0.146	0.083	0.078	-0.309	-0.016
Gauteng	15.4	-0.201	0.0585	0.001	-0.316	-0.086
Kwazulu-Natal	15.9	-0.054	0.0471	0.249	-0.147	0.038
Mpumalanga	15.2	-0.099	0.0764	0.194	-0.250	0.051
North West	27.5	-0.277	0.0557	0.000	-0.386	-0.167
Northern Province	25.8	-0.175	0.0393	0.000	-0.252	-0.098
Western Cape	10.9	-0.274	0.1093	0.013	-0.489	-0.058

Table 5.3C
Wasting concentration indices by province

Province	Rate (%)	C	SE (C)	p-value	95% confidence interval	
					Lower	Higher
Western Cape	4.8	0.123	0.1724	0.476	-0.217	0.0463
Eastern Cape	3.1	-0.093	0.1229	0.450	-0.3343	0.1484
Kwazulu-Natal	8.7	-0.067	0.0664	0.313	-0.063	0.198
Free State	11.5	-0.124	0.115	0.281	-0.350	0.102
Mpumalanga	6.7	-0.226	0.174	0.194	-0.567	0.116
Northern Province	13.1	-0.122	0.060	0.000	-0.240	-0.004
North West	14.7	-0.160	0.086	0.064	-0.329	0.009
Gauteng	10.8	-0.094	0.072	0.193	-0.237	0.048

Table 5.3 also indicates the tendency for higher concentration indices to be found in those provinces that have relatively lower rates of stunting (see for example Western Cape and Gauteng). Thus, even though a relatively low proportion of children under-five in the Western Cape and Gauteng experience stunting, these provinces have the greatest disparities in malnutrition between income groups. In contrast, provinces such as Kwazulu-Natal and the Eastern Cape that have relatively high rates of stunting have smaller differences in malnutrition across income groups.

5.3.2 SOCIO-ECONOMIC DETERMINANTS OF MALNUTRITION

The results of the probit regression indicate the changes in the probability of each of the three forms of malnutrition (dF/dx) for a discrete change of each of the variables from zero to one, when the explanatory variables are categorical variables. In this case, all other variables are held constant at their modal value. With continuous variables (only AGE in this case), the results show the probability of malnutrition for each month increase in the child's age, when all other variables are held constant at their mean values.

It should be noted that the results of the probit estimation in this analysis are not presented as parameter estimates, as is commonly done in OLS regression. This is for the simple reason that presenting parameters and their statistical significance ignores the link function in generalized linear models. Given a statistically significant parameter, a positive sign suggests the likelihood of malnutrition increases with the level or presence of an independent variable, all others held constant. Conversely a negative sign of the estimate indicates the likelihood of malnutrition decreases with the level or presence of an independent variable. However, such interpretation is vague and non-informative, as we do not know by how much the dependent variable increases or decreases the outcome variable (malnutrition). Thus, to avoid this problem, the marginal effects on the probability of malnutrition are presented as is shown in Table 5.4.

Table 5.4
Probit estimation results

Independent variable	STUNTING		UNDERWEIGHT		WASTING	
	dF/dX	P-value	dF/dX	P-value	dF/dX	P-value
Male	0.0526	0.002	-0.0019	0.896	0.0110	0.331
AGE	0.0014	0.007	0.0022	0.000	0.0008	0.022
COLOURED	0.0297	0.441	-0.0310	0.314	-0.0445	0.037
WHITE	-0.0945	0.032	-0.1059	0.002	-0.0470	0.038
URBAN	-0.0093	0.768	0.0135	0.627	0.0251	0.250
METRO	0.0222	0.576	0.0543	0.128	0.0554	0.046
QUINTILE2	-0.0340	0.217	-0.0278	0.195	-0.0004	0.983
QUINTILE3	-0.0973	0.000	-0.0572	0.007	-0.0120	0.507
QUINTILE4	-0.0923	0.000	-0.0747	0.001	-0.0115	0.542
QUINTILE5	-0.1636	0.000	-0.1181	0.000	-0.0380	0.081
NOT-PIPED	0.0037	0.863	-0.0092	0.620	-0.0038	0.798
SHACK	-0.0381	0.058	-0.0132	0.447	-0.0015	0.913
NON-FLUSH	0.0096	0.677	0.0448	0.029	0.0351	0.042
FLUSH	-0.0294	0.481	0.0126	0.733	0.0172	0.553
CLINICAL CARD	-0.0271	0.169	0.0131	0.431	0.0178	0.169
PRIMARY EDUC	-0.0449	0.029	-0.0272	0.127	-0.0306	0.029
SECONDARY EDUC	-0.0649	0.005	-0.0355	0.077	-0.0092	0.554
TERTIARY EDUC	-0.1104	0.033	-0.0421	0.360	0.0092	0.784
$LR\ \chi^2_{18}$	183.94		113.02		35.88	
P-value	0.0000		0.0000		0.0318	

The likelihood ratio χ^2 statistic evaluates the goodness of fit of probit models. It tests the null hypothesis that all the regression coefficients in the model except the intercept are equal to zero. It can be seen that in all three models the null hypothesis is rejected, and thus, we can conclude that the variance explained is greater than zero, and that at least one regression coefficient is significant.

The results indicate that the probability of stunting increases significantly in boys as compared to girls. This, however, is not observed in the case of underweight and wasting. Age of the child has a significant negative effect on the child's nutritional status. There is a significant increase in the probabilities of all three forms of malnutrition with increases in the age of the child. Rates of the three forms of malnutrition computed after dichotomising the *AGE* variable into those below the age of one year who predominantly depend on breast milk, and those one year and above (the total population consists of under-fives) indicate rates which are more than twofold in the older age group.

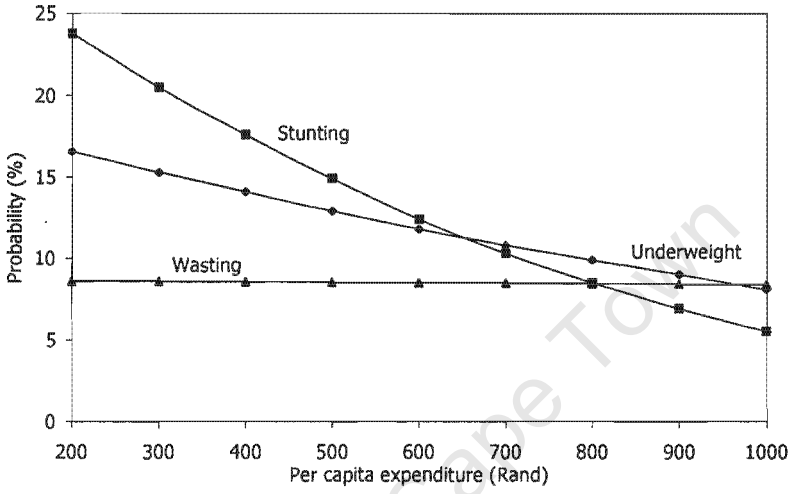
With respect to population group, no significant difference in probability is observed between African and Coloured children in stunting and underweight – groups which in the South African context are regarded as historically disadvantaged. However, a divergence is observed in the case of wasting, where probability decreases by about 4.5 percentage points ($p<0.05$) in Coloured children compared to Africans. Compared to the base category, the African child, all three forms of malnutrition manifest significant and substantial decline in probability in the White child. This indicates the existence of a substantial difference in the levels of child malnutrition between those regarded as historically advantaged and those that are disadvantaged.

The measure of socio-economic status, which in this case is represented by the household's per capita expenditure quintile category, has the expected sign and statistical significance in stunting and underweight. However, in wasting, although with the expected sign, it is of no statistical significance.

As one moves from the base category, the poorest quintile, to the richest the rate of stunting drops by 18.2 percentage points. The corresponding figure for low weight-for-age is 9.0

percentage points. To illustrate the effects of household income as measured by household per capita expenditure on the probabilities of malnutrition, a plot of the probabilities predicted by the models is presented in Figure 5.3.

Figure 5. 3
Predicted probabilities for malnutrition

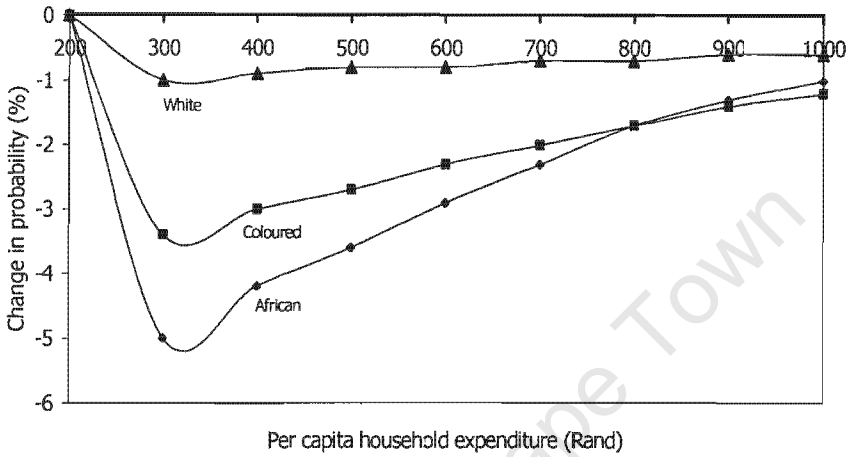


From figure 5.3 we can discern a number of important issues on the relationship between income and the probabilities of a child suffering from malnutrition. First, it is observed that of the three types of malnutrition, stunting is the most income-responsive. There is a dramatic reduction in the rate of stunting with increases in the household's per capita expenditure. More importantly, it can be seen that in households at the lower end of the SES scale, the drop in the prevalence of stunting is very remarkable for each Rand increment in per capita household expenditure, as compared to the decreases in the rate of stunting on the other end.

Second, low weight-for-age is also responsive to increases in household income, but the decrease in probability is much smaller than that seen in stunting. As seen in stunting, there is a waning of the income effect at relatively higher levels of household per capita expenditure, signifying that the marginal change in the probability of being underweight gets smaller with

the further increases in income. On the other hand, wasting does not appear to be responsive to changes in household income levels. The income-response, however, is not uniform across the various population groups. This is depicted in Figure 5.4 with reference to stunting.

Figure 5.4
Changes in predicted probabilities for stunting



At lower levels of SES, it is observed that stunting is highly responsive to increases in household income in the African and Coloured population groups. In the Whites, the same trend is seen, *albeit* at a smaller rate.

The probability of wasting increases by more than 5 percentage points in children of households residing in metropolitan areas compared to those in the rural. Thus controlling for other factors, the rural child seems to be in a better position than his/her counterparts in the metropolis, as far as wasting is concerned. However, no significant change is observed in stunting and underweight among the various residential locations, despite high average prevalence rates in rural areas. This absence in changes of probabilities may be related to the fact that there is a high intra-urban/metropolitan differential as manifested by relatively high pro-rich inequalities.

Having piped water supply or not and living in shacks or otherwise do not seem to significantly influence the three forms of malnutrition. On the other hand, in a household using a non-flush toilet, the probability for a child being wasted and underweight increases by more than three and four percentage points respectively ($p<0.05$) compared to those with no toilet. This finding is contrary to our *a priori* expectation.

The variable *CLINICAL CARD,* with all its limitations, is included as a proxy for access to child health services, which have a positive contribution to a child's nutritional status. Contrary to expectation, the models indicate that there are no statistically significant differences in the probabilities of stunting, underweight or wasting between children who do not have the clinical card and those who have it. The variable lacks significance even at the 10 percent level of significance.

Consistent with expectation, the mother's educational status has a negative effect on the levels of stunting. The probability of stunting decreases substantially and monotonically as the level of education increases. For example, compared to the child whose mother is not educated, the probability of stunting of a child from a mother with post-secondary education decreases by more than 11 percentage points (all other variables held constant at their means and modes).

5.4. DISCUSSION

This chapter attempts to examine the socio-economic inequalities in health with special reference to under-five child malnutrition in an effort to quantify the inequalities and identify some of the factors influencing the various forms of child malnutrition: stunting, underweight and wasting.

It is observed that the three states of protein-energy malnutrition (PEM) in children are a problem in South Africa, as much as they are in other countries in the developing world. The overall rate of stunting found in this study for both male and female children is similar to those reported for South Africa in other reports (WHO 1999, HST 1999). The WHO Global Data Base on Child Growth and Monitoring indicates that a survey conducted in 1999 on a sample of 1556 under-five South African children reveals rates of stunting (24.9 percent), underweight (11.5 percent) and wasting (6.2 percent) in children between the ages of 1-4 years, which are not very different from the current findings. This conforms to the trend seen in many African countries, where the rate of decline in stunting has been slow (de Onis *et al* 2000). As in many other countries, stunting is the greatest problem, followed by underweight and wasting (see for example Table 2 in de Onis *et al* 1993).

The aggregate prevalence rates conceal vital information, and hence do not give a realistic picture of the prevailing situation of PEM. Thus, disaggregating by various indicators of socio-economic status is vital in generating valuable information for policy decisions. The rate differentials in all forms of under-five child malnutrition are highly pronounced when decomposed by various indicators: socio-economic, demographic, and geographic.

The household's economic position is seen to have a highly significant impact on the probability of a child being stunted and underweight. The inverse relationship that is observed between stunting and household's socio-economic status has been well established in the literature. For this reason, the World Health Organization recommends stunting as a measure of overall social deprivation and as one of the indicators to monitor equity in health (WHO Working Group 1986, Braveman 1998).

As one moves up the income ladder, a remarkable drop in the rate of stunting is observed. Improved household income levels are associated with a dramatic drop in the probability of stunting of children. In the poorest households, an increment of Rand 100 in per capita household expenditure results in a fall in predicted probability of stunting of more than 10 percent in African and Coloured children. Several studies have indicated that increasing the income of the poorest is a sound strategy to curb the high rates of stunting in the socio-economically deprived segments of the population (e.g. Sahn 1994, World Bank 1981). In the same vein, recent studies on the effects of South African old age pension on the health of households indicate positive outcomes including child nutrition (Case and Wilson 2001, Duflo 2000).

Systematic inequalities in long-standing under-five malnutrition have far-reaching consequences. Studies have indicated that malnutrition contributes to a significant reduction in lifetime earnings (Behrman and Hoddinot 2000). Consequently, this is likely to perpetuate the already high levels of income inequality in the country. Hence, to address inequities in stunting and underweight, which are likely to continue the cycle of inequalities in income in the future, the implementation of income-generating projects and direct transfers of income to the poor are indispensable measures that must be pursued aggressively. Malnutrition, especially stunting, has a substantial socio-economic dimension, and therefore, should be viewed in a broader context and not merely in a narrow biomedical sense.

While government efforts such as the Primary School Nutrition Programme (PSNP) targeted at school children can offer palliative measures to mitigate the problems associated with school child malnutrition (such as school drop out and ultimately contribution to the improvement of academic performance), they can not have a profound and sustained impact in addressing the deep-rooted proximate determinants of malnutrition. Furthermore, long-standing malnutrition,

especially during the pre-school age is likely to result in irreversible damage to the child's intellectual development. Hence, focus on this age group is essential, as it has a substantial pay-off in the future.

A similar pattern is also observed with respect to population group. The groups who are regarded as historically most disadvantaged (Africans and Coloureds) have rates of stunting, which are in stark contrast with those of Whites. The rates of stunting and underweight in African children are more than five times those of White children. This may partly be explained by the wide income inequalities prevalent among the various population groups. Variations between areas of residence, with considerably higher rates of stunting and underweight in rural than in metropolitan areas, are also partially attributable to rural-urban income differentials.

While these disparities (especially between population groups and geographic areas) are not unexpected within the South African context, the present analysis adds valuable additional insights of policy relevance by a disaggregated concentration index analysis.

The overall concentration indices for stunting and underweight are substantial and statistically significant. They indicate that under-five children from the poorest households bear the greatest burden of malnutrition. Even though this finding is important in its own right, its policy relevance will be further increased if concentration indices are computed for various categories.

When seen by population group, the pro-rich concentration indices for stunting and underweight are substantial and statistically significant in Coloured children. While the concentration indices for White children do not have statistical significance, those of African

children are by far less than those of Coloured children. An important policy implication emanates from this finding. If policies aimed at reducing child malnutrition are based only on average rates, they will target mainly the African children, as they have rates of underweight and stunting that are profoundly high compared to the other two groups. However, this will lead to errors of omission in targeting, as the relatively higher concentration indices in Coloured children imply that the lot of those coming from the poorest households is no better than the Africans from the poorest households. Thus, it is imperative for policy makers to take account of not only inter-group differences, but also intra-group rate differentials. The present finding suggests the need for focusing not only on African children, but also the poorest of Coloured children.

The pro-rich concentration indices are the highest in the metropolitan compared to rural areas – more than two times. This underscores, the need for targeting pockets of abject poverty in the metropolitan and other urban areas. This finding is in keeping with those of others, who have demonstrated the existence of substantial concentrations of ill-health among the urban poor. The fact that urban populations experience more variations in nutritional status, poverty, morbidity and mortality compared to rural populations has been shown by a number of studies (Basta 1977, Bradley *et al* 1992). Basta (1977) has argued that using global averages to characterize poverty and child malnutrition in urban areas may be misleading, because city averages do not capture the large heterogeneity found between social classes. Hadad *et al* (1999) using the Demographic and Health Survey (DHS) data for a dozen developing countries in Africa, Asia and Latin America also demonstrate that the ratio of stunting prevalence between poorer and wealthiest quintiles is greater within urban than within rural areas. In the same line, Menon *et al* (2000) used DHS data for 11 countries in Africa, Asia and Latin America to test the hypothesis that socio-economic differentials in stunting are consistently larger in urban than in rural areas. Their findings unequivocally supported this hypothesis.

Thus dependence on global averages can be particularly misleading in countries like South Africa with high levels of socio-economic inequalities.

The above argument applies equally when the rates and concentration indices of malnutrition are decomposed by province. There is a considerable socio-economic inequality in stunting that favours the rich in all provinces. The rate of stunting in the province with the highest figure is almost twice that of the province with the lowest stunting rate. However, the concentration index of the province with the lowest rate is more than twice that of the province with the highest rate of stunting. Provinces such as the Western Cape and Gauteng with relatively lower rates of stunting have the highest concentration indices. On the other hand, provinces such as the Eastern Cape and Kwazulu-Natal that have relatively higher rates of malnutrition have the lowest concentration indices. Thus, any programme or intervention that aims at reducing under-five child malnutrition has to also consider provinces with high intra-province concentration indices in order to avoid undercoverage of those who deserve it. Hence, in addition to its focus on those provinces with relatively higher rates of stunting, government should also give due attention to those with lower rates of stunting but relatively higher intra-province variation.

Of the influencing variables, educational status of the mother has also emerged very strongly in reducing rates of stunting. The rate of stunting drops monotonically with the increase in the mother's level of education. This negative influence of maternal education on stunting is well established in the literature (see for example, Phimmasone *et al* 1996, Garrett and Ruel 1999, Cebu Study Team 1991). Not unexpectedly, the distribution of the education variable among the various expenditure quintiles is uneven. It is concentrated among the higher income groups. Thus, given the strong role that it plays in preventing child stunting and its far-reaching consequences, government policies aimed at reducing long-standing malnutrition

need to put efforts to improve the educational status of the poorest women. The role of education is not prominent in underweight. There are no rate differences in underweight between children from non-educated mothers and those from mothers having various levels of education. However, in the case of wasting, while the probability of wasting of a child from a mother with primary education drops by a little more than 3 percentage points over a child whose mother has no education ($p<0.05$), the protective effect of education wanes as the level of education increases beyond the primary level. This may perhaps reflect the fact that there is a higher probability for a mother beyond primary education to work, and thus, leave the child for a caretaker, who may not always be a complete substitute for the mother in caring for the child.

Similarly, use of a non-flush toilet (compared to "no toilet") by the household increases the probability of wasting and underweight by about 4 percent, while it has no effect on the probabilities of stunting. A study in Ethiopia shows an increase in childhood morbidity when the household uses an open pit latrine (Ali *et al* 2001). This brings to our attention the fact that toilets that are not maintained hygienically (such as the pit latrines) are likely to be risk factors for precipitating acute malnutrition rather than being protective. This is a point worthy of note in formulating plans that enhance sanitary conditions of the community as part of the efforts to control acute malnutrition in under-five children.

This analysis, contrary to *a priori* expectation, has revealed that the probability of chronic malnutrition is higher in boys than in girls ($p=0.002$). Studies conducted in Asia indicate gender bias in malnutrition that favours boys (Choudhury *et al* 2000, Pal 1999). However, in line with the findings of the current study, a number of research projects conducted in Africa and elsewhere indicate that the various forms of malnutrition are higher in boys (Skoufias

1998, El-Sayed *et al* 2001, Ngare and Muttunga 1999). Thus, in our case there seems to be gender-based inequality in stunting that is to the disadvantage of male children.

Malnutrition affects child survival negatively (Thomas *et al.* 1990, Pelletier *et al.* 1995). It may also adversely affect health status and productivity in a later adult life (Thomas *et al* 1990). Thus, the repercussions of socio-economic inequalities in child nutritional status are likely to be self-perpetuating. The lack of well-timed and targeted action against socio-economic inequalities in child nutritional status may have a neutralising effect on policies that are intended to rectify socio-economic injustices inherited from the previous political system. Hence, in South Africa, there is a pressing need for policies to remedy this situation as part of the overall efforts to redress past inequities.

The study has demonstrated that stunting and, to a certain degree, underweight are amenable to interventions aimed at raising the incomes of the poorest. Thus, augmenting the incomes of the poorest through direct transfers and other income generating mechanisms is necessary. It is also observed that the income effect on the two types of under-five malnutrition wanes after a certain level, implying that it is subject to diminishing returns.

In South Africa, although malnutrition between race groups has been highlighted previously, by undertaking a more extensive analysis, the current study demonstrates that reliance on average values would be misleading. For example, while there is a marked difference in the average rates of stunting between African and Coloured children, the pro-rich concentration index in the latter group is extremely high suggesting that there are many Coloured children whose condition is no better than that of African children. Therefore, the study emphasizes that nutrition policies and interventions in South Africa need to do an analysis of rate

differentials in stunting and underweight among the various income quintiles within a population group so as to minimize errors of targeting.

The study also reveals that provinces with the lowest levels of malnutrition (namely, Gauteng and Western Cape) have the highest pro-rich concentration indices (much higher than those of provinces with relatively lower rates of malnutrition). Hence, in targeting resources to address the problem of chronic under-five child malnutrition, it is imperative that the degree of socio-economic inequality in malnutrition within a province be taken in conjunction with the average rate.

In summary, this chapter highlights the existence of significant differences in child malnutrition (stunting and underweight) that are unnecessary, avoidable and unjust. Inequalities in the prevalence of stunting and underweight are observed among the various income quintiles, population groups and areas of residence. These inequalities are regarded as inequities, because they are associated with factors that are potentially avoidable.

As it has been discussed in the previous chapters, inequities in health manifest themselves at all stages in the life-course. Hence to have a comprehensive view of inequities, Chapter 6 will dwell on inequities in adult illness and utilization of services.

CHAPTER 6

EQUITY IN SELF-REPORTED ADULT ILLNESS AND HEALTH SERVICE UTILIZATION

6.1. INTRODUCTION

Socio-economic inequalities in health manifest themselves at all stages in the life span of individuals. Therefore, an analysis of equity in health has to examine all stages in the life course, as the nature and magnitude of inequalities may show variations. Many studies of socio-economic inequalities in adult health conducted in the developed world have consistently shown a high concentration of ill-health and death among those in the lowest socio-economic strata (Alberts *et al* 1997). However, in low- and middle-income countries, the evidence on socio-economic inequalities in adult health is scanty (Wagstaff 2000). This is specially so in sub-Saharan Africa, where studies conducted in this area are very few.

This chapter has dual objectives. First, it attempts to quantify inequalities in self-reported adult illness, which are accounted for by differences in socio-economic status and second, it assesses inequalities in care seeking and type of provider used, conditional on having reported sickness.

6.2. METHODS

6.2.1. SOURCE OF DATA

For this analysis, data are derived from the Living Standards and Development Survey of 1993 (LSDS 1993) and the October Household Surveys of 1995 and 1998 (OHS 1995 and OHS 1998). Although there are differences in some of the health-related variables used in each study, they are not widely divergent to exclude comparability. Furthermore, the time lapse between the data sets makes evaluation of policies plausible, as a considerable time interval is needed before the effects of policies start to have perceptible effects.

6.2.2. THE MEASUREMENT OF ILL-HEALTH

Although many studies of socio-economic inequalities in adult health depend on mortality data, those that use measures of morbidity mainly depend on self-reported illness. This analysis also depends on self-reported illness. Therefore, before embarking on the analysis, it would be worthwhile to discuss the strengths and weaknesses of self-reported illness as a measure of morbidity.

Measures of self-reported illness are subjective and depend upon the recall period used (2-4 week recall span) and awareness of the respondents (Newbold *et al* 1995, Wagstaff 2000), and are highly influenced by transitory factors (Wagstaff 2000). The subjects tend to respond not only to the underlying health condition, but also to perceptions, expectations, behavioural response to perceived problems and to propensity to report the perceived problems (Jack, 1999).

It is assumed that due to a greater concentration of illness among the poorest, what is self-reported as illness by the poorest is mostly more severe than what is reported by the economically better-off – a condition which is likely to distort or underestimate the extent of health inequalities (Blane *et al* 1994). Individuals who have lived all their lives with frequent bouts of severe malaria or intestinal parasites may not report a mild episode of malaria or the presence of visible helminths as an illness, thus resulting in underestimation of the volume of illness. This may perhaps explain why a higher rate of self-reported morbidity is observed in the United States than in India (Murray *et al* 1992). As Sen (2001) has aptly put it, a person living in a disease endemic area and having no knowledge of other places and experiences may consider the suffering to be part of life rather than viewing it as avoidable through preventive and curative interventions.

Due to this potential problem of self-reported measures of morbidity, some researchers prefer using measures of mortality. This implies that caution needs to be exercised when using self-reported illness in a cross-sectional analysis, especially when the subjects are of diverse socio-economic and cultural backgrounds.

The subjective nature of self-reported measures of health should not, however, be overemphasized to discount their use. Some studies have found a high correlation between the perceived assessment of one's own health and objectively verifiable health problems and survival (Idler and Angel 1990). On the other hand, objective measures of health also have their own limitations. A study in Ghana (Belcher *et al* 1976) found that only 1 in 15 of those who reported lower-back pain were confirmed objectively. This points to the deficiency of the objective measures in the sense that they cannot capture the multidimensional nature of ill-health. From the sociological perspective, it is argued that self-reported illness represents an individual's well being more than an objective, medically confirmed disease within the context of developed countries (Lahelma *et al* 1994).

Self-perceived illness also has major effects on the use of health services, and thus important implications for policy (Lairson *et al* 1995). If people perceive themselves as ill (even if objective tests may indicate to the contrary), then, *ceteris paribus*, there is a greater likelihood for them to seek care. On the other hand, even in the presence of objectively verifiable illness, people may not seek care unless they perceive it as illness. In a market-oriented system where consumer sovereignty is upheld, it is patient's rather than health professional's opinion that counts more in decisions on whether or not to consume and pay for health services (Gwatkin 2001). Thus, the role of self-reported illness/health status in health services planning is highly important.

By and large, both the subjective and objective measures have their own strengths and flaws. With respect to the subjective measures, the differential reporting of illness experience is likely to confound the presence of actual inequalities, where in some instances, it may obscure part of the objective inequalities while in others it may inflate observed health inequalities (Humphries and van Doorslaer 2000).

Some studies have found counter-intuitive results in their studies of socio-economic inequalities in self-reported illness. Baker and van der Gaag (1993) in their study of equity in a sample of developing countries found that the rich were more likely to report themselves as sick compared to the poor. A South African study (Gilson and McIntyre 2001) also indicates the same trend with respect to overall reported illness, but not for a specific condition – diarrhoea. However, both of these studies did not standardize the reported illness for age and sex and thus, did not take account of the confounding effect of these demographic variables.

The LSDS and the OHS series data depend on self-reported illness using either a two- or a four-week recall span. In both data sets, questions are asked on acute illnesses and utilisation of various health care services conditional on having reported an illness. In addition, the OHS data include self-reported injury (short-recall period) and chronic illness or disability (year recall period). Thus, the health status measure, particularly that for acute illness and injury, provides only a snapshot at one point in time. It has to be noted that estimations and assessments of health inequalities are sensitive to the measure of health used (Anand *et al* 2001). Inequities in health that are not prominent when using a particular measure of health status may be more pronounced when using another measure. Thus, recommendations emanating from the different measures of health status may sometimes lead to conflicting results and consequently incorrect policy recommendations. Further research into alternative measures of health status that allow for a broader view over a longer time perspective,

including self-assessed health status over a longer specified period (rather than reported illness using short recall periods), would be valuable. Summary measures of health that incorporate both morbidity and mortality such as Quality-Adjusted Life Years (QALYs) are being increasingly used to address the above-mentioned problem. However, since QALYs have not been widely used in developing countries, more research is needed to determine the validity and robustness of equity assessments based on them.

6.2.3. THE MEASUREMENT OF INEQUITIES

As in the preceding topics on inequalities in health in children, the concentration index is used to measure income-related inequalities in self-reported illness (acute, chronic/disability and injury) as well as care seeking and type and place of care sought.

In the equity analyses of the under-fives, a fairly similar risk of illness was assumed, as the age of the children spans a narrow interval of five years. In the adult population, which in this case is operationally defined to include those who are 18 years and above, however, this assumption may not be tenable as the age range is wide. Thus it is necessary that the concentration indices calculated take account of the confounding effect of age and gender. This is done by means of the techniques of direct and indirect standardisation.

In this study the indirect method of standardisation is used, as the study draws from individual-level data. This is done by means of running a binary logit model, where the dependent variable, self-reported illness – acute, chronic (disability) and injury – is regressed against age and gender and retaining the predicted values from the regression. The resultant standardisation implies that a person's degree of illness is replaced by the degree of illness suffered on average by persons of the same age and gender (Kakwani *et al* 1997). The indirect standardisation technique can be used on individual-level as well as grouped data.

This is in contrast to the direct technique, which requires grouped data. The requirement for grouped data is a disadvantage, as the number of socio-economic groups used is likely to affect the numerical value of the concentration index. The difference between the unstandardised and standardized concentration indices measures the extent of avoidable inequalities in self-reported illness, that is:

$$I^* = C_M - C_N \tag{6.1}$$

where,

I^* = a measure of avoidable inequality;

C_M = unstandardised concentration index; and

C_N = standardised concentration index.

I^* takes negative values if there are avoidable inequalities favouring the rich, and positive values if there are pro-poor avoidable inequalities (Kakwani *et al* 1997).

Following van Doorslaer *et al* (2000), in measuring inequity in utilisation of health care, utilisation concentration indices are computed. The focus in this case is on horizontal equity, where people in equal need of care are to be treated the same, regardless of their income status. The method compares the concentration curve for the actual health care utilisation with the concentration curve expected on the basis of need. The expected utilisation is derived through the technique of indirect standardisation. This is done by running a logit regression of the dependent variable – health care utilisation – on need indicator variables. In the case of the LSDS data, the indicators of need considered are age, gender and reported sickness. On the other hand, since the OHS series data have more categories of self-reported illness, utilisation is regressed on the following need indicators: age, gender, self-reported acute

sickness (less than four weeks), disability (chronic) (of more than six months duration) and injury.

The predicted utilisation from the above regressions gives the need-expected utilisation of health care, that is, utilisation after taking into account the need for health care. The Wagstaff-van Doorslaer index of horizontal inequity (HI_{WV}) (Wagstaff and van Doorslaer 2000) is then computed as the difference between the unstandardised utilisation (i.e. unadjusted for need) and the need-expected utilisation, i.e.:

$$HI_{WV} = C_M - C_N \tag{6.2}$$

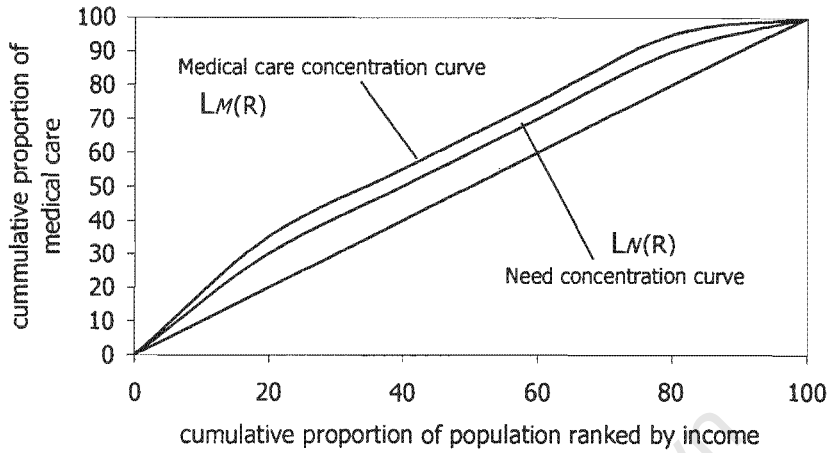
where,

C_M = medical care concentration index unadjusted for need; and

C_N = concentration index for need, that is indirectly standardized medical care

A positive value of HI_{WV} indicates horizontal inequity favouring the rich, whereas, a negative value implies horizontal inequity favouring the worse off. This is shown diagrammatically in Figure 6.1.

Figure 6.1
Concentration curves for actual and expected utilization



In the above figure, the medical care concentration curve $L_M(R)$ depicts the distribution of medical care by income. $L_N(R)$ represents need-adjusted medical care concentration curve. The value of HI_{WV} depends on the size of the area between $L_M(R)$ and $L_N(R)$.

6.3. RESULTS

6.3.1. INEQUITIES IN SELF-REPORTED ILLNESS

The 1993 LSDS data indicate pro-poor self-reported acute adult illness, when the confounding effect of age and sex is not taken into account. This implies that when using unstandardised reported illness, the rich report themselves sick more frequently than do the poor. By taking the confounding effect of the two demographic variables, it is apparent that the poor would be expected to report more sickness, given their age and sex composition. The difference between the unstandardised and standardised concentration indices, denoted as I^* indicates the presence of statistically significant avoidable inequalities in self-reported illness that are to the advantage of the poor. This implies that there are avoidable excess self-reported illnesses in the rich.

On the other hand, both the OHS-95 and OHS-98 data have values of I^* which are negative, thus signifying the presence of avoidable inequalities that are in favour of the rich.

Furthermore, the LSDS data indicate that in addition to the existence of pro-rich inequalities in self-reported illness, the poor report more days of sickness than the rich and spend more days out of work. This is a very important point that has to be considered seriously, as the opportunity cost in terms of forgone earnings is very high to the poor. It clearly indicates the impoverishing effect of illness among the poor population. The OHS data, however, do not include a question on the number of days sick or out of work.

Another dimension of health assessed in OHS 95 and OHS 98 is self-reported injury. While the OHS-95 data do not show any statistically significant difference in avoidable self-reported injury between the richest and the worse off, the OHS-98 data reveal significant avoidable inequalities that are pro-rich. (I^* is negative and statistically significant). This implies that there is a higher burden of self-reported injury among the poorest of society, and that this inequality has increased substantially since 1995.

Measures of chronic illness/disability give a better understanding of inequalities in adult illness both in the developed and developing countries (Wagstaff 2001). According to OHS-98, disability is defined as any long-term (lasting six months or more) physical or mental condition that limits the person in his/her daily activities at home, at work or at school (Statistics South Africa 1998). The categories of disability included in the questionnaire are detailed. In contrast, in OHS-95 no definition of disability is given as far as duration is concerned, and the types of disability are lumped into a few categories, thus making comparison between the two time periods difficult. In both time periods there are avoidable inequalities in disability that favour the rich. The concentration indices for the reported illness are given in Table 6.1.

Table 6.1
Illness concentration indices

LSDS 1993			
Variable	C[1] (t-ratio)	C*[2] (t-ratio)	I*[3] (t-ratio)
Self-reported adult illness	0.0248 (2.394)	-0.005 (-2.934)	0.0240 (2.3)
Days sick	-0.0201 (-5.655)	-	-
Days out of work	-0.0341 (-6.004)	-	-
OHS 1995			
Self-reported adult illness	-0.0134 (-2.463)	-0.0011 (-1.227)	-0.0123 (-2.288)
Self-reported adult injury	-0.0051 (-0.330)	0.0060 (15.423)	-0.0111 (-0.772)
Self-reported disability	-0.5326 (-60.086)	-0.0142 (-58.054)	-0.5184 (-60.301)
OHS 1998			
Variable	C (t-ratio)	C* (t-ratio)	I* (t-ratio)
Self-reported adult illness	-0.0364 (-7.676)	0.0019 (2.297)	-0.0382 (-8.185)
Self-reported adult injury	-0.0822 (-4.658)	-0.0021 (-2.141)	-0.0801 (-4.546)
Disability	-0.0543 (-7.229)	-0.0009 (-0.901)	-0.0531 (-7.130)

[1] Unstandardized concentration index [2] Standardized concentration index [3] Avoidable inequality

6.3.2. Equity in access to and utilization of health services

As seen above for self-reported illness and injury, concentration indices for utilization also reveal marked inequalities that sometimes favour the poor and at other times the better off. The LSDS questionnaire included a question about whether or not a person sought care, conditional on having reported an illness. The index of *horizontal inequity* (HI_{wv}) reveals a highly significant inequity that favours the poor. This, however, is irrespective of provider type. Analysis disaggregated by provider type would be more revealing as is attempted below.

With respect to doctor use, there is horizontal inequity favouring the rich. Hence, the poor utilize the services of a doctor less than what would be expected given their need levels. The same holds true for hospital services. Other important points worth noting are issues of

inequity related to the use of primary care facilities. As may well be expected, the concentration index for the utilization of primary health care services reveals pro-poor horizontal inequity. Hence the poor make more use of these first-level services even compared to their need. This contrasts with the utilization of doctor services, which showed pro-rich horizontal inequity.

One of the measures of access to health care is the time taken to reach a health facility to get treatment. The LSDS data indicate a highly significant pro-rich concentration index. In this dimension, the poor need more time to reach a health facility to get treatment. However, with respect to time taken to get treatment, there is no pro-rich bias.

As seen in the LSDS data, the OHS-95 data reveal horizontal inequities in seeking care favouring the disadvantaged, although the magnitude is much smaller. This is however, reversed in 1998, where it is found that there are statistically significant horizontal inequities favouring the rich. As indicated previously, it should be borne in mind that this is provider non-specific and needs to be disaggregated by provider type in order to have a clearer picture.

In the utilization of doctor's services, the findings of the OHS-98 data coincide with those of the LSDS-1993. There are significant horizontal inequities favouring the rich. Data from OHS-95, however, reveal a discord. The horizontal inequities in doctor use favour the disadvantaged.

The concentration indices for the utilisation of public (health) facilities indicate significant pro-poor horizontal inequities. This implies that the poor use public health facilities more than the rich. This finding is consistent in both the OHS-95 and OHS-98 data. However, the degree of horizontal inequity favouring the poor has increased tremendously in 1998 compared to 1995.

With respect to hospitalisation, there are horizontal inequities in both 1995 and 1998 that are biased towards the rich. The utilization concentration indices are presented in Table 6.2.

Table 6.2
Health care utilisation concentration indices

LSDS 1993			
Variable	C^1 (t-ratio)	C^{*2} (t-ratio)	HI_{wv}^3 (t-ratio)
Seek care	-0.1605 (-12.132)	-0.0003 (-1.228)	-0.1583 (-12.132)
Time to reach health facility	-0.1374 (-18.998)	-	-
Time to get treatment	-0.0142 (-1.539)	-	-
Doctor use	0.0815 (5.376)	-0.0001 (0.864)	0.0814 (5.370)
Hospital use	0.0845 (4.680)	-0.0006 (-0.911)	0.0323 (4.760)
PHC facility use	-0.1885 (-7.429)	0.0006 (0.885)	-0.2343 (-7.452)
OHS 1995			
Seek care	-0.0279 (-25.326)	-0.0226 (-25.646)	-0.0053 (-7.809)
Doctor use	-0.0443 (-33.679)	-0.0320 (-32.447)	-0.0123 (-13.79)
Hospitalisation	0.2686 (15.751)	0.2099 (27.612)	0.0587 (3.667)
Public health facility use	0.1460 (39.730)	0.1632 (53.014)	-0.0172 (-9.323)
OHS 1998			
Variable	C (t-ratio)	C* (t-ratio)	HI_{wv} (t-ratio)
Seek care	-0.0112 (-2.342)	-0.0296 (-8.010)	0.0176 (5.726)
Doctor use	0.0611 (14.724)	0.0003 (2.942)	0.0580 (13.961)
Hospitalisation	-0.0052 (-0.756)	-0.0218 (-9.220)	0.0164 (2.494)
Use of public health facility	-0.0527 (-13.767)	-0.0014 (-6.690)	-0.0431 (-11.292)

[1]Utilization concentration index (unadjusted for need)
[2]Utilization concentration index (adjusted for need)
[3]Horizontal inequality measure

As can be seen from the above table, inequities are observed to sometimes favour the poor and at other times the rich. For ease of understanding, it would be worthwhile to highlight consistencies and some irregularities, which will be discussed in the discussion section. With

respect to seeking care (provider non-specific), while in both 1993 and 1995 there was a horizontal inequity favouring the poor, this was reversed in 1998 to become pro-rich horizontal inequity. Utilisation of doctor services was pro-rich in the 1993 and 1998 data. However, counter-intuitively, the 1995 data showed horizontal inequities to the advantage of the poor. In all three data sets, use of hospitals and clinics exhibited a consistent trend. While hospitalisation showed a pro-rich horizontal inequity, the use of clinics was in favour of the poor.

6.4. DISCUSSION

It is shown that overall there are systematic inequalities in self-reported adult illness that are related to income status. In most instances the inequalities are found to be pro-rich. These are regarded as inequities, as they are related to one's socio-economic status and are avoidable given appropriate social policies and interventions. With respect to self-reported acute adult illness, the 1993 LSDS data indicate the presence of inequalities that are to the advantage of the poor, that is, excess avoidable self-reported illness exists among the better off. Although this finding appears anomalous, it is in line with the findings of many studies that have assessed equity in self-reported illness. Gilson and McIntyre (2001) using rate ratios between the highest and lowest income quintiles, report a similar finding in their study of equity in self-reported illness. In the same vein, Wagstaff (2001) using the illness concentration index reports the same for a number of developing countries including South Africa.

This counter-intuitive result does not conform to the commonly held view that the poor fall sick more often than do the rich. As stated in the introduction, self-reported adult illness may be influenced by a multiplicity of socio-economic factors including income, education, and previous experience with sickness and recall span. It is expected that people with higher levels of education and previous experience with illness tend to report illness episodes more

frequently (Henderson *et al* 1994). Furthermore, people with no access to health care and scanty education of health issues can regard certain bodily conditions as inevitable (Sen 2001), and thus refrain from reporting them as sicknesses. Given the apartheid era of discriminatory social policies, this may be a plausible explanation in this case.

For various socio-economic reasons, the poor may get used to, or may not be tempted to report a non-fatal illness. A person who is brought up in a community with a high burden of diseases and limited access to health care is likely to take some symptoms as normal when indeed, they are medically manageable (Sen 2001). Hence, it would be a mistake to consider low perception of illness as positive evidence of good health (*ibid*). A statement made by an elderly man from Bosnia and Herzogovina in the World Bank's consultation exercise *The Voices of the Poor*, (Narayan *et* al 2000) is but one of the many experiences that may result in under-reporting of sicknesses by the economically disadvantaged groups in society:

> *We are not allowed to get sick anymore because we have to pay for medication... What with?*

A study in Egypt (Nandakumar *et al* 2000), has also found that among other factors, income and education have a positive effect on illness reporting. This implies the tendency for the worse-off and the less educated (which in most cases also tend to be the poorest) to report illness episodes less frequently compared to the better off. To add to this evidence, a study of health system inequalities and inequities in Latin America and the Caribbean (Suárez-Berenguela 2000) reports that differences in self-reported health status by socio-economic groups are relatively small compared to large differences observed when health status is measured by objective measures of disease incidence and mortality. Thus, given the evidence, the counter-intuitive pro-poor inequality in self-reported illness should not be taken at face value. It needs further probing into the surrounding contextual factors.

In the 1995 and 1998 OHS data, however, there are statistically significant avoidable inequalities that are to the disadvantage of the poor. Furthermore, the strength of the avoidable inequalities has markedly increased in 1998. This indicates the existence of excess reported illness amongst the worse off that are avoidable and is therefore inequitable. This falls in line with expectation and common sense and may indicate a genuine increase in pro-rich inequity that may perhaps be related to the growing problem of HIV/AIDS. This finding may also partly be related to changes in the political climate of the country that was marked by the process of democratisation and installation of a government that upholds the welfare of the majority. This period is marked by the formulation of a democratic constitution, the Bill of rights, which among other things includes the right of access to health care, and allows for the development of a Patient's charter. The social context in which one lives heavily influences the self-perception of health and illness (Anand *et al* 2001). It is likely that reporting of illness as assessed by oneself is likely to vary with the social and political characteristics of a country. This may possibly have increased the awareness of the poorest that the diseases they are suffering from are not a normal part of human life and that they are avoidable through various interventions – curative and/or preventive.

Furthermore, increased access to health care by the poorest may also increase the likelihood of reporting an illness, as people are aware of the fact that they can do something to address it. There is substantial evidence that in states that provide more education and good health services, people are in a better position to perceive and report their illnesses (Sen 2002). The achievements in health care in South Africa in the post-1994 period may perhaps explain the trend. For example, between 1994 and 1997, 393 new clinics have been built mainly in rural areas, thus increasing physical access to services (Ntsaluba and Pillay 1998). Furthermore, the importation of Cuban doctors to fill posts mainly in under-served areas meant an increase in the availability of services.

The LSDS data further indicate that the poor report more days of sickness than the rich. This may probably indicate that the poor report a sickness when it is in its advanced stages. Furthermore, the poor spend more days out of work compared to the well off. These findings need to be given due attention in the formulation of policies that are geared towards poverty reduction. Again from the World Bank's consultation exercise, the Voices of the Poor, the following statement gives support to these findings:

> *...Poorer people are more often sick, sick for longer periods of time than the less poor...*

As may be expected, the longer time of sickness and days out of work that the poor experience are likely to result in aggravation of their economic disadvantage. The maintenance of the health status of the poor should, thus, assume centre-stage in poverty reduction programmes that are aimed at improving the living conditions of the poor. Similarly, health equity policies need to have a comprehensive approach, as the consequences of inequities in health are multidimensional and touch upon many areas of individual and household welfare.

The OHS 1998 data indicate significant avoidable inequalities in self-reported injury that favour the rich. There is a higher prevalence of self-reported injury among the poor than among their rich counterparts even after controlling for the confounding effect of age and sex. This is in line with studies from the developed world, which have also demonstrated higher rates of injury among those of lower socio-economic status (Zwi 2001). This may be a result of community, household, as well as individual-level factors. It may imply that the poor live in accident-prone environments or the individual persons may be highly susceptible to accidents due to lack of accident-preventing behaviour. These possible causal factors may, in turn, depend upon the socio-economic circumstances of the individuals, households and communities that are beyond their control. However, these hypothesized explanations need further probing in future research.

Self-reported chronic illnesses/disabilities show a higher concentration among the poorest. As these are long-standing illnesses that may limit the individual's productivity, their role in deepening existing poverty levels cannot be underestimated. This problem may pose a real threat to the government's efforts to improve the socio-economic conditions of the poor and to break the vicious circle of poverty and ill-health.

With respect to utilization, both the LSDS and OHS-95 data indicate horizontal inequity in seeking care (non-specific provider) that is favourable to the poor. The implication is that the rich do not seek health care as much as their health status (need) warrants. The magnitude of the pro-poor horizontal inequity, however, has diminished in 1995 compared to 1993. In 1998, the horizontal inequities in seeking care turned in favour of the rich. This, however, does not necessarily imply that that the rate of utilization by the poor has declined. It may well mean that utilization rate of the rich has increased. However, It has to be emphasized that since seeking care is not specific of provider type, seeking care from a traditional healer and from modern health care providers are combined, making this non-specific indicator less reliable.

Furthermore, the poor need more time to reach a health facility where they can get treatment. This depicts a supply-side problem- it indicates the presence of physical barriers to access. This, however, reflects the situation prior to 1994, the time before the change of government. It is expected that this problem has been minimized since then, as the overriding objective of the new government has been to redress past inequities in all spheres of life. As discussed previously the extensive clinic building campaign and the importation of doctors are but a few of the achievements that can be mentioned in this case. The time taken to get treatment once in the facility, however, does not show any statistically significant pro-rich bias. The absence of socio-economic inequalities in the time needed to get treatment may also indicate good standards of professional ethics among the health professionals.

Utilisation of services is broken down by provider type in all data sets. It is revealed that the poorest have less access to doctor services, in all data sets except OHS-95, which reveal a pro-poor horizontal inequity. It should be noted at this juncture that in the context of developing countries, specific questions of this nature may not always reveal the reality, as the distinction between a doctor and a non-doctor provider may not always be clear to the majority of the populace. This can be regarded as a potential weakness of household survey data and is likely to lead to distortions in the results. The erratic nature of the horizontal inequity in doctor use may possibly be confounded by this factor.

The horizontal inequities in the use of public health facilities that are consistently observed to be pro-poor suggest that government resources target the poorest of society. This finding is in line with empirical studies from Indonesia that indicated that subsidies to primary health care centres provide the best way of reaching the poor (van de Walle 1995). The post-apartheid government's focus on primary care facilities as evidenced by the aggressive clinic-building programme is, thus, a step in the right direction.

In summary, this chapter has shown the pervasiveness of pro-rich inequalities in self-reported adult illness and utilization of services. It has also revealed the potentially impoverishing effect of the higher concentration of illnesses in the poor as manifested by a protracted period of absence from work and days of illness. These are avoidable inequalities, that will be regarded as inequities. It is concluded that policy of health equity needs to have a holistic approach that takes account of the multidimensionality of the consequences of inequities in health, such as comprehensive poverty reduction programmes that create an enabling environment for the production of health and utilization of services. A major change is seen in self-reported acute adult illness in 1995 and 1998 as compared to 1993. Whereas the avoidable inequalities were in favour of the poor in 1993, pro-rich inequalities where observed in 1995 and 1998.

It should be borne in mind that subjective measures of health/ill-health (self-assessed health and self-reported illness) are not sufficient on their own to give a full picture of equity in adult illness and health. Self-perceived illness reporting is influenced by the socio-economic context, availability of health services and health-related information (Sen 2001). To complement the findings from studies that use subjective measures of health, it is necessary that future studies assess inequities in adult health by using objective measures. These may focus on the most common health problems facing the adult population in South Africa.

The analysis of trends in inequities both in self-reported illness and health care utilization is a significant contribution of the study to the existing literature on inequities in adult illness in South Africa. which is mostly based on cross-sectional data. Furthermore, while most studies on equity in health and health care in South Africa have not controlled for the confounding effect of the demographic variables age and sex, this study by using the indirect standardisation technique takes account of these confounding variables.

It has been shown that pre-1994 (before the change of government), avoidable inequalities in self-reported illness favoured the poor. However, the poor reported more days of sickness. In contrast in the post-1994 period, these have turned to be in favour of the rich, implying that the poor started reporting sickness more than the rich do. This coincides with the democratisation process of the country and the installation of a government that is committed to uplifting the welfare of the majority of South Africans.

The study also indicates the existence of pro-rich avoidable inequalities in self-reported injury and disability. This is also a notable contribution of the study, as most previous studies' focus was on multivariate analysis of the factors influencing acute and chronic illness only.

The study also makes contribution to the analysis of social policies that are aimed at allocating resources. It consistently reveals the existence of pro-poor horizontal inequities in the use of public health facilities implying the effectiveness of government targeting.

This chapter concludes the analyses of equity in health and health care. It is the aim of this dissertation to give a view of inequities using different dimensions of morbidity and mortality at the various stages in the life cycle of individuals. The analyses have shown the existence of considerable inequities to the disadvantage of the poorest of society. Given the escalating needs for health care on the one hand, and dwindling resources, on the other, it is necessary to scrutinize the efficiency and productivity of the health system. Efficiency gains can go a long way in releasing resources for redressing inequities. Thus, for this reason, the next two chapters will examine the technical efficiency and productivity of hospitals, which consume the lion's share of resources for health. The next chapter (Chapter 7) will deal with conceptual and methodological issues surrounding efficiency and productivity, while Chapter 8 will present the empirical analysis.

CHAPTER 7
EFFICIENCY AND PRODUCTIVITY: CONCEPTS AND MEASUREMENT

7.1. INTRODUCTION

The measurement of efficiency in the health care industry is complicated by the nature of the productive process. Measurement of the ideal output – improved health status – is difficult (Grosskopf and Valdmanis 1987). This is further complicated by the fact that health status is a function of many variables, many of which are exogeneous to the health sector.

The concept of efficiency, in addition to having varied definitions, is more often than not confused with a related term, *productivity*. The efficiency of a production unit signifies comparison between the observed and optimal quantities of its inputs and outputs (Fried *et al* 1993). This comparison may take three forms:

i. It could be a ratio of *actual* output to *potential* maximum output obtainable for a given input level; or

ii. the ratio of the *feasible* minimum input level required to the *observed* input consumption to produce a given output level; or

iii. The ratio may be a combination of the above two, having both input and output orientations.

The above descriptions refer to *technical efficiency* as the focus is on the measurement of production feasibilities. Efficiency could also be defined by reference to the production unit's behavioural objectives, such as profit maximisation or input minimisation. This denotes the notion of *economic efficiency*. Economic efficiency is measured by comparing the observed and optimum levels of the objective that the production unit is set to pursue subject to applicable constraints on quantities and prices.

Productivity, on the other hand, refers to output-to-input ratios. It is easily computable in a single input, single output scenario. The productivity of a production unit is influenced by three major factors. These include:

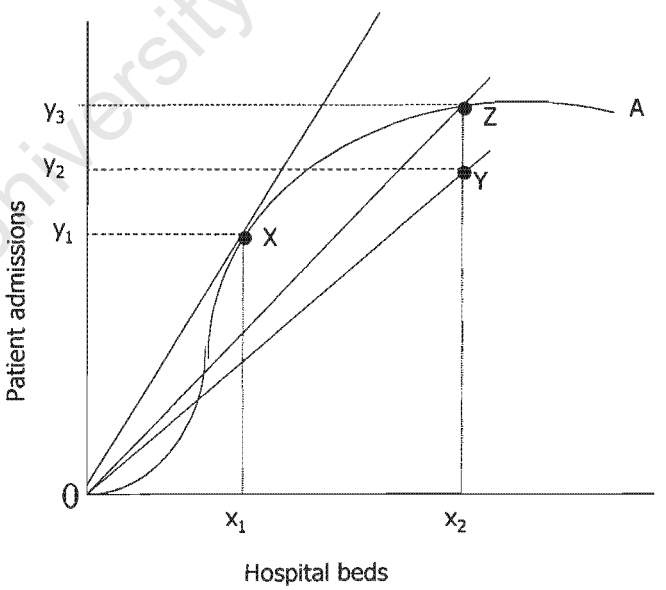
i. the production technology;

ii. efficiency of the production process; and

iii. environmental factors surrounding the operation.

Productivity is therefore related to efficiency by one of the above factors, that is, change of efficiency. The mere presence of technical efficiency does not necessarily mean that the maximum attainable level of productivity is reached. This is best illustrated in the figure below using the classical total product curve.

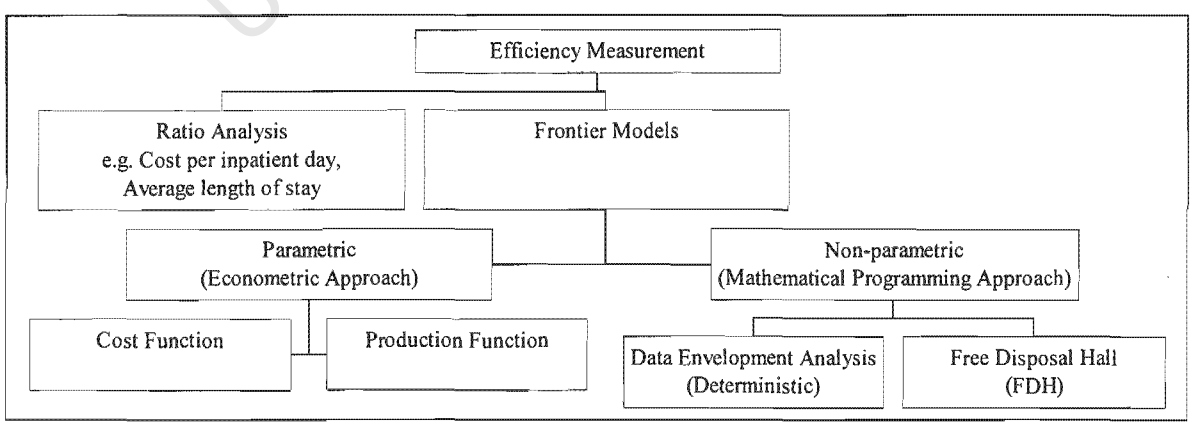
Figure 7.1
Productivity and technical efficiency



In the figure above, $0A$ represents the total product curve (TP), and *X*, *Y*, and *Z* represent three hospitals with their respective input-output combinations. Hospital *Y* that is below the TP curve is inefficient as it is feasible to increase its output without having to increase its input requirement. On the other hand, hospitals *X* and *Z* are technically efficient as they are located at the feasible maximum production point. Productivity of the hospitals in the above case is given by the slope of the line that connects the production point with the origin (*e.g.*, Y_2/X_2 for hospital *Y*.). If *Y* is to be relocated at point *Z*, both its efficiency and productivity increase ($Y_3/X_2 > Y_2/X_2$). However, although *Z* is technically efficient, there still exists room for further improvements in productivity by movement to *X* (the slope of the line connecting *X* to the origin is greater than that of *Z*). This introduces into the picture the concept of *optimal scale* of production. Thus, although hospital *Z* is efficient, by moving to point *X* it reaps economies of scale and consequently, boosts its productivity.

Efficiency measurement in health care may be performed using *ratio analysis* or *frontier models* (Smith and Maytson 1987). The techniques available are summarized in Figure 7.2 below.

Figure 7.2
Techniques of efficiency measurement



For a comprehensive understanding, some of the commonly used ratios in health care efficiency assessment are briefly described in Appendix 7.1. However, as this study makes use of frontier techniques (specifically data envelopment analysis), the forthcoming sections will dwell on the microeconomics of efficiency and productivity.

7.2. DISTANCE FUNCTIONS AND EFFICIENCY MEASURES

Farrell (1957), drawing upon the work of Debreu (1951) and Koopmans (1951), introduced a measure of productive efficiency that avoids the problems associated with traditional average productivity measures (ratios). He refuted the idea of an absolute measure of efficiency and proposed that efficiency be measured relative to a best-performance frontier determined by a representative peer group. Furthermore, he provided the definitions and computational framework for evaluating technical and allocative (in)efficiency. Thus, the starting point for any discussion of frontier methods of efficiency measurement is Farrell's seminal work.

The first step in efficiency measurement is that of defining the reference technology relative to which performance will be evaluated. This makes use of production sets and distance functions. Distance functions allow describing a multiple input/output technology without having to specify a behavioural objective such as cost-minimisation or profit-maximisation (Coelli *et al* 1998). Both input and output distance functions may be specified. An input distance function considers a minimal proportional contraction of the input vector, given an output vector. On the other hand, an output distance function characterises the production technology by looking at a maximal proportional expansion of the output vector, given an input vector.

The multiple-input, multiple-output production technology may be defined using the technology set, T as (Coelli *et al* 1998, Linna 1999):

$$T = \{(x, y) : x \text{ can produce } y\} \tag{7.1}$$

This represents the technically feasible set of all input-output vectors.

Where:

$x = (x_1, \ldots, x_n) \in R_+^n$ denotes a non-negative $n \times 1$ vector of inputs; and

$y = (y_1, \ldots, y_m) \in R_+^m$ denotes a non-negative *mx1* vector of outputs.

The production technology set, T can be represented using two equivalent forms: the *input requirement set*

$$L(y) = \{x : (x, y) \in T\} \tag{7.2}$$

and the *output set*

$$P(x) = \{y : (x, y) \in T\} \tag{7.3}$$

The production technology set has the following properties:

i. $0 \in P(x)$: inactivity is a possible option, that is, it is possible to produce zero output out of a given set of inputs.

ii. $y \notin P(0)$: it is not possible to produce outputs with zero level of inputs.

iii. $x > x^*, x^* \in L(y) \Rightarrow x \in L(y)$: The input requirement set satisfies strong disposability of inputs. This implies that increasing any of the inputs does not decrease output.

iv. $y \leq y^*, y^* \in P(x) \Rightarrow y \in P(x)$: the output set satisfies strong disposability of outputs.

v. $P(x)$ is bounded: we can not produce unlimited levels of outputs with a given set of inputs.

vi. T is a closed set.

vii. $P(x)$ and $L(y)$ are a convex set: if two combinations of output levels can be produced with a given input vector x then any average of these output vectors can also be produced.

In the following discussion, a description of the production technology set will be provided, and the concepts and measurement of distance functions and the associated measures of technical efficiency elaborated.

A. Input requirement set

Production technology can be represented using an input requirement set as:

$$L(y) = \{x : (x, y) \in T\} \tag{7.4}$$

This representation denotes all input bundles that are capable of producing at least y units of output. The input requirement set has an isoquant represented by:

$$IsoqL(y) = \{x : x \in L(y), \lambda x \notin L(y), \lambda \in [0,1)\}. \tag{7.5}$$

The isoquant, $IsoqL(y)$, gives all input bundles that give exactly y units of output. In turn, the isoquant has its efficient subset which is defined as:

$$Eff\ L(y) = \{x : x \in L(y),\ x' \leq x \Rightarrow x' \notin L(y)\} \tag{7.6}$$

In an input-based approach, a functional representation of a multiple output production technology is provided by the input distance function, which was introduced by Shephard (1970). The input distance function is defined by:

$$D_I(x, y) = max\left\{\lambda : \left(\frac{x}{\lambda}\right) \in L(y)\right\} \tag{7.7}$$

$D_I(x,y)$ has a value greater than or equal to one, and it follows from definition (7.5) that

$$IsoqL(y) = \{x : D_I(x,y) = 1\}. \tag{7.8}$$

This leads to defining the Farrell input-oriented measure of technical efficiency as

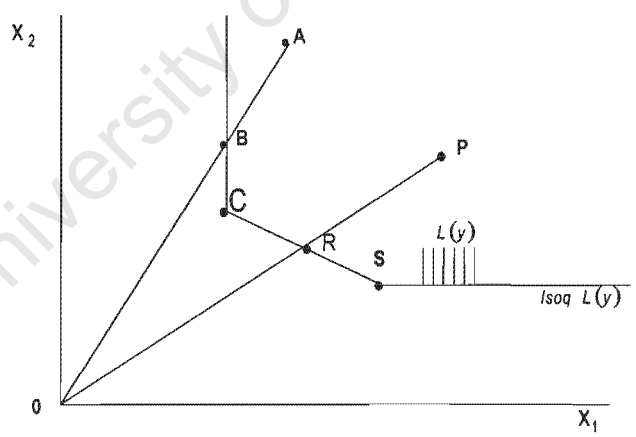
$$TE_I(x,y) = \min\{\lambda : \lambda x \in L(y)\} \tag{7.9}$$

$TE_I(x,y) \leq 1$, and is the reciprocal of the input distance function. That is

$$TE_I(x,y) = \frac{1}{D_I(x,y)} \tag{7.10}$$

Figure 7.3 below is presented to illustrate the concept and computational aspects of the input distance function and its relationship with technical efficiency.

Figure 7.3
Input distance function and technical efficiency



Input vectors represented by points such as A represent a combination of inputs X_1 and X_2 (*e.g.*, physician hours and beds) used by a specific hospital to produce output vector Y (*e.g.*, inpatient days). The input requirement set $L(y)$ is bound from below by the isoquant $Isoq\ L(y)$.

As can be seen from the figure above, the values of the input distance function for the hospitals using input vectors A and P are:

$$\lambda_A = \frac{0A}{0B} \tag{7.11}$$

$$\lambda_P = \frac{0P}{0R}. \tag{7.12}$$

It can be seen clearly that values of the input distance function are always greater than or equal to one. While they are greater than one in the above case for those hospitals using input vectors that fall on the isoquant (B, C, R and S) the input distance function takes values of one.

Furthermore, in the previous definitions it has been stated that input-oriented technical efficiency is just the reciprocal of the input distance function. It therefore follows that the values of the input-oriented technical efficiency for hospitals using input vectors A and P will be:

$$TE_I^A = \frac{0B}{0A} \tag{7.13}$$

$$TE_I^P = \frac{0R}{0P}. \tag{7.14}$$

In both cases the value of the technical efficiency is less than one. Technical efficiency for those hospitals using input vectors located on the isoquant is equal to one. Thus, the input distance function and Farrell technical efficiency have identical values for production units that are on the isoquant.

Three major observations are in order from Figure 7.3.

i. Input vectors A and P can be contracted radially[1] and still remain capable of producing the output level Y represented by the isoquant. Hence they are technically inefficient.

ii. Input vectors located on the isoquant (B, C, R, and S) can not be radially contracted and remain capable of producing output vector y.

iii. Although the hospital using input vector B is efficient in the Farrell sense of technical efficiency, there still exists room for improvement. It is observed that it can reduce its use of input X_2 without increasing the amount of input X_1 required and remain capable of producing the given output vector Y. In other words, it contains a *slack* in input X_2. Thus, input vector B is not part of the efficient subset $EFF\ L(y)$ (see Equation 7.6).

According to Koopmans (1951), a production unit is technically efficient if:

i. It is impossible to increase the production of any output without reducing the production of at least one other output or increasing the use of at least one input; and

ii. A reduction in any input requires an increase in at least one input or a reduction in at least one output.

Referring back to Figure 7.3, this denotes that the hospital using input vector B does not qualify to be called technically efficient in the sense of Koopman's definition of technical efficiency. The presence of the slack in input X_2 disqualifies it. The implication worth emphasising is that the Farrell technical efficiency is a *necessary* but not *sufficient* condition for Koopman's technical efficiency.

[1] Along a line/radius that extends from the origin to the point of the input vector under consideration.

B. Output distance function and output-oriented measure of technical efficiency

The foregoing discussion on the input distance function and input-oriented measure of technical efficiency has spelt out the basics of the issue of technical efficiency. Therefore, this discussion on output distance function and output-based technical efficiency will just highlight the peculiarities of output orientation to efficiency measurement.

In the measurement of technical efficiency oriented toward output augmentation, the production technology can be represented with an output set (production possibility set):

$$P(x) = \{x : (x, y) \in T\}. \tag{7.15}$$

An output distance function considers a maximal proportional expansion of the output vector given an input vector. On the production possibility set, $P(x)$the distance function is defined as:

$$D_o(x, y) = \min\left\{\theta : \left(\frac{y}{\theta}\right) \in P(x)\right\}. \tag{7.16}$$

$D_o(x, y) \leq 1$, if the output y belongs to the production possibility set, $P(x)$. Its value is equal to one for those output vectors that are located on the production possibility frontier. The distance measure is the inverse of the factor by which the production of all output quantities could be increased while still remaining within the feasible production possibility set for the given input level. The corresponding Farrell output-oriented measure of technical efficiency, which is the inverse of the output distance function is defined as:

$$TE_o(x, y) = \max\{\theta : \theta y \in P(x)\}. \tag{7.17}$$

7.3 TECHNICAL AND ALLOCATIVE EFFICIENCY

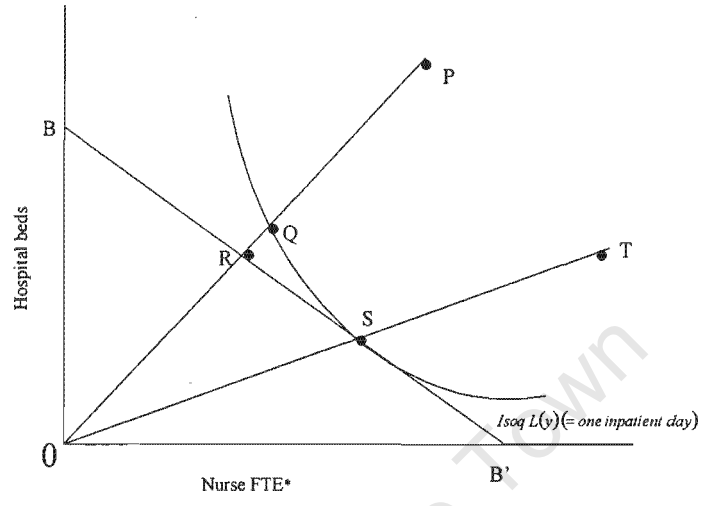
The efficiency of a production unit consists of two components, technical and allocative efficiency (Farrell 1957). Technical efficiency reflects the ability of a production unit to maximise output from a given set of inputs. It is a technological concept and focuses on the production process and on task organisation. Allocative efficiency on the other hand, reflects the extent to which inputs are used in optimal proportions given their respective prices and marginal productivities. While technical efficiency focuses on input and output quantities, allocative efficiency considers input and output prices. The overall efficiency of a production is referred to as *economic efficiency* and incorporates both technical and allocative efficiencies.

As discussed in the previous sections, efficiency measures may have input or output orientation. The focus of an input-oriented measure of technical efficiency is on the possibility of a proportional reduction of inputs used without changing the output quantities. The output-oriented measure addresses the question: by how much could output quantities be expanded without changing the quantity of inputs used?

Figure 7.4 below depicts the relationship between technical and allocative efficiency measures proposed by Farrell (1957) using a set of hospitals using two inputs to produce one output under the assumption of constant returns to scale (CRS)[2].

[2] CRS implies that when all input quantities increase by a certain proportion, output increases by the same proportion. If output increases by a greater (smaller) proportion, it is called increasing (decreasing) returns to scale.

Figure 7.4
Technical and allocative efficiency



In the above illustration, a hospital produces its output (one inpatient day) using a combination of two inputs (nurse FTE and hospital beds). Being technically efficient means 'locating on an *isoquant*', that is on the frontier. Thus, hospitals operating at points *Q* and *S* are regarded as technically efficient, while hospitals operating at points *P* and *T* are technically inefficient. For the hospital operating at point *P*, the input-oriented technical efficiency is given as:

$$TE_I^P = \frac{0Q}{0P} \tag{7.18}$$

This represents the ratio of the minimal input required to the actual input use, given the input mix used by *P*. The ratio $\frac{QP}{0P}$ represents the percentage by which all inputs could be reduced without a reduction in output. If the hospital producing at point 'P' is to be efficient, it has to relocate itself at point *Q*. Technical efficiency takes values between zero and one ($0 \leq TE_I \leq 1$).

Technically inefficient production units have a TE_I value less than one, while the efficient ones have a TE_I value of 1. This can be seen by calculating the *technical efficiency* of the hospital producing at *S*:

$$TE_I^S = \frac{0S}{0S} = 1 \tag{7.19}$$

Given the input prices, the *isocost* line *BB'* represents the minimum cost of producing one unit of output. Allocative efficiency demands that production should take place at the point where the *isoquant* is *tangential* to the *isocost line*. Given this definition, the hospital producing at point *Q*, which is regarded as technically efficient, is *allocatively inefficient*. Only the hospital operating at point *S* is both *technically and allocatively efficient*. The allocative efficiency (AE) for the hospital at point *P* is given as:

$$AE^P = \frac{0R}{0Q} \tag{7.20}$$

The ratio $\frac{RQ}{0Q}$ represents the percentage reduction in production costs that would occur if production were to occur in the allocatively efficient point *S*. The overall (economic) efficiency (EE) for the same hospital is:

$$EE^P = \frac{0R}{0P}. \tag{7.21}$$

The measure of overall efficiency has the advantage that it easily decomposes into technical and allocative efficiencies:

$$EE^P = \frac{0R}{0P} = \frac{0Q}{0P} \times \frac{0R}{0Q} \tag{7.22}$$

that is, $EE = TE \times AE$.

The above measures represent *input-oriented, radial* measures of efficiency. They are input-oriented, as their focus is on the measurement of variations in input use between different

firms for a standardised output. The measures are radial as they are taken along a ray from the origin in the input-output space. This implies that the current input-output mix determines the firm's technology and, any possible increase in efficiency will be achieved if inputs are reduced proportionally, with output proportions held constant (Valdmanis 1992). The radial nature of the efficiency measures allows comparison of firms with similar input-output mixes.

Output -oriented measures can also be illustrated on the input-output space by taking an example of a production process involving one input and two outputs. The two measures of efficiency are equivalent under the assumption of constant returns to scale (CRS).

The constant returns to scale assumption is only appropriate when all production units are operating at an optimal scale. In the presence of variable returns to scale (VRS), which may be increasing or decreasing returns to scale, technical efficiency is decomposed into *pure technical* efficiency and *scale* efficiency. A difference between the CRS and VRS technical efficiency scores of a production unit suggests the presence of scale inefficiency. This is best illustrated using the following figure (Figure 7.5).

Figure 7.5
Scale efficiency

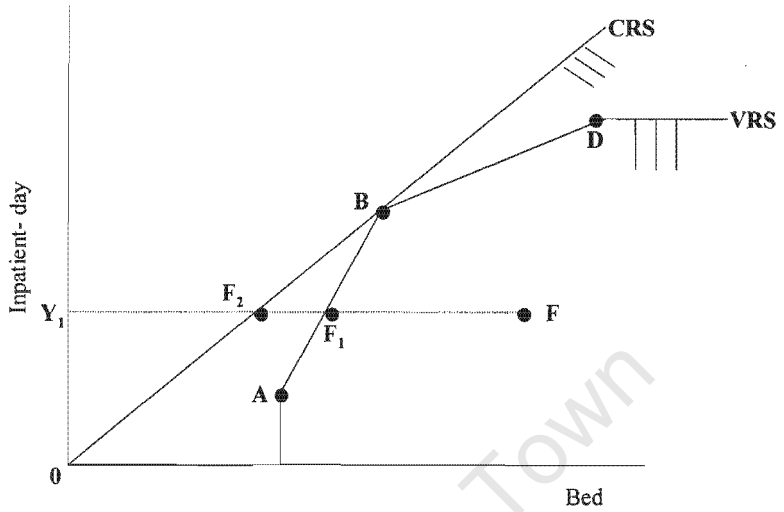


Figure 7.5, is a representation of a single input, single output production technology, with both the CRS and VRS assumptions. By inspecting the figure and based on the previous discussion on the various concepts of technical efficiency, the efficiency status of the hospitals can be described as follows (Table 7.1):

Figure 7.1
Various forms of technical efficiency

Hospital	Overall technical efficiency (CRS)	Pure technical efficiency (VRS)	Scale efficiency	Returns to scale
A	-	+	-	IRS
B	+	+	+	CRS
D	-	+	-	DRS
F	-	-	-	IRS

- = No + = Yes
CRS = constant returns to scale; DRS =decreasing returns to scale; IRS = increasing returns to scale

Calculation of scale efficiency is demonstrated as follows using the case of the inefficient hospital *F*.

$$TE_{CRS} = \frac{Y_1F_2}{Y_1F} \tag{7.23}$$

$$TE_{VRS} = \frac{Y_1F_1}{Y_1F} \tag{7.24}$$

The difference between the overall technical efficiency (CRS) and pure technical efficiency (VRS) is a measure of scale efficiency, and is given as:

$$Scale\ efficiency = \frac{Y_1F_2}{Y_1F_1} = \frac{CRS\ TE}{VRS\ TE}. \tag{7.25}$$

The scale efficiency measure of hospital *F*, in other words, is the ratio of the *average product* of a hospital operating at F_1 to that of the hospital operating at the point of optimal scale, B.

As seen in the foregoing discussion, empirical estimates of efficiency measures involve two steps:

i. estimation of the frontier; and

ii. calculating individual production unit deviations from the frontier.

So far, we have assumed that the production frontier against which the efficiency of a hospital is measured is known. However, in practice we have to fit the frontier empirically from sample data. Thus the efficiency scores computed are from best performance observed within the sample, and therefore, are measures of relative efficiency.

Currently there are two main approaches used in estimating the production frontier (Seiford and Thrall 1990, Coelli *et al* 1998):

i. Parametric approach

In this method, *a priori* well-defined functional form such as the Cobb-Douglas form is fitted to the data such that all productive units lie on or below it.

ii. Non-parametric approach

In this approach, no *a priori* assumptions are made of the functional form of the underlying technology. The main technique in this group is the method of Data Envelopment Analysis (DEA) that estimates efficiency scores by using linear programming techniques. This technique is to be used in efficiency estimation in this study. The study being a policy-oriented action research, it is preferred to use this technique (DEA) rather than simple ratios or the econometric approaches of cost and production functions, as the approach can provide clear answers to such important managerial questions as (Fried and Lovell 1994):

- Which hospitals are the most efficient in the system?
- If the inefficient hospitals were to operate as efficiently as their efficient peers, by how much could resource use be minimized, so that resources are released for other pressing needs? (Or conversely, by how much can we augment outputs without any additional resources?)
- What are the characteristics of the efficient hospitals that need to be emulated to guide management to improve its productive efficiency?
- How do we select appropriate 'role models" that could serve as possible benchmarks for a programme of performance improvement in the hospital system?
- What is the desired scale of operations? Which facilities need to be expanded or downsized to achieve the optimum scale of operations? And by how much should they be trimmed or expanded?
- How do we account for differences in performance brought about by non-discretionary inputs/outputs?

Answers to the above-mentioned issues are some of the crucial factors that make the use of DEA in performance assessment of South African hospitals the preferred option. Therefore, an attempt is made to thoroughly discuss this technique in the following section.

7.4 DATA ENVELOPMENT ANALYSIS (DEA)

It was stated earlier that efficiency measurement entails the estimation of an empirical production function (efficiency frontier) also known as the *envelopment surface*, from observed sample data. Before proceeding to the DEA computational formulae, it would be worthwhile to have an understanding of the mechanism by which the envelopment surface is determined using a simple hypothetical example.

Assume we have five maternity hospitals using two inputs (nurse FTE and beds) to produce an output of 100 child deliveries. The quantity of the inputs utilised by each of the five hospitals is given in Table 7.2 below.

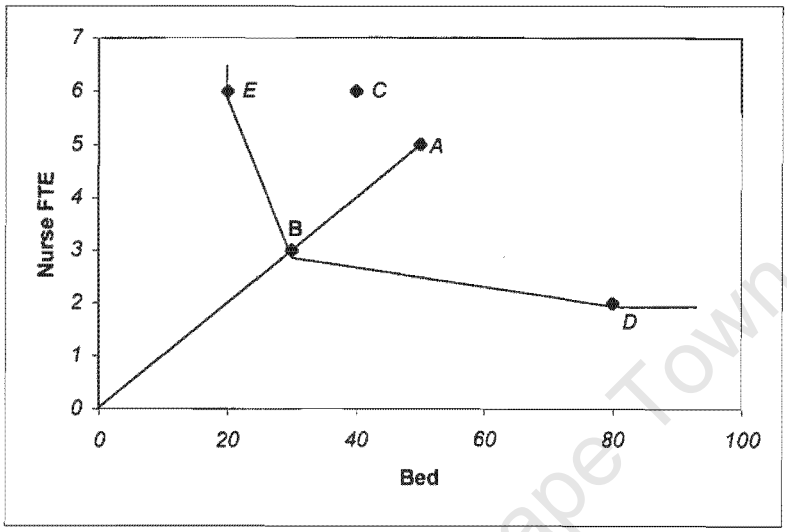
Table 7.2
Hypothetical input-output data

Hospital	Input one (Beds)	Input two (Nurse FTE)
A	50	5
B	30	3
C	40	6
D	80	2
E	20	6

From the input data given in Table 7.2, it can be seen that hospital *C* is outperformed by hospital *E*, as *E* uses less of input one and no more of input two to produce the same output of 100 child deliveries. Similarly, hospital *A* is dominated by hospital *B*, as *A* uses more of both inputs to produce the same level of output. Thus, the envelopment surface is

formed by the efficient hospitals *B, D,* and *E* as observed in the following figure (Figure 7.6).

Figure 7.6
Construction of the envelopment surface



Hospitals *A* and *C* are operating outside the efficient frontier, and are therefore regarded as relatively inefficient hospitals. The above figure is just an illustration of how the DEA method establishes the efficient frontier against which the efficiency of a unit in the group is to be assessed. The following, sections discuss the DEA computational formulae under the various assumptions of returns to scale.

7.4.1. CONSTANT RETURNS TO SCALE (CRS) DEA MODEL

Building on Farrell's seminal work, Charness *et al* (1978) proposed the non-parametric technique of *DEA* for measuring the relative efficiencies of decision-making units (DMUs)[3] such as schools, post offices and hospitals. *DEA* uses linear programming methods to establish the frontier from sample data. The efficiency of a DMU is then measured relative to the efficiency of all others in the group, subject to the restriction that all DMUs lie on or below the frontier

[3] Intended to emphasise an orientation toward managed entities in the public and /or not-for-profit sectors.

(Bjurek *et al* 1990, Seiford and thrall 1990, Coelli *et al* 1998). This is performed by solving a series of LP problems.

An intuitive way to introduce DEA is using a ratio form. The measure of efficiency of any decision making unit (DMU) as proposed by Charness *et al* (1978) is obtained as:

$$Efficiency = \frac{weighted\ sum\ of\ outputs}{weighted\ sum\ of\ inputs}. \tag{7.26}$$

The optimal input and output weights are obtained by solving the following mathematical programming problem:

$$Max\ h_0 = \frac{\sum_{r=1}^{s} u_r y_{rj_0}}{\sum_{i=1}^{m} v_i x_{ij_0}} \tag{7.27}$$

Subject to:

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, \quad j = 1...., j_0, ..., n$$

$$u_r \geq 0; \quad r = 1, ..., s$$

$$v_i \geq 0; \quad i = 1, ..., m$$

Where:

y_{rj} = amount of output r from unit j

x_{ij} = amount of input i to unit j.

u_r = weight given to output r

v_i = weight given to input i

n = number of DMUs

s = number of outputs

m = number of inputs

The above formulation involves finding values for u and v which are most favourable to the unit being studied and maximise its efficiency. A non-linear ratio model, however, poses a

problem as it has an infinite number of solutions. To avoid this, it is converted to a linear programming *multiplier* problem as follows:

$$Max\ h_0 = \sum_{r=1}^{s} u_r y_{rj_0} \tag{7.28}$$

Subject to:

$$\sum_{i=1}^{m} v_i x_{ij_0} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum v_i x_{ij} \leq 0, \quad j = 1,...,n$$

$$u_r, v_i \geq 0$$

For each DMU a linear programming (LP) problem is solved by maximisation of the weighted sum of outputs for DMU j_0 with the restriction that the weighted sum of its inputs equals one. A further restriction is that for all DMUs, the weighted sum of outputs minus the weighted sum of inputs must be less than or equal to zero. This last constraint implies that all DMUs are on or below the production possibility frontier.

7.4.2. Variable returns to scale (VRS) DEA model

As indicated previously, where the CRS assumption does not hold, the TE measure is confounded by scale efficiency. To disentangle the effects of scale efficiency, it is necessary to use a DEA model with a variable returns to scale assumption. To this end Banker *et al* (1984) developed an extension of the original CRS model. In a VRS DEA model, the LP problem to be solved is:

$$Max\ h_0 = \sum_{r=1}^{s} u_r y_{rj_0} + u_0 \tag{7.29}$$

Subject to:

$$\sum_{i=1}^{m} v_i x_{ij_0} = 1$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum v_i x_{ij} + u_0 \leq 0, \quad j = 1,\ldots,n$$

$$u_r, v_i \geq 0$$

$$u_0 \lesseqgtr 0$$

The new term introduced, u_0, corresponds to an intercept (Bjurek *et al* 1992). If U_0=0 it implies constant returns to scale; and values less than zero and greater than zero indicate decreasing and increasing returns to scale respectively.

Compared to ratio analysis and the parametric methods, DEA has the following advantages (Charness *et al* 1996, Seiford and Thrall 1990, Ferrier and Valdmanis 1996, Thanassoulis and Dyson 1992): it

i. Does not impose restrictions on the functional form of the technology (input-output relationship) or the distribution of the inefficiency term. DEA allows each DMU to have different functional forms. This is in contrast to the stochastic frontier models (SFM), where *a priori* assumptions are made about the input-output relationships (technology) (e.g. Cobb-Douglas, CES) and the distribution of the decomposed error term (e.g. half-normal). DMUs are not allowed to take the functional form that suits them.

ii. DEA easily accommodates multiple inputs and outputs, without the requirement for homogeneous measurement units. In ratio analyses, the inability to accommodate multiple inputs and outputs at the same time requires that measurement units be homogeneous. For example in assessing hospital average costs using two outputs (outpatient visits and inpatient days), there is a requirement that the two outputs be expressed in the same denominator. This entails weighing the outputs by pre-selected weights, which in the case of South Africa is done by attaching a weight of 1:3 (inpatient day: outpatient visit) to arrive at a common denominator, *patient day*

equivalent (PDE). This raises a host of questions such as justifying the 1:3 ratio of inpatient day to outpatient visits. This can confound the resulting efficiency ratings, as we don't know how much of the efficiency rating is due to the pre-selected weights. On the other hand, in the SFMs, the accommodation of many inputs and outputs requires a large data set, for there may arise problems of degrees of freedom. Moreover, there is a potential problem of multicollinearity. Furthermore, in a multiple output situation, the parametric techniques require input price data, which most often are difficult to obtain in the hospital industry.

iii. Focuses on observed best-practice frontiers rather than central tendencies (averages), as is the case in the econometric methods, where the frontier passes through a line that averages the performance of the best and the least performers. DEA is oriented toward individual DMUs that are regarded as responsible for utilizing inputs to produce outputs.

iv. Unlike the econometric techniques, DEA does not make behavioural assumptions (e.g. profit maximization) about the DMUs.

v. Offers a performance measure that is convenient and free of monetary factors, and yet has a straightforward cost interpretation.

vi. Produces specific input-output targets that would render an inefficient DMU relatively efficient. It furthermore identifies efficient "peers" for those DMUs that are not efficient. This helps the inefficient DMUs to emulate the functioning of their efficient peers so as to improve their efficiency. Ratio analyses and the econometric methods do not help identify a "role model".

vii. DEA helps identify both the sources (input and output) and amounts of inefficiencies. This is a deficiency of the econometric approaches, which do not provide information on sources and estimates of the inefficiency associated with these sources. Thus, no guides of remedial actions are provided.

viii. DEA also allows for weight restrictions so as to incorporate managerial preferences in terms of relative importance levels of various inputs and outputs (Cooper *et al* 2000). For example, if output 1 is regarded at least twice as important as output 2, this can be incorporated into the DEA model by adding the constraint: $\mu_1 \geq 2\mu_2$.

However, despite these strengths, there are two major drawbacks to this method (Coelli *et al* 1998). Firstly, DEA is *non-stochastic*. It does not capture random noise (e.g. epidemics, strike, unforeseen maintenance expenditure). Any deviation from the estimated frontier is regarded as being due to inefficiency. The econometric methods handle this problem by decomposing the error term into two: one capturing statistical noise outside the control of the DMU, and another one gauging inefficiency. Secondly, DEA is non-statistical, in the sense that it is not possible to conduct tests of hypotheses regarding the inefficiency and the structure of the production technology. It does not provide the usual diagnostic tools with which to judge the goodness-of-fit of the model specifications. Moreover, unlike the econometric approach, DEA does not provide a model for predicting the performance of an organization for years that are not included in the evaluation or for evaluating DMUs that are not part of the sample.

Ferrier and Valdmanis (1996), however, argue that, these drawbacks may not be as serious as they seem to be. Firstly, as there is no *a priori* specification of the functional form of the technology, specification error that might show up as a noise is ruled out. Secondly, as inputs and outputs are measured in their natural physical units, a measurement error is most unlikely. Moreover, a recent breakthrough has been achieved in remedying the non-statistical and non-stochastic nature of DEA (see for example Sengupta 1998).

Despite the criticisms levelled at it, DEA remains the preferred method in the non-profit sector where (Coelli *et al* 1998):

i. random noise is less of a problem;

ii. multiple-output production is relevant;

iii. price data is difficult to find; and

iv. setting behavioural assumptions such as profit (cost) maximisation (minimisation) is difficult.

7.5. THE MALMQUIST PRODUCTIVITY INDEX

The Malmquist productivity index (MPI) that was proposed by Caves *et al.* (1982), measures total factor productivity (TFP)[4] change between two data points in terms of ratios of distance functions. A Malmquist index greater than one indicates growth in productivity, while a value of less than one indicates a decline. The Malmquist index approach does not require *a priori* behavioural assumptions about the production technology. It also does not require input and output prices. These characteristics make it more appealing for measuring productivity in the public sector.

Following Caves *et al.* (1982), the output-oriented Malmquist productivity index is defined as:

$$M_0^t\left(y^t, x^t, y^{t+1}, x^{t+1}\right) = \frac{D_0^t\left(y^{t+1}, x^{t+1}\right)}{D_0^t\left(y^t, x^t\right)} \tag{7.30}$$

This ratio index measures productivity changes at time periods t and $t+1$ resulting from changes in technical efficiency with reference to the technology in time period t. It can also be measured with respect to the technology in time $t+1$ as:

[4] It is the average product of all inputs (in contrast, partial factor productivity is the average product of a single input; e.g. child deliveries per midwife).

$$M_0^t\left(y^t, x^t, y^{t+1}, x^{t+1}\right) = \frac{D_0^{t+1}\left(y^{t+1}, x^{t+1}\right)}{D_0^{t+1}\left(y^t, x^t\right)} \tag{7.31}$$

Färe *et al.* (1994) defined the output-oriented Malmquist productivity index as the geometric mean of the two indices (7.30 and 7.31 above):

$$M_0^{t,t+1}\left(y^t, x^t, y^{t+1}, x^{t+1}\right) = \left[\frac{D_0^t\left(y^{t+1}, x^{t+1}\right)}{D_0^t\left(y^t, x^t\right)} \times \frac{D_0^{t+1}\left(y^{t+1}, x^{t+1}\right)}{D_0^{t+1}\left(y^t, x^t\right)}\right]^{1/2} \tag{7.32}$$

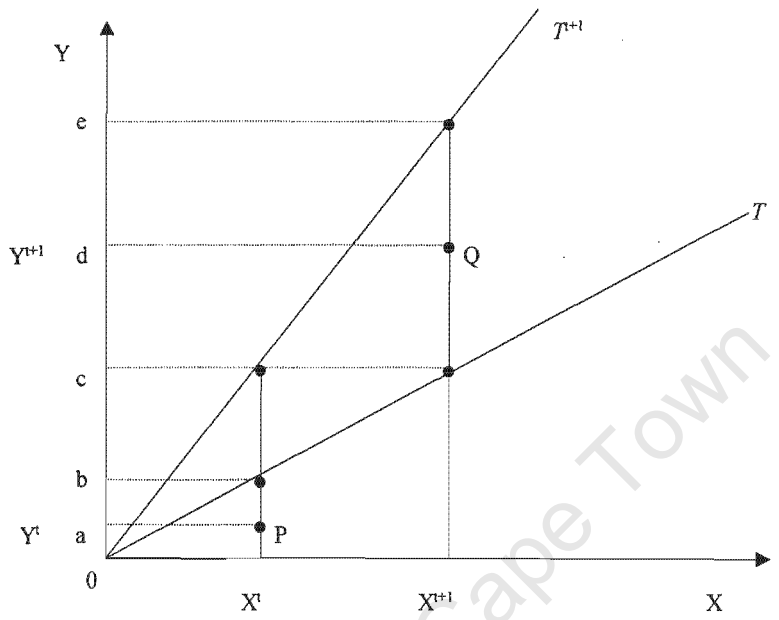
Färe *et al.* (1994), further decomposed the Malmquist productivity index into two parts: efficiency change and change in production technology. Thus,

$$M_0\left(y^t, x^t, y^{t+1}, x^{t+1}\right) = \left[\frac{D_0^{t+1}\left(y^{t+1}, x^{t+1}\right)}{D_0^t\left(y^t, x^t\right)}\right]\left[\frac{D_0^t\left(y^{t+1}, x^{t+1}\right)}{D_0^{t+1}\left(y^{t+1}, x^{t+1}\right)} \times \frac{D_0^t\left(y^t, x^t\right)}{D_0^{t+1}\left(y^t, x^t\right)}\right]^{1/2} \tag{7.33}$$

where the first term on the right hand side measures *efficiency change* and the second term measures *technical change*.

Färe *et al.* (1994) listed a number of different methods to calculate the Malmquist productivity index. However, the most preferred of these methods is the one that uses DEA-like linear programming techniques. Four linear programming problems are solved for each hospital to compute four distance functions to measure the total factor productivity change between two periods under a constant returns to scale technology (*ibid*). The technical efficiency change can further be decomposed into pure efficiency-change component and scale-change component by solving two additional linear programming problems under variable returns to scale technology (Coelli *et al.* 1998). The definition and measurement of the MPI is illustrated in Figure 7.7.

Figure 7.7
Output-based Malmquist productivity index



T^t and T^{t+1} represent the production technology in two periods, *t* and *t+1*. The hospital produces at point *P* in period *t* and at point *Q* in period *t+1*. Using the last formula, the decomposition of the MPI from the above figure is given as:

$$Efficiency\ change = \frac{0d/0e}{0a/0b} \tag{7.34}$$

That is, the efficiency change is the ratio of the Farrell technical efficiency in period *t+1* to that in period *t*. The technical change is the geometric mean of the shift in technology evaluated at x^{t+1} and the shift in technology evaluated at x^t is as follows:

$$Technical\ change = \left[\frac{0d/0c}{0d/0e} \times \frac{0a/0b}{0a/0c}\right]^{1/2} \tag{7.35}$$

7.6. HOSPITAL INPUTS AND OUTPUTS

The selection of inputs and outputs for a DEA study requires careful thought as the distribution of efficiency is likely to be affected by the definition of outputs and the number of inputs and outputs included (Magnussen 1996).

Two schools of thought dominate the discussion on the definition and measurement of the output of health care organisations (Mersha 1989):

i. the *process approach*, which asserts that the output of a health care organisation consists of services provided by the different units such as the X-rays, laboratory procedures, patient days *etc*; and

ii. the *outcomes approach*, regards the above processes only as intermediate steps leading to the desired change in patient's health status. According to this approach, therefore, output should be measured in terms of the end result or outcome, that is improved health.

Although there is a general consensus that the ultimate measure of output should be an improvement in the quantity and quality of life, practical difficulties limit the use of the outcomes approach (Mersha 1989, Clewer and Perkins 1998). First, it is easier to measure and define processes (services) in health care than changes in health status. Second, changes in health outcome can not be entirely attributed to health care. Health is multi-dimensional and affected significantly by a host of other socio-economic factors. Moreover it may take considerable time for full health outcome improvements to be evident. Consequently, output is measured as an array of intermediate outputs (health services) that supposedly improve health status (Grosskopf and Valdmanis 1987).

Buttler (1995) classifies hospital output into four broad categories: inpatient treatment, outpatient treatment, teaching and research.

Measuring hospital output by such variables as inpatient days or outpatient visits, does not capture the case-mix and the quality of service rendered. Even though the use of Diagnosis-related groups (DRGs) may handle the problem of hospital case-mix, the absence of data makes its use limited in most developing countries. Within the context of developing countries, stratifying hospitals according to their level may to some degree take account of the case-mix and factors such as staffing pattern and medical technology used that are likely to affect the quality of care delivered.

Inputs in hospital production are classified as labour, capital and supplies. The labour input can be disaggregated into the various professional groups such as physician, nurse and administrative staff. In most studies, capital is proxied by the number of hospital beds.

In selecting inputs and outputs for a DEA study, the following issues need to be noted (Golany and Roll 1989):

i. the factor (input/output) should be related to one or more objectives of the study;
ii. the factor should convey pertinent information not included in other factors; and
iii. data on the factors should be readily available and reliable.

7.7 A BRIEF SURVEY OF PREVIOUS STUDIES OF HOSPITAL EFFICIENCY AND PRODUCTIVITY

There are very few efficiency and productivity studies of health care programmes using frontier models in developing countries, especially in Sub-Saharan Africa. This is evidenced by the dearth of literature on this subject. Most studies for policy and management purposes

employ ratio analysis. However, given the limitations of such a technique, it is not possible to obtain comprehensive and reliable information on the state of efficiency and productivity in a multiple input-output industry.

A few econometric analyses of cost and efficiency of health facilities exist in developing countries. These studies have found the existence of significant inefficiency (e.g. Anderson 1980, Barnum and Kutzin 1990).

A study of the efficiency of public sector hospitals in three provinces of South Africa (McMurchy 1995) using econometric techniques (production and cost functions) revealed the existence of technical and allocative inefficiency. However, this study used a deterministic Cobb-Douglas model, which is a very restrictive functional form and other *ad hoc* cost functions. Estimation of the models was done using the method of ordinary least squares (OLS), in which case the estimate of the intercept is inconsistent and biased downward (Lovell 1993). The study classifies hospitals with a zero residual (i.e. the actual expenditure equals the predicted) as efficient, and those with a positive residual (the actual greater than the predicted) as inefficient. Using the above method, it was not possible to provide information on the actual level of technical efficiency. Thus, a difference between the actual and predicted expenditure of 1 percent and another of 50 percent will all be regarded as equally inefficient. From a policy viewpoint this is a serious shortcoming, for the study does not explicitly state by how much resources are to be contracted (or conversely, by how much output can be maximised) to get the inefficient hospitals to the efficient frontier.

Despite a relatively wide application of the non-parametric technique of data envelopment analysis in health care efficiency assessment, there is no indication that it has been applied in

Sub-Saharan African health care systems. However, in the developed world DEA has been relatively widely used in the field of health care.

Among the first to apply DEA in health care include: Nunamaker (1983), Sherman (1984) and Banker *et al* (1984). The various DEA studies in health care have focussed on assessing efficiency and productivity, the dynamics of productivity change, and the determinants of efficiency and productivity. Examples of issues examined include, assessment of the efficiency of urban and rural hospitals (e.g. Ferrier and Valdmanis 1996, Ozcan and Luke 1993), efficiency across ownership types (e.g. Burgess and Wilson 1996 compared profit and not-for-profit hospitals, academic and non-academic hospitals, *etc.*).

Most studies have indicated the presence of inefficiency of varying magnitudes, and therefore, the potential for conservation of scarce health care resources. For example a study of Turkish hospitals (Ersoy *et al* 1997) found that about 91 percent of the 573 acute care general hospitals studied were inefficient compared to their peers. A similar study of Greek hospitals (Giokas 2001) reveals that inefficiencies contribute to about 4.1 percent of health care costs in the gross domestic product.

Studies of the determinants of (in)efficiency have shown that institutional factors at the discretion of the management, as well as environmental factors beyond its control (e.g. health-related legislation, socio-economic factors of the catchment population) affect the health facilities' performance (e.g. see Ferrier and Valdmanis 1996, Luoma *et al* 1996, Ozcan and Luke 1993, Rosko *et al* 1995, Valdmanis 1992). Some of the factors that influence efficiency cited in the literature include: ownership, location of health facility, teaching status, payment source, occupancy rate and quality.

With respect to the determinants of hospital (in)efficiency, the most debated area has been the influence of ownership type. Many studies of the effect of ownership type on efficiency have come up with equivocal or counter-intuitive, results (see for example, Ozcan *et al* 1992, Mobley and Bradford 1997, Buergess and Wilson 1996). This may partly be attributed to confounding factors not accounted for in the study design. For example, Mobley and Bradford (1997) in their study of the behavioural differences among hospitals in California concluded that location-specific factors confounded the relationship between ownership type and efficiency.

Despite the relatively large number of frontier-based studies of efficiency, few studies have used frontier techniques to measure hospital productivity (Linna 1999). As is the case with efficiency studies, there is no evidence of the use of frontier methods of productivity assessment in a developing country setting.

From the foregoing discussion, it can be seen that hospital inefficiency is rampant, and that there is a great potential for releasing resources to be used for other purposes. Previous studies have also found that factors both within and outside the hospital organization, i.e. factors controllable by management and factors beyond its control, influence the technical efficiency and productivity of hospitals positively or negatively. This discussion will form the basis for the empirical analysis on the determinants of inefficiency in the next chapter.

7.8. CHAPTER SUMMARY

In this section an attempt has been made to discuss the various issues and theoretical models underpinning the measurement of efficiency and productivity in health care. Specifically the following have been surveyed:

i. Definitions of the concepts of efficiency and productivity and their inter-relationship;

ii. Definitions and implications of commonly used ratios and rates of hospital performance, their merits and demerits;

iii. Concepts and maxims in production economics that underlie the measurement of efficiency and productivity using the frontier approach;

iv. The DEA method of efficiency measurement and DEA-based Malmquist productivity index used to decompose productivity change into efficiency change and change of technology;

v. Issues surrounding the definition and measurement of hospital outputs, as the controversies surrounding its measurement need to be highlighted; and

vi. A brief survey of previous studies of hospital efficiency and productivity and their determinants so as put the present study in context

CHAPTER 8

TECHNICAL EFFICIENCY AND PRODUCTIVITY OF SOUTH AFRICAN HOSPITALS

8.1. INTRODUCTION

This chapter deals with the assessment of the technical efficiency and productivity of a sample of public sector hospitals. In a situation where there is slackening of economic growth and an increase in the demand for services, it is necessary to be efficient in order to address inequities inherited from a system that was unjust. This chapter will demonstrate the magnitude and form of existing inefficiency with a view to appreciating the extent of efficiency gains that policy makers could potentially reap.

It can safely be said that within the context of developing countries, a health system would be regarded as efficient if its hospital sector is efficient. The hospital sector is a large consumer of scarce health care resources. Although the actual percentage varies from country to country, hospitals in developing countries absorb an average of 50-80 per cent of the public sector health resources (Barnum and Kutzin 1993). In 1992/93, hospitals in SA consumed about 89 per cent of the total public sector expenditure on health (McIntyre *et al.* 1995, Castro-Leal *et al.* 1999). Given that hospitals account for such a considerable fraction of resources, their evaluation is integral to the assessment of efficiency of the system.

South Africa is still grappling with the legacy of the apartheid system. There are glaring disparities in health indicators and access to health care between the most and the least privileged population groups (Gilson and McIntyre 2001). In an effort to redress these disparities, a key policy goal of the Department of Health is to achieve universal access to primary health care services (South Africa 1997). This goal has to be achieved within the context of stagnation in real per capita health budgets associated with the stringent budget deficit reduction targets set by the South African government. Thus, the development of primary care services has to be funded through resource redistribution

from hospitals. Given the macroeconomic and socio-demographic realities of the country, the need for assessing the efficiency and productivity of hospitals and its correlates cannot be overemphasised.

It is common knowledge that the health care system (especially hospitals) in developing countries is inefficient. The World Bank's policy study on *Financing Health Services in Developing Countries* (Akin *et al.* 1987) indicates that one of the major problems of African health care systems is the inefficiency of government health programmes, the others being problems of allocation and inequity.

In the presence of inefficiency, costs of service delivery are inflated. This undermines the cost-recovery ratio and any other stated benefits of cost-sharing schemes. Furthermore, given the economic realities of SSA countries, the task of redressing inequalities in access to health care cannot be achieved without a concomitant improvement in efficiency. Inefficiency is more likely to breed further inequity.

This chapter seeks to examine the technical efficiency and productivity of a sample of hospitals in South Africa. The findings will help deepen the understanding of the magnitude of inefficiency and its causes in SSA. Its specific objectives are to:

(i) evaluate the technical and scale efficiency of non-academic acute care hospitals in the Eastern, Northern and Western Cape provinces of South Africa;

(ii) identify some of the factors that are likely to influence the (in)efficiency of hospitals; and

(iii) assess changes in the productivity of acute care hospitals in the Western Cape province.

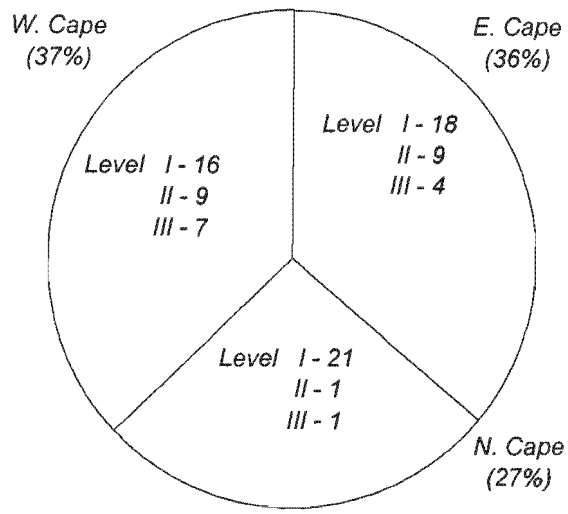
8.2. METHODS

8.2.1. SOURCE OF DATA

Data are derived from the annual statistical publications of the former Department of Health of the Cape Province and the new Department of Health, Provincial Administration of the Western Cape. In assessing technical efficiency, data for the year 1992/93 is used. This covers three provinces, which at that time were under the same administration. However, the panel data used in assessing productivity is limited to hospitals in the Western Cape. This covered the period between 1992/93 and 1996/97. The selection of the time period is entirely dictated by the availability and completeness of the data. The selection of hospitals within these three provinces is mainly based on the availability of useable data.

The reports include data on inputs, outputs and other relevant hospital service statistics. For the purpose of this study, on the basis of their size and scope of activity, the hospitals are classified into three groups. These in ascending order of their size and complexity are Level I (n=55), Level II (n=19), and Level III (n=12). For assessing changes in productivity, only 10 Western Cape provincial hospitals are used. This is dictated by the availability of data. The provincial distribution of the hospitals is as shown in Figure 8.1 below.

Figure 8.1
Provincial distribution of the sampled hospitals

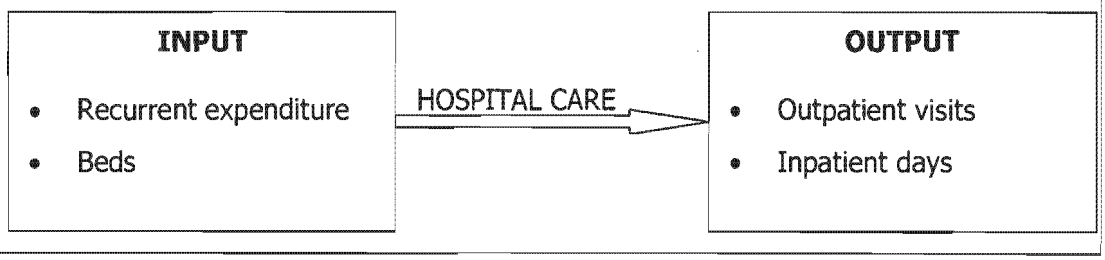


8.2.2. THE EMPIRICAL DEA MODEL

Input-oriented constant and variable returns to scale DEA models are used in computing the efficiency scores. The choice between input-/output-oriented DEA models is made according to the flexibility of inputs or outputs. An input-oriented model is preferred in this study, because hospital managers are unlikely to have control of the demand side factors which are determined partly by the health-care seeking behaviour of the public. As detailed personnel and supplies data were not available, expenditure was used to represent recurrent inputs in this evaluation. As in most other studies, capital is proxied by the number of beds. Out of the various measures of capital that they used, Grosskopf and Valdmanis (1987) regard the number of beds as a reliable measure of capital. This may partly be due to the fact that beds have a prominent influence on the potential production of inpatient days. The empirical model is outlined diagrammatically in Figure 8.2.

Figure 8.2

The DEA model



8.2.3. THE MEASUREMENT OF TECHNICAL EFFICIENCY AND PRODUCTIVITY

Data envelopment analysis (DEA) and the DEA-based Malmquist productivity index are used in assessing the technical efficiency and productivity of the sampled hospitals (see details in Chapter 7 for detailed discussion on methodology). Because DEA is a non-parametric technique, the statistical properties of the DEA efficiency measures are not fully understood. Thus, to test for the robustness of the technical efficiency estimates the jackknife analysis is used. This technique helps to assess if there were extreme outliers which affected the frontier and efficiency scores. In the Jackknife analysis, a limited number of samples are obtained by omitting one observation at a time (Efron 1982). In this case, each efficient hospital is dropped one at a time from the analysis and the efficiency scores re-estimated. To this end, six additional DEA models are estimated in Level I hospitals, three and two models in Levels II and III hospitals respectively. The similarity of the efficiency rankings between the model with all the hospitals included and those based on dropping each efficient hospital one at a time is tested by using the Spearman rank correlation coefficient. A correlation coefficient of 1 implies that the rankings are exactly the same. A value of 0 indicates the absence of relation between the rankings and reverse ranking is implied by a value of –1.

The technical efficiency scores and productivity indices are calculated using the data envelopment analysis programme, version 2.1 (DEAP 2.1) (Coelli 1996).

8.2.4. THE ECONOMETRIC MODEL OF THE DETERMINANTS OF INEFFICIENCY

The efficiency scores of level I hospitals only are examined using a censored Tobit model to identify factors that influence inefficiency. The other two levels are excluded, as their numbers are not sufficiently large to undertake a meaningful analysis. In the econometrics literature, distributions similar to those of DEA scores are best regarded as censored normal distributions. Censoring occurs when the independent variables are observed for the entire sample, but for some observations we have only limited information about the dependent variable (Madala 1983). In assessing factors influencing hospital efficiency, some hospitals are regarded as inefficient, as their DEA efficiency scores assume positive values that are strictly less than one. Those that are regarded as efficient have their DEA efficiency scores clustered at one, because DEA efficiency scores have an upper limit of one.

In the presence of censoring, estimates of the parameters obtained using the method of Ordinary Least Squares (OLS) will be inconsistent (Long 1997, Gujarati 1995). To overcome this problem, the Tobit model is estimated using the method of maximum likelihood (ML).

In the Tobit model, for computational convenience, it is preferred to assume a censoring point at zero (Greene 1993). To this end, the DEA technical efficiency scores are transformed into inefficiency scores, left-censored at zero using the formula:

$$Inefficiency\ score = \left(\frac{1}{TE\ score}\right) - 1. \tag{8.1}$$

The Tobit model is defined as follows:

$$y_i^* = \beta_i x_i + u_i \tag{8.2}$$

$$y_i = y_i^* \quad if\ y_i^* > 0 \tag{8.3}$$

$$y_i = 0 \quad if\ y_i^* \leq 0 \tag{8.4}$$

where $u_i \sim N(0,\sigma^2)$, and

y_i^* is a latent variable that is observed for values greater than zero and is censored for values less than or equal to zero.

y_i is the observed inefficiency score

β_i is a $k \times 1$ vector of unknown parameters

x_i is a $k \times 1$ vector of explanatory variables. The independent variables (x's) are observed for all cases.

A very common interpretation of the Tobit model is in terms of an underlying latent variable (y^*), of which y is the realized observation (Breen 1996). The change in the latent variable (y^*) with respect to x_k is given as (Long 1997):

$$\frac{\partial E(y^* \mid x)}{x_k} = \beta_k \tag{8.5}$$

The interpretation of the coefficients is the same as those of the OLS, except that in the above, the independent variable is latent. They show the effect of a change in a given independent variable (x_k) on the expected value of the unobserved latent variable (y^*), holding all other independent variables constant.

Due to data constraints, some important variables within the hospitals and their operating environment such as location (urban/rural), payment source, quality of care, teaching status, *etc*, which may influence (in)efficiency have been omitted. This is to be regarded as a limitation of the model. Data constraint is a pervasive problem in developing country health systems. Moreover, since analytical models of this kind have rarely been used in a developing country setting (specifically sub-Saharan Africa), the health information systems are not designed to capture most of the data that are vital to assess performance of the health system and its determinants. Hence, caution has to be exercised in interpreting the results of the model. Nevertheless, the study being one of the pioneers, does demonstrate that by utilizing the available sources of information, important evidence

for policy can be generated in data-scarce settings of sub-Saharan Africa, and that such efforts can also induce changes in the health information systems. The available explanatory variables included and their expected signs are given in Table 8.1.

Table 8.1
Variable influencing hospital inefficiency: measurement and expected signs

Variable	Measurement	Expected sign
Occupancy rate (OCC)	%	-
Average length of stay (ALS)	Days	+/-
Outpatient visits as a proportion of inpatient days (OUTPRO)	%	-
Location dummy (PROV1*)	Prov1 (Eastern Cape)=1 0 otherwise;	+/-
Location dummy (PROV2)*	Prov2 (Northern Cape)=1 0 otherwise	+/-

*The reference province is the Western Cape

The occupancy rate is a composite index that incorporates inpatient admissions, the average length of stay and the number of beds. However, multicollinearity is not expected to be a problem, as the value of the occupancy rate is determined by the relative position of each of its components and not a single one. A simple correlation analysis between the occupancy rate and the average length of stay also suggests the absence of a significant relationship (r=0.1291, p=0.3476). The literature suggests that hospitals with a higher occupancy rate are also found more efficient when the DEA methodology is used, compared to their counterparts with a lower occupancy rate (Giokas 2001). Hence it is expected that the coefficient of *OCC* will have a negative sign indicating an inverse relationship between *OCC* and hospital inefficiency.

ALS has an indeterminate expected sign. On the one hand, as the *ALS* composes part of the *OCC* numerator, other things constant, a higher *ALS* implies a higher *OCC.* In this situation, it will therefore be expected to assume a negative sign. On the other hand, an unduly long *ALS* associated with a lower *OCC might* indicate managerial inefficiency (e.g.

delays in therapeutic and diagnostic interventions due to, for example, improper scheduling, poor nursing care, *etc*), and therefore may manifest a positive sign.

The variable *OUTPRO* is expected to have a negative sign if hospitals have an underutilised inpatient service capacity. Furthermore, if the outpatient volume of the hospitals under scrutiny is very limited, an increase in the production of outpatient visits may, ceteris paribus, increase efficiency, as providing both out- and in-patient services under the same roof may curb unnecessary resource consumption (e.g. transporting patients for admission to the hospital from a polyclinic outside the hospital).

The sign for provincial location is indeterminate *a priori*, as it is influenced by a number of socio-economic, demographic, geographic and epidemiological factors whose effects may not be easily determined. It is included to control for confounding effects that may stem from province-specific factors that are not easy to capture.

Statistical analyses are performed using STATA 5 statistical software (Statacorp 1997).

8.3. RESULTS

8.3.1. General characteristics

The three levels of hospitals are found to have different sizes as measured by the bed-size. Level I community hospitals are the smallest. Those of Level II are about twice as large and Level III hospitals have the largest size, which is about eight times that of Level I.

There is also a marked gap in the activity levels of the three groups of hospitals. The volume of output in terms of admissions, outpatient visits and inpatient days is the highest in Level III hospitals. The mean recurrent expenditure in Level III hospitals is about twenty three times those of Level I hospitals. The summary statistics are presented in Appendix 5.

8.3.2. Technical efficiency

The DEA models estimated for the three groups of hospitals indicate the presence of a marked deviation of the efficiency scores from the respective best-practice frontiers. Only 13 percent of the sampled hospitals operate efficiently as compared to their peers. The overall level of technical inefficiency in the three groups of hospitals is in the range of 35.1 to 46.8 percent. This implies that on average, the hospitals use a level of resources, which is 35.1 to 46.8 percent in excess of what is required for the given level of outputs. Level I hospitals have the highest mean technical efficiency score (Table 8.2).

TABLE 8.2
TECHNICAL EFFICIENCY SCORES

Technical efficiency measure	Mean	SD	Min	Max	Hospitals on frontier
Level I Hospital					
CRS	0.740	0.124	0.518	1	6
VRS	0.828	0.174	0.468	1	17
Scale	0.900	0.124	0.518	1	6
Level II Hospitals					
CRS	0.681	0.204	0.283	1	3
VRS	0.825	0.192	0.442	1	8
Scale	0.825	0.147	0.508	1	3
Level III Hospitals					
CRS	0.695	0.162	0.516	1	2
VRS	0.820	0.125	0.671	1	3
Scale	0.845	0.140	0.641	1	2

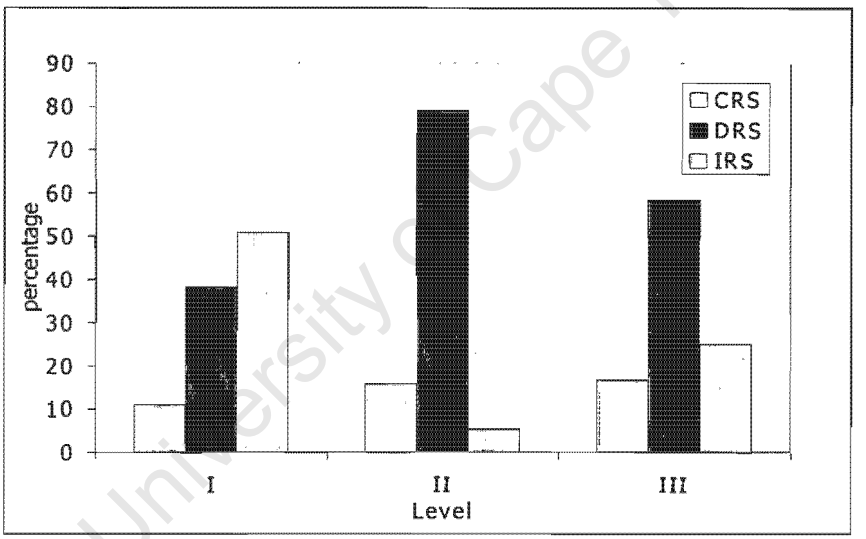
Decomposition of the overall levels of technical efficiency shows that while pure technical efficiency in all levels is more or less similar, scale efficiency is higher in Level I hospitals. Significant proportions of hospitals in Levels II and III operate at a non-optimal scale (Table 8.3).

TABLE 8.3
DECOMPOSITION OF OVERALL TECHNICAL INEFFICIENCY

Hospital	Pure technical inefficiency (%)	Scale inefficiency (%)
Level I	20.8	11.1
Level II	21.2	21.2
Level III	22.0	18.3

Most of the hospitals operate at variable returns to scale. Decreasing returns to scale is predominant in Levels II and III hospitals, while increasing returns to scale is more prevalent in Level I hospitals (Figure 8.3).

Figure 8.3
Returns to scale by hospital level



About half of the hospitals experience decreasing returns to scale. On the cost side, this implies that half of them experience diseconomies of scale. Only 13 percent of the hospitals operate at an optimal scale.

8.3.3. Input savings

If the relatively inefficient hospitals operate as efficiently as their peers on the frontier, the total saving in recurrent expenditure in Level I hospitals would have been about R29.5 million in 1992/93. This amounts to about 26 percent of the recurrent expenditure on

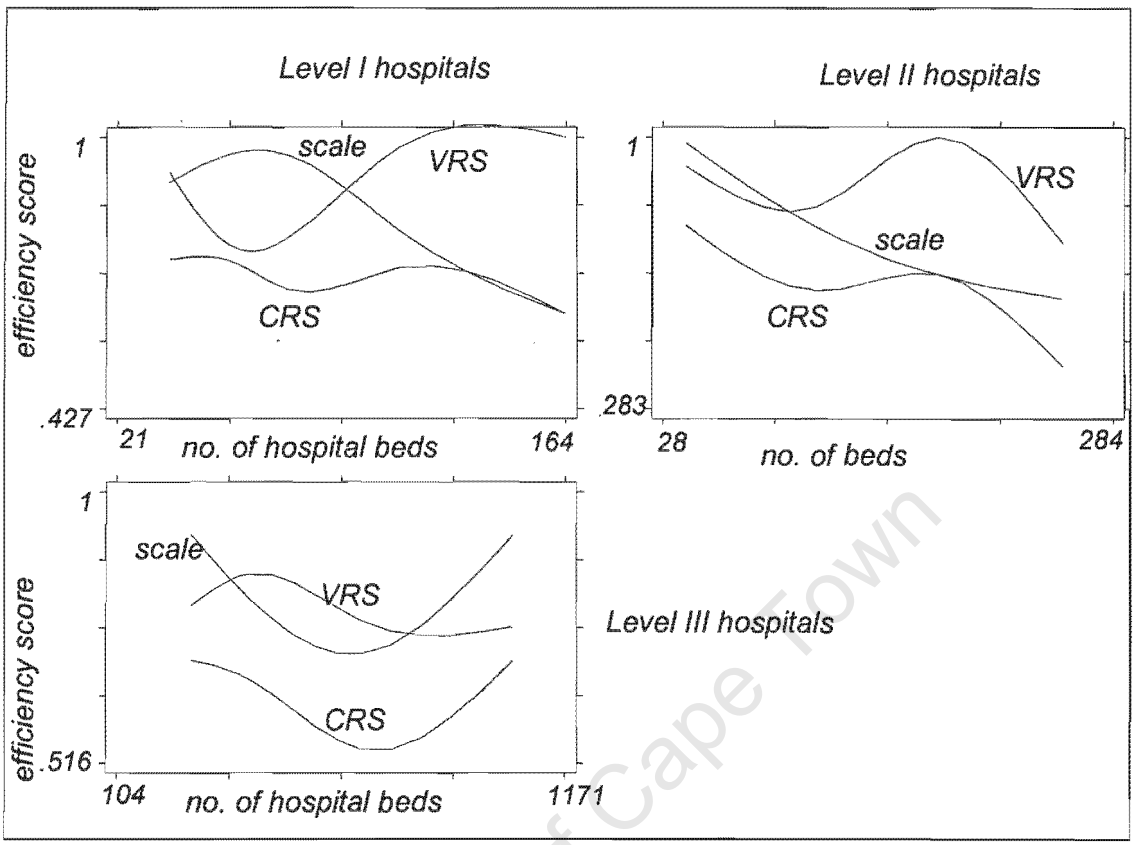
these hospitals in that year. The number of beds could also be reduced by about 30 percent. In Level II hospitals, efficiency savings would amount to about 33 percent of the total recurrent expenditure, and bed-size could be trimmed by about 39 percent. Similarly in Level III hospitals, efficiency savings to the amount of 33 percent of the recurrent expenditure could be achieved. Bed-size could also be reduced by the same proportion.

8.3.4. Technical efficiency and bed-size

In relation to the number of beds, the overall technical efficiency in Level I hospitals reaches its highest mean value of 0.795 when the bed size is between 40-60. This is when the scale efficiency is on its increasing side and the pure technical efficiency declines. Efficient scale for these group of hospitals seems to be located within a bed size range of 60-80. Pure technical efficiency declines progressively until about a bed size of 80 and picks up thereafter. Thus the smaller (< 40 beds) and larger (> 80 beds) hospitals of Level I seem to have a higher degree of pure technical efficiency compared with their peers having a bed-size in the middle. However, the rise in technical efficiency at the higher levels of bed size is overshadowed by the steeply declining scale efficiency. Thus overall technical efficiency decreases when the increasing returns to scale at the smaller bed size levels are exhausted.

Out of 31 Level I hospitals that had a bed size of less than 60, about 71 per cent experienced increasing returns to scale, indicating the existence of economies of scale. The proportion of those that experienced decreasing and constant returns to scale respectively was 9.7 and 19.3 per cent. The relationship between bed-size and technical efficiency is presented in Figure 8.4.

FIGURE 8.4
TECHNICAL EFFICIENCY AND HOSPITAL BED-SIZE



As in the case of Level I hospitals, the CRS technical efficiency in Level II hospitals also exhibits a trend of decline with increasing number of beds. The pure technical efficiency seems relatively better with small (< 50) and large (> 150) bed sizes. The scale efficiency for this category of hospitals shows a progressive rate of decline, with the largest drop in efficiency (of about 26 per cent) occurring when the number of beds increases to over 150. Constant returns to scale is observed in only three hospitals (3/19) with bed-size of less than 50. Fifteen of the hospitals (15/19) with a bed-size of more than 50 have decreasing returns to scale, which in other words implies that they experience diseconomies of scale.

In Level III hospitals the CRS and scale efficiency plots assume a U-shape. Both Levels reach a relative minimum level at a bed-size of 600-800. After this minimum level there is a sustained increase in efficiency. This is contrary to Levels I and II where the minimal

increase in CRS efficiency is not maintained as a result of a sharp drop in scale efficiency at higher levels of bed size. As can be seen from Figure 8.4, the pure technical efficiency behaves in a manner which is diametrically opposed to those of Levels I and II. It has a positive slope at the smaller bed size levels (< 400 beds) followed by various rates of decline as the bed size increases. Thus, whereas the pure technical efficiency increases at higher bed size levels in Levels I and II, in Level III hospitals, it decreases tremendously. The decrease in pure technical efficiency is, however, more than offset by a greater increase in scale efficiency.

8.3.5. Provincial variations in technical efficiency

There is some degree of variation in efficiency levels among the three provinces. In Level I and Level III hospitals, the mean efficiency scores of the Western Cape province are higher than the other two provinces. However, in Level II hospitals those of the Eastern Cape Province are the highest. The distribution of the efficiency scores by province is given in Table 8.4.

TABLE 8.4
DISTRIBUTION OF TECHNICAL EFFICIENCY SCORES BY PROVINCE

Province	Level I hospital			Level II hospital			Level III hospital		
	mean	SD	no.	mean	SD	no.	mean	SD	no.
Eastern Cape	0.69	0.18	18	0.73	0.25	9	0.64	0.09	4
Northern Cape	0.72	0.18	21	0.54	0	1	0.51	0	1
Western Cape	0.81	0.15	16	0.65	0.16	9	0.75	0.18	7

8.3.6. SENSITIVITY ANALYSIS OF THE TECHNICAL EFFICIENCY SCORES

The results of the jackknife analysis and Spearman's rank correlation indicate that the DEA technical efficiency scores are robust and stable. In all levels of hospitals the Spearman's rank correlation coefficients indicate a statistically significant degree of concordance in the

rankings of the original model and each of the DEA models estimated using the jackknife approach. The results of the sensitivity analysis are depicted in Table 8.5.

Table 8.5
The stability of DEA results in regard to outlier hospitals

	Level I hospitals	
	Spearman's rho	P-value
Model 1	0.9946	0.0000
Model 2	0.9856	0.0000
Model 3	0.9973	0.0000
Model 4	0.9979	0.0000
Model 5	0.9445	0.0000
Model 6	0.9674	0.0006
	Level II hospitals	
Model 1	0.9757	0.0000
Model 2	0.9881	0.0000
Model 3	0.9488	0.0000
	Level III hospitals	
Model 1	0.8702	0.0005
Model 2	0.6055	0.0484

The correlation coefficients of all the models in Levels I and II hospitals are seen to be very close to 1, suggesting that the efficiency rankings between each of these models and the original model where all of the hospitals are included, are nearly identical. The p-values indicate rejection of the null hypotheses that the efficiency scores of each of the models and the original estimate are independent. This stability of the technical efficiency scores gives credence to the estimated DEA model. In Level III hospitals, although the Spearman's rho is statistically significant and shows stability in the rankings of the technical efficiency scores, the correlation is not as strong as it is in the two lower-level hospitals. The smallness in the number of this level of hospitals may perhaps have an effect on the stability of the DEA scores, as the sample size in these type of hospitals is barely adequate, given the number of inputs and outputs used in the model.

8.3.7. THE DETERMINANTS OF INEFFICIENCY

The regression results of the econometric model of the factors influencing inefficiency is presented in Table 8.6.

TABLE 8.6
ESTIMATION RESULTS FOR TOBIT MODEL

Variable	Coefficient	t-ratio
constant	1.9112	8.216
OCC	-0.0171	-7.940
ALS	-0.0319	-1.289
OUTPRO	-1.5110	-5.407
PROV1	0.0282	0.292
PROV2	-0.0819	-0.842
$\chi^2_{5)}$ (p-value)	56.72 (0.0000)	

The likelihood ratio χ^2 statistic rejects the null hypothesis that all the coefficients except the intercept are not significantly different from zero. Thus, despite the ommission of a number of variables the model offers some explanation for the variations in the tendency to be inefficient.

The bed occupancy variable has a sign consistent with expectation. It is negatively related to inefficiency. This implies that higher occupancy levels are associated with higher levels of efficiency. A one percentage point increase in the occupancy rate results in a 1.7 percentage points decrease in the tendency to be inefficient. The coefficient of *ALS* is not statistically significant even at the 10 percent level of significance. The number of outpatient visits as a proportion of inpatient days (*outpro*) has a very high statistical significance. In Level I hospitals, an increase in the number of outpatient visits relative to inpatient days is likely to result in an increase in overall efficiency levels. A one percentage point increase in the ratio of outpatients to inpatients would lead to about 15 percentage points increase in the tendency to be technically inefficient. The provincial location of a hospital (*prov1*, *prov2*) has no significant bearing on efficiency.

8.3.8. Productivity growth

The interval of time used includes a period of structural transformation, that is a structural break, in South African history. It encompasses both the apartheid and post-apartheid periods. Over the sample period, total factor productivity (TFP) dropped by 12.1 per cent. As can be observed from Table 8.7, this is largely due to a decline in technical progress. The drop in technical efficiency is marginal. Technical efficiency increased in the two immediate years after 1994/95.

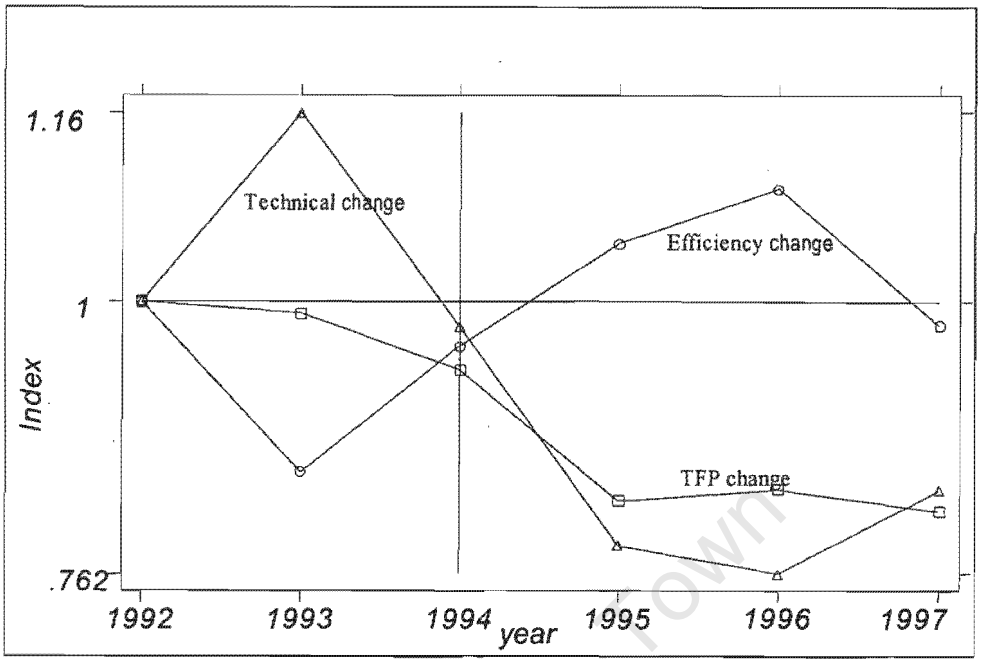
TABLE 8.7
MPI SUMMARY OF ANNUAL MEANS

Year	Efficiency change	Technical change	TFP change
1992/93[1]	1	1	1
1993/94	0.851	1.164	0.990
1994/95	0.961	0.978	0.940
1995/96	1.050	0.787	0.826
1996/97	1.098	0.762	0.836
1997/98	0.979	0.835	0.817
Mean	**0.984**	**0.893**	**0.879**

Over the years 1992/93-1997/98, technical efficiency dropped by an average of 2.1 percent, as opposed to a 16.5 percent decrease in technological growth. Efficiency change and technical change are observed to move in opposite directions (Figure 8.5).

1 1992/93 is the base year.

FIGURE 8.5
PRODUCTIVITY CHANGE, 1992/93-1997/98



8.4. DISCUSSION

Technical inefficiency in health facilities is present in varying degrees in both developed and developing countries (see for example Wouters 1993, McMurchy 1996, Ersoy *et al* 1997, Ferrier and Valdmanis 1996, Hao and Pegles 1994, Ozcan *et al* 1996, Rosko and Chilingerian 1999). However, in sub-Saharan Africa, not many hospital efficiency studies have been conducted using frontier models as evidenced by the dearth of published literature. Thus, there is no clear and quantifiable evidence on the type and degree of health service in efficiency in the African context.

The results presented here suggest that significant numbers of the hospitals included in this study operate at technical efficiency levels well below the efficient frontier. The level of technical inefficiency, which is in the range of 35 to 47 percent, indicates wastage of significant amounts of health resources. If the efficient hospitals were to operate on the best-practice frontier, there would be an immense saving in terms of all key resources.

Given the tight fiscal constraints and resulting stagnant real per capita health budgets in South Africa, extending and improving the quality of primary health care services has to be funded through health service efficiency gains (particularly in hospitals as they account for the vast majority of expenditure) and/or increased health service revenue from non-tax sources. At present, the main source of such revenue is that of user fees at public sector hospitals. In 1992/93, fee revenue was equivalent to approximately 9 percent of public sector hospitals' recurrent expenditure (McIntyre *et al.* 1995) and fee revenue has declined dramatically since then largely due to the exodus of medical scheme members from public hospitals. The potential efficiency savings estimated in this study amount to more than three times that of the fee revenue collected. Thus, very high levels of user fees would be required to generate revenues that could match the potential efficiency savings. Given the limited health insurance coverage in South Africa, substantial increases in public hospital user fees are likely to negatively impact on equity.

How could these efficiency savings be achieved? Bed-sizes in all three groups of hospitals appear to exceed what is required for the given levels of outputs. However, a caveat is in order here. The finding that the number of beds could be reduced to produce the same output at lower cost does not imply that the existing number of beds exceeds the population's potential need for hospital services. The South African public sector hospital bed to population ratio is in fact relatively low by international standards (McIntyre *et al.* 1995). The utilisation of the existing hospital facilities, however, depends not only on supply-side factors, but also on factors related to demand. Therefore, it can safely be argued that, given current demand levels, the existing number of beds is in excess of what is required for efficient inpatient service provision. The results of a number of hospital efficiency studies in other countries indicate that redundancy of beds is a common problem (Ozcan *et al.* 1996, Brownell and Roos 1995).

Bed closures should particularly be considered in those hospitals that experience decreasing returns to scale (about half of the hospitals in this study). Decreasing returns to scale implies that a hospital has an inefficiently large size. To curb its inflated unit costs, the hospital needs to scale down its size (beds and staff). Pursuing this option is likely to promote the government's moves towards re-allocation of resources away from hospital-based services to cost-effective non-hospital services, in line with the primary health care strategy.

In contrast, increasing returns to scale is observed in about 37 percent of the hospitals evaluated. In the presence of increasing returns to scale, expansion of outputs reduces unit costs. A hospital with increasing returns to scale will, therefore, benefit by augmenting its scale of operations. However, increasing the level of outputs requires an increase in demand, which is beyond the hospital management's control. Merger of hospitals that are in close proximity to one another may be an option worth consideration. This option may, however, pose some problems, especially in sparsely populated rural areas. If a few hospitals of a bigger size are to be established in centrally located places, residents of such areas may incur additional costs in travel expenditure and in delayed treatment of emergency cases. These problems may to some extent be minimised by establishing primary care units that have a link with the centrally located hospitals through an effective referral and patient transport system. It should, however, be emphasised that any initiatives undertaken to reap economies of scale must be implemented only after careful appraisal of the circumstances surrounding the operation of the hospital(s) under consideration, as well as the potential equity implications of merging existing hospitals.

The estimated Tobit model indicates that the tendency for technical inefficiency significantly decreases with an increase in the occupancy rate. This finding corroborates that of Ferrier and Valdmanis (1996). As health managers most frequently use the

occupancy rate to evaluate hospital performance, its significant positive association with the DEA-based measure of technical efficiency is reassuring.

The number of outpatient visits as a proportion of inpatient days has a significantly positive impact on efficiency. This may indicate the presence of scope economies between outpatient and inpatient care. In the presence of economies of scale, it is cheaper to produce both inpatient days and outpatient visits together rather than separately. Since the outpatient activity of Level I hospitals is limited, increasing the activities of the outpatient's department is likely to promote technical efficiency. This finding is of great importance to policy makers as it suggests that rather than constructing new clinics near existing hospitals, increased use of district hospital (Level I) outpatient departments for primary ambulatory care will improve overall health sector efficiency.

The decrease in total factor productivity, which is largely due to technical regress, is worrying, as it has overshadowed the modest growth in efficiency that was seen in the period after 1994/95. Thus, the emphasis should be on addressing scale and scope inefficiencies in hospitals to facilitate level of care redistribution while preventing further technological decrements and concomitant productivity losses. The decline in productivity observed in this small sample of hospitals indicates the need for an extensive assessment of the hospital system as a whole. If productivity losses of this magnitude are found in the system at large, all efforts of the government to redress past injustices and increase access to services of acceptable quality will be jeopardised.

This study contributes to and strengthens the existing literature on the technical efficiency and productivity of hospitals in South Africa by using more robust techniques of measurement based on micro-economic theory. Previous studies of hospital efficiency in South Africa have used simple ratio analyses, which in most instances do not give a comprehensive picture of the magnitude and sources of inefficiency. Furthermore, previous

studies have not assessed changes in total productivity and its components - efficiency and technological changes. The study gives a clear quantitative value of the magnitude of technical inefficiency (35-47 percent) that is prevalent in South African hospitals. It also provides an evidence base on the extent of scale inefficiency which forms the basis for either down-sizing or merger of hospitals. The total factor productivity of hospitals decomposed into efficiency and technological change shows that there was a decline in TFP, which is largely attributed to technical regress. This is a new contribution of the study within the South African context.

CHAPTER 9
SUMMARY, CONCLUSION AND RECOMMENDATIONS

This study has explored the status of the various aspects of equity and efficiency within the context of the South African health system. It has empirically assessed the *status quo* and trends in equity with respect to morbidity and mortality in early life and morbidity and utilization of care in adulthood. This helps capture the state of equity during the different stages in the life-course. In a country besieged by extensive, systematic social inequalities, slow or absent economic growth and negative effects of economic restructuring plans, redressing the equity backlog through mobilizing additional resources may be very difficult. In such a scenario, it is of paramount importance to improve the efficiency (both technical and allocative) of the health system and plough back the efficiency savings towards meeting the resource gap in the fight against inequity. This, however, needs a thorough examination of the state of efficiency and trends over time. To this end an attempt is made to examine the state of technical efficiency and productivity of a sample public sector hospitals to demonstrate the magnitude of inefficiency and the possible efficiency gains that may be reaped to bridge the existing equity gap. Hospitals are targeted in this case, as they are the largest consumers of health sector resources.

South Africa has emerged from decades of apartheid rule. The backlog of inequity inherited from a system, in which race played a major role in determining access to health and health enhancing resources, is tremendous. The mammoth task that lies ahead requires the formulation of appropriate resource reallocation policies and plans. In turn, the formulation of effective policies and interventions to redress inequities requires an empirical evidence base that would help to objectively assess the magnitude, trends and determinants of equity, efficiency and productivity. Policies and interventions that are not founded on a strong

evidence base are likely to lead to misguided conclusions with regard to the potential effectiveness of government policies and interventions and their proper targeting.

The study of equity is divided into three sections:

1. Equity in infant and under-five mortality between two time periods (1993 and 1998), which is meant to capture equity in survival in early childhood. It is a well-established fact that mortality in early life is influenced by a host of factors many of which may be outside the health sector. Furthermore, mortality in early life is taken as one of the many indicators of a country's level of development. Thus, by assessing inequities in infant and under-five mortality rates, it is purported to demonstrate inequalities in these sensitive indicators during the early years of the life course.

2. Equity in child malnutrition, which is aimed at capturing inequalities in child health. It may be the case that even if children escape death (and thus apparently the absence of income-related inequalities in infant and under-five mortality) their quality of life may be seriously undermined and may be experiencing sub-optimal health. Thus by including child malnutrition an attempt is made to consider their quality of life. Child nutritional status is one of the most sensitive indicators of child health.

3. Equity in self-reported adult illness, which is meant to provide insights into inequalities that take place during adulthood. Adulthood in this case is operationally defined as inclusive of all aged 18 years and above. This will give a picture of the state of equity that may occur in the later years of life. Moreover, equity in self-reported illness is also related to utilization of health services in an attempt to test for horizontal inequities, that is whether people in equal need get the same treatment or not. This is done by examining

utilization/care-seeking behaviour that is conditional on reporting sickness. Utilization of services is further disaggregated by provider type including traditional healers.

Inequities in health and health care utilization have different dimensions and manifestations at the various stages of the life span of individuals. Hence assessing socio-economic inequalities in health (both morbidity and mortality) during childhood as well as the adulthood period becomes essential.

The equity analyses are followed by consideration of efficiency issues, with particular focus on the technical aspect of efficiency and some of its determinants. This is an attempt to examine if there are substantial inefficiencies, which in the end are likely to perpetuate inequities. Moreover, since the country has seen a change of government in 1994, panel data are used to assess productivity. The change in productivity is decomposed into two components: efficiency change and technical change. Other things being constant, a positive efficiency change implies a growth in productivity resulting from changes in technical efficiency. On the other hand, a positive technical change may imply growth in productivity emanating from changes in the production technology.

Overall, the findings of this study indicate that the huge socio-economic inequalities in health and health care that existed prior to the change of government in 1994 have been reduced significantly in the years that followed the installation of the new government. This signifies that the efforts of the government to improve the health and welfare of the poorest, who sustained the greatest burden of morbidity and mortality as a result of the exclusionary policies of the pre-1994 political regime, have been fruitful. This cannot be overemphasized, given that a government that upholds the interests of the majority was installed and a number of steps were taken to improve access to health and other health-enhancing resources (such

as the free health care for children under six-years of age, the Primary School Nutrition Programme, the aggressive campaign of clinic building and the Cuban Doctor programme, to mention but a few).

Analyses of the 1993 LSDS data indicate significant inequities that favour the rich in all the dimensions of equity in health and health care utilization. The findings also reveal the impoverishing effect of inequities in health and health care utilization among the poor. These include prolonged days of sickness, physical barriers to access as seen by the longer time needed to reach a health facility and relatively more use of pharmacy services, which in most instances entail out-of-pocket payments.

The "race" (population group) factor seems to play a prominent role in inequalities in under-five mortality and malnutrition. Inequalities, especially in under-five mortality that are not even evident when household income (proxied by expenditure) is used, emerge when population group is used as an indicator of socio-economic status. The two groups, Africans and Coloureds, that are considered as disadvantaged in the South African context exhibit relatively higher rates of under-five mortality and malnutrition (stunting and underweight). Thus, even if it may sound unpalatable, in targeting resources to promote society's welfare it is necessary to use "population group" (being African and Coloured) as a proxy indicator of socio-economic disadvantage. Rural areas also have significantly higher rates of under-five mortality and malnutrition compared to urban settings. Furthermore, Gauteng and the Western Cape provinces, which may be regarded as the economic hubs of the country have consistently, lower rates of under-five mortality and malnutrition. This implies that rural and to a certain degree provincial locations could also be used to identify the disadvantaged.

With respect to socio-economic inequalities in stunting and underweight, there is a consistent trend for the highest pro-rich inequalities to be seen in Coloured children, metropolitan areas, and provinces with the lowest rates of malnutrition (Gauteng and the Western Cape). This indicates that higher pro-rich inequalities are found coupled with relatively lower average rates of malnutrition. Hence, policies need to take account of the presence of pockets of high levels of malnutrition in metropolitan and other urban areas, in Coloured children and the two provinces with lower average rates, so as not to commit errors of omission in targeting.

Aggregate level of poverty in the residential province seems to highly influence under-five mortality. Therefore policies geared towards avoiding socio-economic inequalities in under-five mortality also need to see beyond the health sector and have a perspective that places poverty reduction at the centre.

High degrees of pro-rich inequality in stunting and underweight are witnessed in this analysis. However, since the data refer to the period before 1994, it is necessary to conduct a household survey of the LSDS type with measures of nutritional status and extensive household socio-economic data to assess changes since 1994. Since the more recent Demographic and Health Survey data do not include household income/expenditure status, it is not possible to assess changes that have taken place after 1994.

The models on the determinants of malnutrition have indicated that household income and education of the mother have a negative effect on malnutrition. Increasing the household income and educational levels decrease the probabilities of chronic malnutrition. Thus, the fight against inequities in child health has to be multi-pronged and address simultaneously the factors that contribute to perpetuating inequity. In this regard augmenting household income and mother's education should be given prominence in the policy agenda.

The analyses on self-reported adult illness and utilization of services indicate changes in many of the measures used in 1995 and 1998 compared to the situation that existed in 1993. For example, the paradoxical pro-poor inequalities in self-reported illness seen in 1993 are turned pro-rich in 1995 and 1998. This change, which conforms to our intuition, may possibly be an indication of the radically changed political climate, which among other things saw the formulation and implementation of a democratic constitution, relevant policies and a Bill of Rights, that may have raised the awareness of the poorest regarding health matters and their entitlements to health care. It may also indicate increased access to health care seen after the installation of the new government.

The utilization data indicate that in all the three periods (1993, 1995 and 1998) there are pro-poor horizontal inequities in the utilization of primary and other public health facilities. This implies an appropriate targeting of public sector health care resources. **Although still pro-poor, the** magnitude of the horizontal inequity has diminished in 1995 and 1998 compared to the situation in 1993. A declining trend may, in the end, result in leakages of government subsidies to the non-poor, and thus efforts need to be intensified to reinforce the 1993 pro-poor horizontal inequities in the use of public sector health facilities.

Given the highly segmented nature of the South African health system that has a relatively large private sector characterised by declining population coverage and rapidly escalating costs (Cornell *et al* 2001), a growing proportion of the population will rely on the public sector for health care. The increase in the proportion of the population resorting to the public sector (particularly in relation to hospital care) requires that appropriate targeting techniques be put in place to avoid excessive leakages of the benefits to the non-poor or under-coverage and jeopardise the equity objectives of the system. Special attention should be given to updating targeting mechanisms regularly, in order to check for under-coverage of those who are

dropping out from the private sector because of inability to pay and/or possible adverse selection. The identification of appropriate targeting mechanisms is a very important component of studies of equity in health care that falls beyond the scope of the present study. It is therefore recommended that this vital component be rigorously investigated in future research.

Overall, although there are promising improvements in equity in health and health care in South Africa as measured by under-five mortality and malnutrition and adult illness and utilization of services, major equity challenges still remain. As seen in the present analyses, inequities that are masked when using certain measures of socio-economic status and ill-health, may be prominent when other measures of socio-economic status and health are used. Hence there is a need for extensive studies of inequities, including inequities in the most common problems of the country. It is also seen that redressing inequities in health is not only the sole responsibility of the health sector – a point that was also endorsed by the Alma Ata Declaration (Primary Health Care) in 1978. For example, some inequities in health may need augmentation of household income or improving levels of literacy and education. These issues transcend the health sector and need a concerted multisectoral approach.

As stated in the introduction to this chapter, South Africa is not only grappling with a huge backlog of inequities, but also an unpromising macro-economic performance and some of the untoward effects of economic restructuring plans which, among other things, entail cutbacks in public spending. The HIV/AIDS epidemic is also taking its toll in terms of the consumption of scarce health care resources. In this scenario, injection of significant additional resources to the health sector from the public purse may be very difficult. It is, hence, necessary to be able to maximize benefits from whatever scarce resources are already devoted to this system. In other words, efficiency in production and allocation will play a synergistic role in releasing

badly needed resources that can be used to improve access to health care and health-enhancing interventions. It is with this intention that this study attempts to examine the technical efficiency and productivity of a sample of hospitals in three provinces. The focus is on hospitals, as they absorb the lion's share of the resources available for health care, and it is assumed that efficiency savings from these institutions can make a perceptible impact in terms of releasing resources.

The study suggests the potential for immense savings in resources from the sample of hospitals analysed. The hospitals could have produced their output levels in terms of outpatient visits and inpatient days with ½ to 2/3 of the resources that they consumed. This implies savings ranging from about 1/3 to ½ of their actual resource endowment. The efficiency savings are by far higher than what hospitals in South Africa were able to collect as fee revenue from patients.

The above levels of efficiency savings can impact on equity in a number of ways. First, the fact that the savings surpass fee revenues by a big margin suggests that user fees for hospitals can be set at levels that are affordable to the poorest. Additionally, the application of exemption criteria that are meant to minimize leakages of the benefits to those who don't qualify (errors of commission) need not be too strenuously imposed due to their adverse impact on equity. Second, with the levels of efficiency savings from these hospitals it is possible to construct and operate the much needed primary care clinics that would improve access of the poorest and/or improve quality of care. If the findings from this sample of hospitals could be extrapolated to the whole population of hospitals in the country, then the resource savings will really go a long way in addressing equity issues.

Significant numbers of the hospitals suffer not only from pure technical inefficiency, but also inefficiencies emanating from a non-optimal scale. About half of the studied hospitals experience decreasing returns to scale. The usual therapeutic prescription for this malady would be downsizing, as large size is contributing to their inefficiency. This move may also promote the reallocation of resources away from the chronic problem of hospital bias that is a typical feature of the health systems of developing countries. The process of downsizing may, however, lead to more efficiency gains if measures are taken to bolster public-private partnership. Although the closure of beds may bring about savings in terms of personnel costs, the efficiency savings may be neutralised if the already existing physical infrastructure remains idle. Given the increasing presence of private health care providers in rural areas (Soderlund *et al* 2001), leasing beds/wards vacated through downscaling to the private sector is an option that has to be explored. This will help the public sector generate additional revenue in addition to the cost savings. Moreover, from a macro-level perspective, the rational use of existing hospital resources will help society free its scarce resources for other pressing needs rather than duplicating investments in hospital infrastructure which in the end lead to sub-optimal capacity utilisation.

On the other hand, substantial numbers of the hospitals (a little more than a third) suffer from another form of problem due to size. They experience increasing returns to scale, implying that they need to scale up their operation. Local departments of health, however, need to scrutinize the situation on the ground before making this move. Hospital outputs cannot be stored for future use, and therefore increasing the supply when the demand does not exist would mean introducing more inefficiency. The preferred option to consider here is the merger of hospitals that serve catchment areas that are contiguous. The merger of hospitals should, however, be complemented by primary care facilities that are linked to the large (merged) hospital through an efficient referral and patient transportation system.

Economies of scope seem to be present between the two output lines in district hospitals: outpatient visits and inpatient days. This implies that the joint production of the two outputs remedies the inefficiency that results from a single product line technology. This signals to policy makers to consider using the outpatient departments of existing hospitals fully for ambulatory care, rather than constructing new clinics near underutilized hospital facilities in the name of upholding the Primary health Care strategy.

The period 1992-98 has seen a drop in total factor productivity in the sample of hospitals, which was due to technical regress. The arrested investment in technology overshadowed modest improvements in efficiency. An important implication of this finding is that in channelling resources away from hospitals, which historically have been consuming huge amounts of resources, consideration of the complementarity of inputs is essential. For example, it would lead to a decrease in labour productivity to slash the budget used to procure consumable items that are needed to operate high-technology diagnostic/therapeutic equipment that for one reason or another has been installed, as the technical mobility of labour is very minimal. Thus, budget decrements should always take into account alternate uses of the technology and human resources trained to operate it. Sudden reductions in budget without due consideration of the effects will simply lead to reduced productivity, which in the end implies wastage of scarce resources. Thus, it is necessary for hospitals to earmark some of their budget for ensuring that there is adequate maintenance and improvement of technology so as not to compromise total factor productivity. Alternatively, in the presence of serious resource constraints to continually buy new technology, it is essential to strengthen partnerships with the private sector. By doing so, agreement may be reached for government doctors to use the private sector technology for their patients for certain hours or the private sector may be sub-contracted to provide certain high-technology diagnostic or therapeutic

interventions. Again this has the effect of maximising the utilisation of scarce societal resources, and augmenting efficiency gains.

The following policy recommendations are in order from the foregoing discussion:

- Even if not politically desirable, race is an important predictor of disadvantage in South Africa. Socio-economic inequalities in morbidity and mortality that do not manifest when household income/expenditure are used as proxy for socio-economic status, appear prominently when race is used as a measure of socio-economic status. Therefore, the variable race/population group needs to be used as a measure of socio-economic status along with household income and other relevant attributes for at least the foreseeable future.

- Although rural areas have higher average rates of disease burden, pro-rich inequalities (particularly in stunting and underweight) are much higher in metropolitan and other urban areas. Thus, reliance on average rates of stunting and underweight would lead to errors of targeting, for the focus will only go to rural children. Therefore, it is imperative that in formulating policies to address the problem of malnutrition, policy makers need to take account of measures of inequality along with measures of average prevalence.

- Highest pro-rich inequalities in chronic under-five child malnutrition are found in the provinces with lower average rates of malnutrition, which are also regarded as relatively rich provinces. In contrast, poorer provinces with relatively higher levels of malnutrition have much lower pro-rich inequalities. There are many children in the two provinces that are regarded as well off whose lot is no better than their counterparts in the relatively poorer provinces. Therefore, policies should also focus on the poorest children in the richer provinces so as to address the problem effectively.

- To reduce socio-economic inequalities in health, policies that transcend the health sector need to be designed and implemented. Poverty reduction and augmentation of the incomes of the poorest as well as promoting the education of mothers are among those that deserve special mention.
- Although there are pro-poor horizontal inequities in the use of public health care facilities, the magnitude of this pro-poor inequity has been declining. Therefore, efforts need to be intensified to prevent a situation that may result in a leakage of government subsidies to the non-poor.

- The level of technical inefficiency in the sample of hospitals is so alarming that it is likely to impede initiatives to address inequities. Therefore, there is a need to undertake further studies that compare the characteristics of the efficient and non-efficient hospitals, so as to design measures that enhance the efficiency of hospitals and, ultimately, the health system at large.

- As significant numbers of the hospitals studied are operating at a non-optimal scale, there is a need to take appropriate measures to enhance efficiency. These include measures such as downsizing and merger, which should be implemented after a careful assessment of the situation on the ground (e.g. equity considerations).

- To promote technical efficiency, government should use outpatient departments of existing hospitals fully. Constructing clinics in the vicinity of under-utilized hospitals will simply aggravate the existing inefficiency.

- Technical regress has been a major cause of the decline in productivity over the years included in the study. The move to channel resources away from hospitals should

always take into account input complementarity so that the high technology equipment and highly specialized human resources don't remain underutilised (and thus become an additional source of inefficiency and loss of productivity).

In summary, the study has shown the existence of pro-rich inequities in most measures of health and health care utilization. Inequities among different population groups, residential locations and provinces are also noted. On the other hand, examination of the second component of the twin objectives of health policy also indicates the prevalence of high levels of technical inefficiency and decline in productivity emanating from technical regress. We therefore have a situation where both objectives are compromised.

In such a scenario, it would be very difficult to address the inequity backlogs effectively, as the resource base towards this end is gravely constrained. Technical inefficiency and poor productivity would in effect lead to further shrinkage of resources. On the assumption that the role of health care is to improve health and reduce inequalities in health, the inefficient delivery of health care will therefore impact on equity negatively.

Finally, the study by assessing inequities both in child- and adult-hood using robust inequality measurements adds new insights into policy debates in the area of equity in health and health care. Furthermore, given the tight fiscal constraints to mobilization of additional resources, it empirically illustrates the presence of huge amounts of efficiency savings from within the health system that can go a long way in redressing inequities without having to compromise quality of care.

APPENDICES

APPENDIX1

Philosophical foundations of equity

Theory	Description	Health policy implication
Utilitarianism	Associated with the work of Jeremy Bentham (1789), its principle can be summarized as serving the greatest good for the greatest number. Utilitarianism revolves around three principles (Sen 1973): (i) welfarism – the goodness of a state of affairs is judged only in terms of utility. Other information (e.g. individual needs) are regarded as irrelevant or indirectly relevant; (ii) sum ranking – the goodness of a collection of utilities is simply their sum, without due concern about inequalities in their distribution; and (iii) consequentialism – all choice variables are judged only by the goodness of their outcomes.	The principle of consequentialism is in line with the requirement for effective health interventions. It excludes the provision of services that do not produce the desired results; hence can be defended in health policy. The first two principles, however, are not appropriate for the formulation of sound health equity policies, as they have no concern of distributional issues and needs. As utilitarianism is concerned with maximizing aggregate welfare, it looks more efficiency- rather than equity-enhancing principle.
Egalitarianism	Involves equalizing individual net benefits or opportunities for such benefits, if the benefits cannot physically be distributed (e.g. health status) (Pereira 1993). It has two aspects: equality of welfare and equality of resources.	Its consequentialist aspect is unlikely to promote welfare in health. For example, it will regard an equally bad state of health preferable to a state in which one is in the same bad situation and the other one is in a good health condition. It lacks specificity and universal acceptability. For e.g., does equality of welfare imply equal levels of health? Does equality of resources imply the use of equal amounts of resources, or equal opportunity of access? Should these definitions be applied to public or private health care or both?
Rawlsian Maximin	Also called the *difference* principle is due to John Rawls (1971). It states that behind a veil of ignorance and in their original position, individuals will choose to maximize the primary goods for the least advantaged.	It has an appeal for it takes account of the severity of illness of the worst off. Its flaw, however, is that need is equated with severity of illness, without regard to capacity to benefit. Thus it may result in depletion of limited resources for a very marginal improvement of patients with terminal illness (Mooney & McGuire 1987, Olsen 1997).
Entitlement	This theory, which is due to Nozick (1974), is based on the premise that people are entitled to what they possess, if those possessions are acquired in a just way. It focuses on processes as against outcomes.	It is not defensible in health equity policy, as it attaches no weight to the state of the unfortunate (e.g. those born with congenital anomalies). Moreover, if taken as *entitlement to access*, citizens are only entitled access to health care that is acquired through the market (considers taxation as one form of theft & slavery). Government intervention to increase access to those who use health care inefficiently (e.g. due to poverty, low level of education *etc*), or because of the existence of externalities & public goods is regarded as injustice. Thus it would lead to a distribution, which is inequitable and unfavourable to the poor & the sick.

Decent minimum	Proposes that individuals should not fall below a certain standard, i.e. the *decent minimum*; This implies the setting of a safety net to rectify unfavourable outcomes that are likely to arise from Nozick's theory.	It implies the provision of a minimum standard of health care, given scarcity of resources. Emphasis is laid on the private sector; the government just providing a limited level of health care for the poor. It entails a subjective judgment of what constitutes the decent minimum. There are no explicit and indisputable criteria to explain why some health services are to be excluded from the minimum package of services. Also what is accepted as a minimum package of services may vary temporally with changing circumstances
Envy-free allocations	Proposed by Tinbergen (1953) & Foley (1967), it focuses on distributive justice. According to envy-free criterion, an allocation is regarded as equitable if no individual prefers any other individual's situation to his own (Varian 1974, Baumol 1986). It has its roots in the age-old conception of "how to cut a cake fairly" (LeGrand 1984).	It does not seem to be of use in guiding health equity policies. There could be situations that are inequitable and envy-free, or alternatively, equitable situations where envy exists. Even if it is applied in the health arena despite its shortcomings, we can only consider health care, as health itself is indivisible. It does not also provide necessary information for a complete ranking of alternative states, which is avery important condition in the health domain (Pereira 1993).
Communitarianism	This theory emerged in the 1970s. Individuals have two types of utilities (Mooney 1996): normal goods utility derived from outcomes and utility derived from doing rather than getting (Margolis (1992) calls this *participation utility*). Participation utility underpins communitarian claims. The principle states that individuals obtain satisfaction or increased well-being from actively participating and contributing to the well-being of their society. The participation utility of Margolis is akin to Sen's *istrumental agency success* – an individual gets utility from being instrumental in achieving equal access for all to health care.	According to this approach the public is consulted to give its values about health care priorities. Community's preferences and the weights that it attaches to health gains by the various population groups in society are given serious consideration. Mooney (1998) used this approach in Australia. The approach can only promote equity to the extent that the community from which the ethical values are drawn is benevolent and with good judgments. However, as witnessed in history, there are times when communities can be malevolent (e.g. the community of Nazi Germany), thus detracting from the merits of this approach.

APPENDIX 2

SOME QUANTITATIVE METHODS FOR HEALTH INEQUALITY MEASUREMENT

1. INTRODUCTION

In chapter 3, only the methods that are used in this study in assessing inequalities in health have been discussed. The purpose of this Appendix is to present some of the commonly used quantitative techniques from the repertoire of measurements of inequalities in health and health care. Several methods have been in use to date. Some have their origin in research on income inequality (e.g. Lorenz curve and the associated gini coefficient) (Atkinson 1970, Vagero and Lundberg 1989) or from modifications of these (e.g. concentration index) (Wagstaff *et al* 1989). Other methods are based on measures of association (e.g. index of dissimilarity, slope index of inequality) (Manor *et al* 1997).

2. REGRESSION-BASED MEASURES

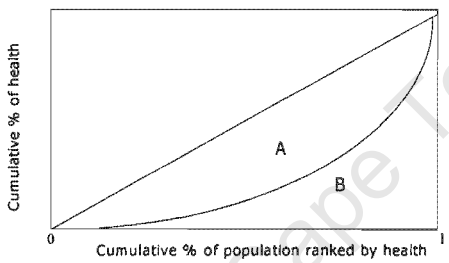
Most of these measures employ econometric models appropriate for categorical dependent variables such as the logit and probit models. The health variable (morbidity/mortality, *etc*) is taken as the dependent variable and regressed against explanatory variables representing socio-economic status.

3. THE LORENZ CURVE AND THE GINI COEFFICIENT

The Gini coefficient has been widely used for assessing economic inequalities since 1912 (Sen 1972). Its use in measuring inequalities in health has been a recent phenomenon. Although various authors have suggested using it, many have not employed the true Lorenz curve (Wagstaff *et al* 1991).

The Lorenz curve plots the cumulative proportions of the population ranked according to the severity of illness beginning from the sickest person and ending with the healthiest, against the cumulative proportions of health. It has to be noted that the Gini coefficient does not measure socio-economic inequalities in health. It only measures the total amount of inequalities in health between individuals in the population. The figure below depicts the Lorenz curve.

Figure A2.1
Health Lorenz curve



The gini coefficient is derived from the Lorenz curve. It is the area between the diagonal and the Lorenz curve (*A*) expressed as a proportion of the total area below the diagonal (*A*+*B*). It ranges from zero (complete equality- the Lorenze curve coinciding with the diagonal) to one (complete inequality). The further the distance of the Lorenz curve from the diagonal, the higher the degree of inequality (moves towards one). The Gini coefficient is exactly one-half of the relative mean difference, that is the mean of the absolute values of differences between all pairs of income or any other variable being evaluated (Sen 1973). It may be computed using the following formula (*ibid*):

$$G = 1 + \left(\frac{1}{n}\right) - \left(\frac{2}{n^2 \mu}\right)\left[y_1 + 2y_2 + \ldots + ny_n\right] \qquad \text{(A2.1)}$$

$$\text{for } y_1 \geq y_2 \geq \ldots \geq y_n$$

The Gini coefficient avoids one of the problems associated with `range' (rate ratio) measures, that is it takes into consideration the health condition of all individuals rather than limiting itself to any two extreme groups (such as the least versus the most well-off). It, however, does not take account of socio-economic status. Distribution of health where the mean level is maintained but the health of the sickest improves and that of the healthiest deteriorates is registered as a reduction in inequality.

4. THE RELATIVE AND SLOPE INDICES OF INEQUALITY

The relative index of inequality (RII), which is commonly used by epidemiologists, is a regression-based technique and is equivalent to the concentration index of Wagstaff *et al* (1991). Like the concentration index, it takes into account the population size and the relative socio-economic position of groups.

The RII is calculated in two stages. First, the SES of each group is quantified. This can be done by assigning them a relative position in the social hierarchy. For example, if the highest social group comprises 10 percent of the population, the relative position of its members is between 0 and 0.1, which on average is 0.05. if the next highest group is also 10 percent of the population, the relative position of its members is between 0.1 and 0.2, which on average will be 0.15. In the second stage, the SES measure is related to the prevalence of a health problem (morbidity/mortality) by means of a regression analysis.

The resulting figure (RII) represents the proportional increase in morbidity/mortality per unit increase in the SES measure. A one-unit increase in the SES measure is equivalent to the difference between the top (0) and the bottom (1) of the social hierarchy. Therefore, the RII represents the proportional increase in morbidity/mortality by moving from top to the bottom

of the hierarchy, taking into account intermediate points. A large value of the RII indicates large differences in morbidity/mortality the high and low positions in the social hierarchy.

In the second stage of RII calculation, different regression techniques have been used. For example, Kunst and Mackenbach (1994) used a poisson regression, Kunst *et al* (1995) used the logistic model. Wagstaff *et al* (1991) advise against the use of the method of Ordinary Least Squares (OLS), as the likelihood of heteroskedasticity in the grouped data is high. They recommend the use of the method of Weighted Least Squares (WLS).

The slope index of inequality is a variant of the RII. It measures health inequality between the top and bottom of the social hierarchy in terms of absolute differences between rates rather than rate ratios.

5. INDEX OF DISSIMILARITY (ID)

The index of dissimilarity indicates the percentage of all cases (e.g. ill individuals or deaths) that need to be redistributed in order to obtain a uniform morbidity/mortality rate for all socio-economic groups. Larger values of the ID mean larger degrees of inequality.

If there are $j = 1,...,J$ socio-economic groups, the ID is given as:

$$ID = \frac{1}{2}\sum\left|S_{jh} - S_{jp}\right| \qquad \text{(A2.2)}$$

where S_{jh} is the j^{th} group share of the population's health, and S_{jp} represents the j^{th} group population share.

As can be discerned from the above, the ID does not take into account the ordered nature of socio-economic position. This is a notable shortcoming of this method.

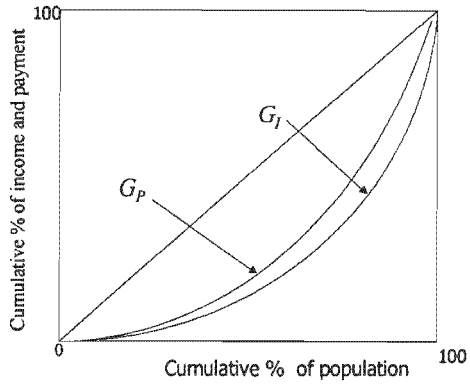
6. KAKWANI INDEX

The Kakwani index is used in assessing equity in health care financing. It measures the progressivity of payments for health care. The progressivity of a health care financing system refers to the relation between income and payments for health care as a proportion of income.

A health care financing system may be progressive, regressive or proportional. In a progressive system, health care payments increase as a proportion of income with increasing incomes. If payments fall as a proportion of income when income is rising, the system is regressive, and if payments account for the same proportion of income for everyone, regardless of one's income, it is called a proportional system.

The Kakwani index not only identifies progressivity, but it also measures the degreeof progressivity. It is based on the extent to which a health financing system deviates from proportionality. The essence of the Kakwani index can be seen from the following figure.

Figure A2.2
Illustration of the Kakwani index



The curve labeled G_I is the Lorenz curve for income. The second curve G_P the concentration curve for payments for health care. In a proportional system of health financing, G_I and G_P coincide. In a progressive system, G_I lies above G_P. The opposite holds true in the case of a regressive system. The degree of progressivity is assessed by looking at the size of the area between G_P and G_I. Thus the Kakwani index (π_k) for health care payment is defined as:

$$\pi_k = G_P - G_I \qquad \text{(A2.3).}$$

This is twice the area between G_I and G_P. The value of π_k ranges from +1.0 (when all income is distributed equally and the burden of payment falls on one person) to -2.0 (when all income is in the hands of one person and someone else has the burden of paying for health care). Thus, π_k assumes positive (negative) values when the system is progressive (regressive), and equals zero in a proportional system.

7. POPULATION ATTRIBUTABLE RISK (PAR)

Although used extensively in the field of epidemiology, its use in the measurement of inequalities in health is fairly recent. PAR indicates the proportional reduction in overall rates of morbidity and mortality if it is assumed that everyone experiences the rates of those with the highest position in the social hierarchy. It is defined as:

$$PAR = 1 - \frac{R_H}{R_T} \qquad \text{(A2.4).}$$

Where,

R_H = the rate in the group with the highest social position; and

R_T = the rate in the total population.

A rate of, for example, 24 percent implies that the overall rate of morbidity/mortality will be reduced by about 24 percent if all persons were to experience the rate of those in the top social position.

It has to be noted that simplicity in the calculation of the PAR as given above is achieved by ignoring the link between socio-economic position and morbidity/mortality in the groups below the one with the highest social position. To avoid this problem, a regression-based PAR can be calculated (Mackenbach and Kunst 1997). This is done in two stages. First, the regression based effect index is calculated, and in the second stage the PAR is calculated. The difference from the simple PAR is that in this case, the reference rate is not the observed rate of the group with the highest social position, but the rate predicted for some high SES value estimated from the fitted regression equation. For example if the predicted rate of morbidity for people with 12 years of education is 28 percent and the overall rate in the population is 40, the PAR will be equal to 30 percent. This indicates that the overall rate of morbidity will

decrease by 30 percent if everyone experiences the rate that corresponds to that of people with 12 years of education.

APPENDIX 3
HOSPITAL EFFICIENCY ANALYSIS USING RATIOS

As discussed in chapter 7 and illustrated in Figure 7.1, the efficiency of hospitals may be assessed using robust frontier techniques (parametric and non-parametric techniques) or using simple ratios. Although ratios may not be robust in measuring efficiency and productivity, their ease of calculation and comprehension makes them more widely used at the operational level. This annex summarizes some of the ratios, which are also referred to as performance indicators.

1. Average length of stay (ALS)

This measure refers to the average number of days that a patient stays in a hospital. It is calculated using the following formula:

$$ALS = \frac{inpatient\ days}{admissions} \qquad \text{(A3.1)}$$

With the assumption that comparison is done within a homogeneous group of hospitals with a similar case-mix, hospitals with lower ALS are regarded as performing well relative to their counterparts with higher ALS. Primary-level hospitals are expected to have a shorter ALS compared to tertiary-level hospitals. If the ALS in tertiary hospitals is lower than that of primary level ones, it may possibly mean that the higher-level hospitals are treating patients who may otherwise have been treated in lower-level facilities.

2. Bed occupancy rate (OCC)

The occupancy rate is a measure of utilization of the available bed capacity. It indicates the percentage of beds occupied by patients in a defined period of time, usually a year. It is computed using the following formula:

$$OCC = \frac{patient\ days}{bed\ days} \times 100 \qquad \text{(A3.2)}$$

Where $patient\ days = admissions \times ALS$; and

$bed\ days = number\ of\ beds \times 365\ (i.e.\ the\ number\ of\ days\ in\ a\ year)$.

This is a method commonly used in assessing hospital performance. Barnum and Kutzin (1993) suggest that hospitals would be operating efficiently at an occupancy rate of 85-90 percent.

Gauging the performance of a hospital's inpatient department solely on the basis of this parameter might prove misleading. As can be seen from the computational formula above (7.1.2), it has a positive relationship with the average length of stay. Thus, it is possible to wrongly classify a facility that keeps patients for an unnecessarily long period of hospitalisation as efficient, although it may not be.

3. Bed turnover ratio (BTR)

The turnover ratio is a measure of bed productivity and represents the number of patients treated per bed in a defined period of time (usually a year). It is computed as:

$$BTR = \frac{total\ patient\ admissions}{number\ of\ beds} \qquad \text{(A3.3)}$$

Turnover ratio in acute care hospitals is expected to be higher than that of chronic hospitals. It is also expected to be higher in lower-level hospitals as compared to higher-level ones.

4. Turnover interval (TI)

This is a measure that is related to the BTR. It measures the average time that beds are unoccupied between successive patients. It is computed as follows:

$$TI = \frac{365}{BTR} - ALS \qquad \text{(A3.4)}$$

The ideal turnover interval is suggested to be 1-3 days.

5. The Pabón Lasso (PL) technique

It has to be stressed that an assessment based on only one of the aforementioned ratios of hospital bed-capacity utilization may be flawed and misleading. Thus, it becomes necessary to make use of all indicators simultaneously, so as to have a better picture. To this end, the method devised by Pabón Lasso (1986) to analyse the performance of a group of hospitals in Colombia is useful.

The PL technique is a graphical method that makes use of the three indicators (BTR, OCC and ALS) concurrently in assessing the relative performance of hospitals. In this method, the occupancy rate (horizontal axis), is plotted against the turnover ratio (on the vertical axis), with vertical and horizontal lines dividing the diagramme into four zones. The horizontal and vertical demarcations represent the mean values of the turnover ratio and occupancy rate. It follows from the functional relationship that exists between the three measures that the slope of the line linking the origin to any of the observations (any point on the graph) represents the reciprocal of the ALS of the hospital under consideration. Figure 7.1.1 represents the possible features of hospitals located in each of the four zones.

Figure A3.1
Simultaneous use of bed utilization measures: characteristics of hospitals

Bed turnover (patients/bed)

Zone II (high BTR, low OCC) • Excess bed capacity • Unnecessary hospitalisation • Many patients admitted for observation • Predominance of normal deliveries	**Zone III (high BTR, high OCC)** • Good quantitative performance • Small proportion of unused beds
Zone I (low BTR, low OCC) • Excess bed supply • Less need for hospitalisation • Low demand/utilization	**Zone IV (low BTR, high OCC)** • Large proportion of severe cases • Predominance of chronic cases • Unnecessarily long stays

Occupancy rate (%)

Source: Adapted from Pabón Lasso (1986)

From the above figure, it can be seen that Zone III, which has relatively high levels of bed occupancy and turnover is the most desirable situation. Zone I is the least desirable. The

setting of the cut-off points at the mean values of the BTR and OCC may be contentious. However, Pabón Lasso (1986) also suggests using other cut-off points (e.g. allowing a margin of one standard deviation from the mean).

6. Average cost per patient day equivalent

In the South African health system, a patient day equivalent (PDE) is defined as the number of inpatient days plus one-third the number of outpatient and casualty visits, i.e.:

$$PDE = inpatient\ days + \frac{1}{3}\ outpatient\ visits \qquad \text{(A3.5)}$$

This is based on the assumption that an inpatient day consumes three times as much resource as an outpatient visit (McIntyre 1997). This premise regarding the weighting is, however, controvertible. In fact, some studies (e.g. Lombard *et al* 1991, McMurchy 1995) have come up with different weighting factors.

Hospitals with a lower cost per PDE are regarded as more efficient than their peers with higher cost per PDE. Unit costs are expected to decrease as one moves down the gradient of hospital levels (McIntyre *et al* 1995).

APPENDIX 4
Input-output data, selected hospital DEA studies

Author (year)	Input(s)	Output(s)
Gerdtham *et al* (1999)	Discharges in surgical department, discharges in short-term internal medicine, surgical operations in short-term care, physician visits in short-term surgical care, physician visits in short-term internal medicine.	Expenditure, beds
Chang (1998)	Physician FTE, Nurse and support personnel FTE, administrative personnel FTE	Number of clinic visits, number of weighted patient days
Ersoy *et al.* (1997)	Beds, number of primary care physicians, number of specialists	Inpatient discharges, outpatient visits, surgical operations
Ferrier and Valdmanis (1996)	FTE personnel, beds	Inpatient days: acute, intensive care, subacute; surgeries, outpatient visits, discharges
Burgess and Wilson (1996)	Acute care beds, long-term beds, registered nurse FTEs, licensed practical nurse FTEs, other clinical labour FTEs, non-clinical labour FTEs,long-term care labour FTEs,	Acute care inpatient days, case-mix weighted acute care inpatient discharges, long-term care inpatient days, number of outpatient visits, ambulatory surgical procedures, inpatient surgical procedures,
Hollingsworth and Parkin (1995)	Medical staff, nursing staff, other staff, capital, drugs	Medical and surgical inpatient days, emergency visits, outpatient visits
Chilingerian (1995)	Total length of stay of each patient, total charges for ancillary services	High severity discharges, low severity discharges
Ozcan and Luke (1993)	Labour: non-physician FTE, weighted number of part-time personnel; supplies (expenses not including payroll, capital or depreciation expenses), capital (beds and number of diagnostic and special services provided)	Case-mix adjusted discharges, outpatient visits, weighted sum of medical and paramedical professionals trained
Ozcan *et al* (1992)	Service complexity, functional beds, non-physician FTEs, weighted part-time personnel, expenditure on supplies	Case-mix adjusted discharges, outpatient visits, training full time equivalents

Author (year)	Input(s)	Output(s)
Morey *et al* (1990)	Beds, type of ownership, case-mix, net plant asset, total annual expenditures	Acute patient days, intensive care patient days, surgeries, outpatient visits, number of residents per attending physician (teaching output)
Grosskopf and Valdmanis (1987)	Number of physicians, Non-physician FTE, admission, net plant assets	Inpatient days: acute and intensive care; surgeries, ambulatory and emergency visits
Sherman (1984)	Bed days, FTEs, supplies	Patient days, nurses trained, interns trained
Nunamaker (1983)	Total inpatient routine costs	Total inpatient aged and paediatric days, total routine maternity days, all other routine days

APPENDIX 5
SUMMARY STATISTICS FOR EFFICIENCY ANALYSIS

Level I hospitals				
Variable	Mean	Standard deviation	Min	Max
Bed	61	26	21	164
Admission	2,626	1,557	530	7,343
Outpatient visit	2,498	2,230	172	9,183
Inpatient day	13,066	6,799	2,572	40,969
Occupancy (%)	58.6	17.9	23.6	117.2
Average length of stay (days)	5.1	1.4	3.4	10.8
Bed turnover rate	42	15.4	12.3	81.9
Expenditure (Rand)	2,069,587	1,311,240	434,913	7,164,874
Level II hospitals				
Bed	112	67	28	284
Admission	5,775	3,115	1,312	13,854
Outpatient visit	37,390	35,986	2,057	146,524
Inpatient day	26,188	14,320	7,891	61,852
Occupancy (%)	71.7	29.2	25.8	159.6
Average length of stay (days)	4.7	1.4	2.6	9.2
Bed turnover rate	52	16	35	107
Expenditure (Rand)	8,987,729	8,056,606	879,665	36,500,000
Level III hospitals				
Bed	483	320	104	1,171
Admission	19,572	10,298	6,436	40,998
Outpatient visit	120,788	96,200	13,264	357,808
Inpatient day	118,994	85,625	35,025	328,806
Occupancy (%)	68.6	12.4	48.5	92.3
Average length of stay (days)	5.9	1.6	3.6	9
Bed turnover rate	46.3	17.7	20	84
Expenditure (Rand)	48,500,000	36,700,000	11,300,000	127,000,000

REFERENCES

Abidoye RO, Ihebuzor NN (2001). Assessment of nutritional status using anthropometric methods on 1-4 year old children in an urban ghetto in Lagos, Nigeria. *Nutrition and Health*, 15 (1): 29-39.

Adetunji JA (ND). *Infant mortality levels in Africa: Does method of estimation matter?* Harvard Center for Population and Development Studies.

Akin J, Birdsall N, De Ferranti D (1987). *Financing Health Services in Developing Countries: An Agenda for Reform.* Washington D.C.: The World Bank.

Alberts Jf, Sanderman R, Eimers JM, van Gen Heuvel WJA (1997). Socio-economic inequality in health care: A study of service utilisation in Curaçao. *Social Science and Medicine*, 45(2): 213-220.

Alderman H (1993). *Including the poor.* Washington D.C.: The World Bank.

Aldrich JH, Nelson FD (1984). *Linear probability, logit, and probit models.* Beverly Hills, CA: SAGE.

Ali M, Assefaw T, Beyene H, Byass P, Hisabu MS, Pedersen FK (2001). A community-based study of childhood morbidity in Tigray, Northern Ethiopia. *Ethiopian Journal of Health Development*, 15(3): 165-172.

Anand S, Diderichsen F, Evans T *et al.* (2001). Measuring disparities in health: Methods and indicators. In: Whitehead M, Evans T, Diderichsen F and Bhuiya A (eds). *Challenging inequities in health: From ethics to action.* New York: Oxford University Press.

ANC (African National Congress) (1994a). *The Reconstruction and development programme – A policy framework.* Johannesburg: Umanyano Publications.

ANC (African National Congress) (1994b). *A National health plan for South Africa.* Johannesburg: ANC.

Anderson DL (1980). A statistical cost function study of public general hospitals in Kenya. *Journal of Developing Countries*, 14: 223-235.

Atkinson AB (1970). On the measurement of income inequality. *Journal of Economic Theory*, 2:244-263.

Bachmann M, London L, Barron P (1996). Infant mortality rate inequality in the Western Cape Province of South Africa. *International Journal of Epidemiology*, 25(5): 966-972.

Baker JL, van der Gaag J (1993). Equity in health care and health care financing: Evidence from five developing countries. In: van Doorslaer E, Wagstaff A, Rutten E (eds). *Equity in the finance and delivery of health care: An international perspective.* Oxford: Oxford University Press.

Banker RD, Charness A, Cooper WW (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(1): 1078-1092.

Barnum H, Kutzin J (1993). *Public Hospitals in Developing Countries: Resource use, Cost and Financing.* Baltimore: Johns Hopkins University press for the World Bank.

Barron P, Strachan K, Ijsselmuiden C (1997). The year in review. In Health Systems Trust. *South African Health Review 1997.* Durban: Health Systems Trust.

Basta SS (1977). Nutrition and health in low-income urban areas of the third world. *Ecology of Food and Nutrition.* 6: 113-124.

Behrman J, Deolalikar A (1988). Health and Nutrition. In: Behrman J, Deolalikar A (eds). *Handbook of development economics.* Amsterdam: North Holland.

Behrman J, Hoddinott J (2000). *An Evaluation of the impact of PROGRESA on pre-school child height.* Washington D.C.: International Food Policy Research Institute. Mimeo.

Belcher DW, Neumann AK, Wurapa FK, Lourie IM (1976). Comparison of Mortality Interviews with a Health Examination Survey in Rural Africa. *American Journal of Tropical Medicine and Hygiene.* 25(5): 751-758.

Bjurek H, Hjalmarsson L, Forsund FR (1990). Deterministic parametric and non-parametric estimation of efficiency in service production: A comparison. *Journal of Econometrics*, 46(1/2): 213-227.

Blane D, Davey SG, Filakti G *et al* (1994). Social patterning of medical mortality in youth and early adulthood. *Social Science and Medicine*, 39: 361-366.

Bradley D, Stephens C, Harpham T, Cairncross S (1992). A review of environmental health impacts in developing country cities. Washington DC: The World Bank.

Bradshaw D (1999). Health for All – Monitoring equity. In Health Systems Trust (eds). *South Africa Health Review 1999*. Durban: Health Systems Trust.

Braveman P (1998). *Monitoring equity in health: A policy-oriented approach in low- and middle-income countries.* WHO/CHS/HSS/98.1, Equity Initiative Paper No. 3. Geneva: World Health Organization.

Breen R (1996). *Regression models: Censored, sample selected, or truncated data*. CA: SAGE

Brockerhoff M, Hewett P (2000). Inequality in child mortality among ethnic groups in sub-Saharan Africa. *Bulletin of the World Health Organization*, 78(1): 30-41.

Brownell MD, Roos NP (1995). Variations in length of stay as a measure of efficiency in Manitoba hospitals. *Canadian Medical Association Journal*, 152(5): 675-682.

Butler JR (1995). *Hospital Cost analysis*. Dordrecht: Kluwer Academic Publishers.

Carr-Hill R (1987). The inequalities in health debate: A critical review of the issues. *Journal of Social Policy*, 16(4): 509-542.

Carr-Hill R (1990). The measurement of inequities in health: Lessons from the British experience. *Social Science and Medicine*, 33(3): 393-404.

Case A, Wilson F (2001). *Health and wellbeing in South Africa: Evidence from the Langeberg Survey*. USA: Princeton University. Mimeo.

Castro-Leal F, Dayton J, Demery L, Mehra K (1999). Public social spending in Africa: Do the poor benefit? *The World Bank Research Observer*, 14(1), 49-72.

Caves D, Christensen L, Diewert E (1982). The economic theory of index numbers and the measurement of input, output and productivity. *Econometrica*, 50(6): 1393-1414.

Cebu Study Team (1991). Underlying and proximate determinants of child health: The Cebu longitudinal health and nutrition study. *American Journal of Epidemiology*, 133: 185-201.

CSS (Central Statistical Service) (1997). *October Household Survey 1997*. Pretoria: CSS.

CSS (Central Statistical Service) (1998).*Statistics in brief: RSA 1997*.

Charness A, Cooper WW, Lewin AY, Seiford LM (1996). *Data Envelopment Analysis: Theory, Methodology and Applications*. Boston: Kluwer Academic Publishers.

Charness A, Cooper WW, Rhodes E (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2: 429-444.

Chen LC, Huq E, D'Souza S (1981). Sex bias in the family allocation of food and health care in rural Bangladesh. *Population and Development Review*, 7:147-163.

Chiang T (1999). Economic transition and changing relation between inequality and mortality in Taiwan: Regression analysis. *British Medical Journal*, 319: 1162-1165.

Choudhury KK, Hanifi MA, Rasheed S, Bhuiya A (2000). Gender inequality and severe malnutrition among children in a remote rural area of Bangladesh. *Journal of Health, Population and Nutrition*. 18(3): 123-130.

Clewer A, Perkins D (1998). *Economics for Health Care Management*. London: Prentice Hall.

Coelli T, Rao DSP, Battese G (1998). *An Introduction to Efficiency and Productivity Analysis*. Boston: Academic Publishers.

Coelli TJ (1996). *A Gude to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Programme*. CEPA Working Paper 96/8. Department of Econometrics, University of New England.

Cooper WW, Seiford LM, Tone K (2000). *Data envelopment analysis*. USA: Kluwer Academic Publishers.

Cornell J, Goudge J, McIntyre D, Mbatsha S (2001). *National Health Accounts: The public sector report.*

Cornia GA, Mwabu G (1997). *Health status and health policy in sub-Saharan Africa: A long-term perspective.* World Institute for Development Economics Research, Working Paper No. 141: Helsinki: UNU/WIDER.

Cravioto J, Arrieta R (1986). Nutrition, mental development and learning, in Falkner F, Tanner J (eds), *Human growth: A comprehensive treatise.* New York: Plenum Press. Volume 3, 2/e.

Culyer AJ (2001). Equity- some theory and its implications. *Journal of Medical Ethics*, 27: 275-283.

Culyer AJ, Wagstaff A (1993). Equity and equality in health and health care. *Journal of Health Economics*, 12(4): 431-457.

Dahlgren G, Whitehead M (1992). *Policies and strategies to promote equity in health.* Copenhagen: World Health Organization, Regional Office for Europe.

Daniels N (1985). *Just health care.* Cambridge: Cambridge University Press.

Day C, Gray A (2002). Health and related indicators. In Health Systems Trust (eds). *South Africa Health Review 2001.* Durban: Health Systems Trust.

Deaton A (1997). *The analysis of Household surveys: A microeconometric approach to development policy.* Washington D.C.: The World Bank.

de Onis M (2000). Measuring nutritional status in relation to mortality. *Bulletin of the World Health Organization*, 78(10): 1271-1280.

de Onis M, Frongillo EA, Blossner M (2000). Is malnutrition declining? An analysis of changes in levels of child malnutrition since 1980. *Bulletin of the World Health Organization*, 78 (10): 1222-1233.

de Onis M, Monteiro C, Akre J, Clugston G (1993). The worldwide magnitude of protein-energy malnutrition: An overview from WHO global database on child growth. *Bulletin of the World Health Organization*, 71 (6): 703-712.

Debreu G (1951). The coefficient of resource utilization. *Econometrica*, 19: 273-292.

Department of Finance (1996). *Growth, employment and redistribution. A macro-economic strategy*. Cape Town: Department of Finance.

Department of Health, Medical Research Council, Macro International (1999). *South Africa Demographic and Health Survey 1998: Preliminary Report*. Pretoria: Department of Health.

Department of Health, Medical Research Council, Macro International (1999). *South Africa Demographic and Health Survey 1998: Full Report*. Pretoria: Department of Health.

Duflo E (2000). Child health and household resources in South Africa: Evidence from the old age pension programme. *The American Economic Review*, 90(2): 393-398, Papers and Proceedings.

Dworkin R (1994). "Will Clinton's plan be fair? *New York Review of Books*, 13 January, 20-25.

Eachus J, Chan P, Pearson N, Propper C, Smith GD (1999). An additional dimension to health inequalities: Disease severity and socio-economic position. *Journal of Epidemiology and Community Health*, 53: 603-611.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.

El-Sayed N, Mohamed AG, Nofal L, Mahfouz A, Zeid HA (2001). Malnutrition among pre-school children in Alexandria, Egypt. *Journal of Health, Population and Nutrition*, 19(4): 275-280.

Ersoy K, Kavuncubasi S, Ozcan YA, Harris JM 2nd (1997). Technical efficiencies of Turkish hospitals: DEA approach. *Journal of Medical Systems*, 21(2): 67-74.

Färe R, Grosskopf S, Lindgren B, Roos P (1994). Productivity developments in Swedish hospitals: A Malmquist index approach, in *Data Envelopment Analysis: Theory, Methodology and Applications*, Charness A, Cooper WW, Lewin AY, Seiford LS (eds), Boston: Kluwer Academic Publishers, 253-272.

Farell MJ (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, 120(3): 283-281.

Feachem RGA (2000). Poverty and inequity: A proper focus for the new century. *Bulletin of the World Health Organization*, 78(1): 1-2.

Fernandez E, Schiaffino A, Rajmil L, Badia X, Segura A (1999). Gender inequalities in health and health care services use in Catalonia (Spain). *Journal of Epidemiology and Community Health*, 53: 218-222.

Ferrier GD, Valdmanis V (1996). Rural hospital performance and its correlates. *The Journal of Productivity Analysis*, 7: 63-80.

Flegg A (1982). Inequality of income, illiteracy and medical care as determinants of mortality in developing countries. *Population Studies*, XXXVI: 441-458.

Fried HO, Lovell CAK, Schmidt SS (eds) (1993). *The measurement of productive efficiency: Techniques and applications*. New York: Oxford University Press.

Gakidou EE, Murray CJL, Frenk J (1999). *A Framework for measuring health inequality*. World Health Organization: GPE Discussion Paper No. 5.

Garret JL, Ruel MT (1999). Are determinants of rural and urban food security and nutritional status different? Some insights from Mozambique. *World Development*, 27 (11): 1955-1975.

Gelband H, Stansfield S (2001). The evidence base for interventions to reduce under five mortality in low and middle-income countries. Commission on Macroeconomics and Health, CMH Working Paper Series No. WG5:9. WHO: Geneva.

Gerdtham U-G (1997). Equity in health care utilisation: Further tests based on Hurdle models and Swedish micro-data. *Health Economics*, 6:303-319.

Getaneh T, Assefa A, Tadesse Z (1998). Protein-energy malnutrition in urban children: Prevalence and determinants. *Ethiopian Medical Journal*, 36 (3): 153-166.

Gillon R (1985). *Philosophical medical ethics*. New York: Wiley.

Gilson L (1988). *Government Health care charges: Is equity being abandoned?* Publication Number 15, Evaluation and Planning Centre for Health Care, London School of Hygiene and Tropical Medicine.

Gilson L, McIntyre D (2001). Experiences from South Africa: dealing with a poor health legacy of apartheid. In: Whitehead M, Evans T, Diderichsen F and Bhuiya A (eds). *Challenging inequities in health: From ethics to action.* New York: Oxford University Press.

Giokas DI (2001). Greek hospitals: How well their resources are used. *Omega International Journal of Management Science.* 29: 73-81.

Glewwe P (1991). Investigating the determinants of household welfare. *Journal of Development Economics.*

Golany B, Roll Y (1989). An application procedure for DEA. *OMEGA International Journal of Management Science*, 17(3): 237-250.

Greene WH (1993). *Econometric analysis.* New York: Macmillan, 2nd edition.

Grosskopf S, Valdmanis V (1987). Measuring hospital performance: A non-parametric approach. *Journal of Health Economics*, 6: 89-107.

Gujarati DN (1995). *Basic econometrics.* USA: McGraw-Hill, Inc.

Gwatkin DR (2000). Health inequalities and the health of the poor: What do we know? What can we do? *Bulletin of the World Organization*, 78(1): 3-18.

Gwatkin DR (2001). Poverty and inequalities in health within developing countries: Filling the information gap. In Leon D, Walt G (eds). *Poverty, inequalities and health: An international perspective.* Oxford: Oxford University Press.

Haddad L, Ruel MT, Garrett JL (1999). Are poverty and urbanization growing? Some Newly assembled evidence. *World Development*, 27 (11): 1891-1904.

Haldenwang BB, Boshof SC (1996). Forecasts of the South African Population, 1991-2026. Stellenbosch: Institute for Futures Research, Stellenbosch University.

Hao S, Pegles CC (1994). Evaluating relative efficiencies of veterans affairs medical centers using data envelopment analysis. *Journal of Medical Systems.* 18(2): 55-67.

Health Systems Trust (1999). *South African Health Review 1998*. Durban: Health Systems Trust.

Health Systems Trust (2000). *South African Health Review 1999*. Durban: Health Systems Trust.

Henderson G, Akin J, Zhiming L, Shuigao J, Haijiang M, Keyou G (1994). Equity and utilization of health services: Report of an eight-province survey in China. *Social Science and Medicine*, 39(5): 687-699.

Hojman DE (1996). Economic and Other determinants of infant and child mortality in small developing countries: the case of Central America and the Caribbean. *Applied Economics*, 28: 281-290.

Humphries KH, van Doorslaer E (2000). Income-related health inequality in Canada. *Social Science and Medicine*, 50: 663-671.

Idler EI, Angel RJ (1990). Self-rated health and mortality in NHANES-1 epidemiological follow-up study. *American Journal of Public Health*, 80: 446-452.

Illsley R, Svensson PG (1990). Social inequalities in health. *Social Science and Medicine*, 31: 223-240.

Jack W (1999). *Principles of health economics for developing countries*. Washington, D.C.: The World Bank.

Kakwani N, Wagstaff A, van Doorslaer E (1997). Socioeconomic inequalities in health: Measurement, computation and statistical inference. *Journal of Econometrics*, 77: 87-103.

Kennedy BP, Kawachi I, Glass R, Prothrow-Stith D (1998). Income distribution, socio-economic status, and self-rated health in the United States: Multilevel analysis. *British Medical Journal*, 317: 917-921.

Koopmans TC (1951). An analysis of production as an efficient combination of activities, in Koopmans TC (eds), *Activity Analysis of Production and Allocation*. New York: John Wiley and sons Inc.

Kunst AE, Geurts JJM, van den Berg J (1995). International variation in Socioeconomic inequalities in self reported health. *Journal of Epidemiology and Community Health*, 49: 117-123.

Kunst AE, Mackenbach JP (1994). International variation in the size of mortality differences associated with occupational status. *International Journal of Epidemiology*, 23(4): 742-750.

Lahelma E, Manderbacka K, Rahkonen O, Karisto A (1994). Comparisons of inequalities in health: Evidence from national surveys in Finland, Norway and Sweden. *Social Science and Medicine*, 38(4): 517-524.

Lairson DR, Hindson P, Hauquitz A (1995). Equity of health care in Australia. *Social Science and Medicine*, 41(4): 475-482.

Le Grand J (1984). Equity as an economic objective. *Journal of Applied Philosophy*, 1: 39-51.

Le Grand J (1996). Equity, efficiency and rationing of health care. In: Culyer AJ, Wagstaff A (eds). *Reforming health care systems: Experiments with the NHS.* Chetenham: Brookfield.

Liao TF (1994). *Interpreting probability models: Logit, probit, and other generalized linear models.* USA: SAGE Publications.

Linna M (1999). *Measuring Hospital Performance: The Productivity, Efficiency and costs of Teaching and Research in Finnish hospitals.* Stakes research report 98.

Long JS (1997). *Regression models for categorical and limited dependent variables.* USA: SAGE Publications.

Lovell CAK (1993). Production frontiers and productive efficiency, in Fried HO, Lovell CAK, Schmidt SS (eds), *The Measurement of Productive Efficiency: Techniques and Applications.*

Maddala GS (1983). *Limited-dependent and qualitative variables in econometrics.* Cambridge: Cambridge University Press.

Magill FN (ed) (1997). *International Encyclopedia of Economics*. London: Fritzory Dearborn publishers.

Magnussen J (1996). Efficiency measurement and the operationalisation of hospital production. *Health Services Research*, 31(1): 21-37.

Makan B (1998). Distribution of health personnel. In Ntuli A (ed). *South African Health Review 1998*. Durban: Health Systems Trust.

Manor A, Matthews S, Power C (1997). Comparing measures of health inequality. *Social Science and Medicine*, 45(5): 761-771.

Manuel TA (2002). *2002 Budget Speech*. www.treasury.gov.za, downloaded on 3 August 2002.

Margolis H (1992). *Selfishness, altruism, and rationality*. Cambridge: Cambridge University Press.

Mastilica M (1990). Health and social inequalities in Yugoslavia. *Social Science and Medicine*, 31(3): 405-412.

May J (ed) (1998). *Poverty and inequality in South Africa*. Report prepared for the Office of the Executive Deputy President and inter-Ministerial Committee for poverty and inequality. Durban: Praxis Publishing.

McIntyre D, Baba L, Makan B (1998a). Equity in public sector health care financing and expenditure in South Africa. In Health Systems Trust (eds). *South Africa Health Review 1998*. Durban: Health Systems Trust.

McIntyre D, Bloom G, Doherty J, Brijlal P (1995). H*ealth expenditure and finance in South Africa*. Durban: Health Systems Trust and the World Bank.

McIntyre D, Gilson L, Valentine N, Soderlund N (1998b). *Equity of health sector revenue generation and allocation: A South African case study*. Washington, DC: Partnerships for Health Reform.

McIntyre D, Muirhead D, Gilson L *et al.* (2001). *Geographic patterns of deprivation and health inequities in South Africa*: Informing public resource allocation strategies. EQUINET Policy Series No. 10.

McLachlan G, Maynard A (eds) 1982. *The public/private mix in health care: The relevance and effects of change*. London: Nuffield Provincial Hospitals Trust.

McMURCHY, D (1996). *Efficiency Studies of Public Sector Hospitals in the Eastern, Northern and Western Cape Provinces*. HEU Working Paper No. 6. Cape Town: Health Economics Unit, University of Cape Town.

Menon P, Ruel MT, Morris S (2000). *Socioeconomic differentials in child stunting are consistently larger in Urban than in Rural areas*. Washington DC: International Food Policy Research Institute, FCND Discussion Paper No.97.

Mersha T (1989). Output and performance measurement in outpatient care. *OMEGA International Journal of Management Science*, 17(2): 159-167.

Mobley LR, Bradford WD (1997). Profit variability among hospitals: Is it ownership, or location. *Applied Economics*, 29: 1125-38.

Mooney G (1983). Equity in health care: Confronting the confusion. *Effective Health Care*, 1(4): 179-185.

Mooney GH (1996). And now for vertical equity? Some concerns arising from aboriginal health in Australia. *Health Economics*, 5:99-103. (Guest editorial).

Murray CJL, Gakidou EE, Frenk J (1999). Health inequalities and social group differences: What should we measure. *Bulletin of the World Health Organisation*, 77(7): 537-542.

Murray CJL, Young G, Qiao X (1992). Adult mortality: Levels, patterns and causes. In: Feachem RGA, Kjellstrom T, Murray CJL (eds), *The health of adults in the developing world*, New York: Oxford University Press.

Nannan N, Bradshaw D, Mazur R, Maphumulo S (1998). What is the infant mortality rate in South Africa? The need for improved data. *South African Medical Journal*, 88(12): 1583-1587.

Nandakumar AK (2000). Utilization of outpatient care in Egypt and its implications for the role of government in health care provision. *World Development*, 28(1): 187-196.

Narayan D, Patel R, Schafft K, Rademacher A, Koch-Schulte S (2000). *Voices of the poor: Can anyone hear us?* New York: Oxford University Press.

Nemer L, Gelband H, Jha P (2001). *The Evidence Base for Interventions to Reduce Malnutrition in Children Under Five and School-age Children in Low and Middle-income Countries.* Geneva: WHO, Commission on Macroeconomics and Health, CMH Working Paper Series, Paper No. WG5: 11.

Newbold KB, Eyles J, Birch S (1995). Equity in health care: Methodological contributions to the analysis of hospital utilization within Canada. *Social Science and Medicine*, 40(9): 1181-1192.

Ngare DK, Muttunga JN (1999). Prevalence of malnutrition in Kenya. *East African Medical Journal*, 76(7): 376-380.

Nozick R (1974). *Anarchy, state and Utopia.* New York: Basic Books.

Ntsaluba A, Pillay Y (1998). Reconstructing and developing the health system – the first 1000 days. *South African Medical Journal*, 88(1): 33-36.

Nunamker T (1983). Measuring routine nursing service efficiency: A comparison of cost per patient day and data envelopment analysis models. *Health Services Research*, 18(2): 183-205.

Olsen JA (1997). Theories of justice and their implications for piority setting in health care. *Health Economics*, 16: 625-639.

Ozcan YA Luke RD (1993). A national study of the efficiency of hospitals in urban markets. *Health Services Research*, 27(6): 719-739.

Ozcan YA, McCue MJ, Okasha AA (1996). Measuring the technical efficiency of psychiatric hospitals. *Journal of Medical Systems*, 23(1): 57-71.

Pal S (1999). An analysis of child malnutrition in rural India: Role of gender, income and other household characteristics. *World Development*, 27(7): 1151-1171.

Peacock D, Devlin N, McGee R (1999). The horizontal equity of health care in New Zealand. *Australian and New Zealand Journal of Public Health*, 23(2): 126-130.

Pelletier DL, Frongillo EA, Schroeder DG, Habicht JP (1995). The effects of malnutrition on child mortality in developing countries. *Bulletin of the World Health Organization*, 73(4): 443-448.

Perèira J (1990). The economics of inequality in health: A bibliography. *Social Science and Medicine*, 31(3): 413-420.

Perèira J (1993). What does equity in health mean? *Journal of social policy*, 22(1): 19-48.

Phimmasone K, Douangpoutha I, Fauveau V, Pholsena P (1996). Nutritional status of children in the Lao PDR. *Journal of Tropical Paediatrics*, 42(1): 5-11.

Pillay YG, Bond P (1995). Health and social policies in the new South Africa. *International Journal of Health Services*, 25(4): 727-743.

Rahkonen O, Lahelma E (1992). Gender, social class and illness among young people. *Social Science and Medicine*, 34: 649-656.

Rawls J (1971). *The theory of justice*. Cambridge: Harvard university press.

Regidor E, Dominguez V, Navarro P, Rodriguez C (1999). The magnitude of differences in perceived general health associated with educational level in the regions of Spain. *Journal of Epidemiolgy and Community Health*, 53: 288-293.

Rogers G (1979). Income and inequality as determinants of mortality: An international cross sectional analysis. *Population Studies*, XXXIII: 343-352.

Rosko MD, Chilingerian JA (1999). Estimating hospital inefficiency: Does case-mix matter? *Journal of Medical Systems*, 23(1): 57-71.

Rosko MD, Chilingerian JA, Zinn JS, Aaronson WE (1995). The effects of ownership, operating environment, and strategic choices on nursing efficiency. *Medical Care*, 33(10): 1001-1021.

Ross NA, Wolfson MC, Dunn JR *et al* (2000). Relations between income inequality and mortality in Canada and in the United States: Cross sectional assessment using census data and vital statistics. *British Medical Journal*, 320: 898-902.

Ruel MT, Haddad L, Garrett JL (1999). Some Urban facts of life: Implications for research and policy. *World Development*, 27 (11): 1917-1938.

Rustein SO (2000). Factors associated with trends in infant and child mortality in developing countries during the 1990s. *Bulletin of the World Health Organization*, 78(10): 1256-1270.

Sahn DE (1994). The contribution of income to improved nutrition in Côte d'Ivoire. *Journal of African Economies*, 3(1): 29-61.

Schultz TP (1993). Mortality decline in the low-income world: Causes and consequences. *American Economic Association Papers and Proceedings*, 83(2): 337-342.

Seiford LM, Thrall RM (1990). Recent developments in DEA: The mathematical programming approach to frontier analysis. *Journal of Econometrics*, 46: 7-38.

Sen A (1973). On *economic inequality*. Oxford: Clarendon press.

Sen A (2001). Health equity: Perspectives, measurability and criteria. In Evans T, Whitehead M, Diderichsen F, Bhuiya A, Wirth M (eds). *Challenging inequities in health: from ethics to action,* New York: Oxford University Press.

Sen A (2002). Health: Perception versus observation. *British Medical Journal*, 324: 860-1.

Sengupta JK (1998). Stochastic data envelopment analysis: A new approach. *Applied Economic Letters*, 5: 287-290.

Shephard RW (1970). *Theory of Cost and Production Functions*. Princeton, N.J.: Princeton University Press.

Sherman HD (1984). Hospital efficiency measurement and evaluation: Empirical test of a new technique. *Medical Care*, 22(10): 922-935.

Sidiropoulos E, Jeffery A, Forgey H *et al.* (1998). *South Africa Survey 1997/98*. Johannesburg: South African Institute of Race Relations.

Skoufias E (1998). Determinants of child health during the economic transition in Romania. *World Development*, 26(11): 2045-2056.

Smith P, Mayston D (1987). Measuring efficiency in the public sector. *Omega, International Journal of Management Science*, 15(3): 181-189.

Soderlund N, Schierhout G, van den Heever A (2001). Private health sector care. In Health Systems Trust (eds). *South Africa Health Review 2001.* Durban: Health Systems Trust.

South Africa (Republic) (1997). *White paper for the transformation of the health system in South Africa.* Government Gazette 382 (17910): Notice 667.

Stansfield S, Gelband H (2001). *The evidence base for interventions to reduce under five mortality in low and middle-income countries.* Geneva: WHO, CMH Working Paper Series, Paper No. WG5: 9.

Stanton B (1994). Child health: equity in the non-industrialized countries. *Social Science and Medicine*, 38(10): 1375-1383.

Statacorp (1997). *Stata statistical software: Release 5.0.* TX: Stata corporation.

Statistics South Africa (1995). *October Household Survey 1994.* Pretoria: Statistics South Africa.

Statistics South Africa (1996). *The people of South Africa –Population census 1996.* Census in brief. Pretoria: Statistics South Africa.

Statistics South Africa (1998). *October Household Survey 1998.* Pretoria: Statistics South Africa.

Suárez-Berenguela RM (2000). *Health system inequalities and inequities in Latin America and the Caribbean: Findings and policy implications.* Working document prepared for the Health and Human Development Division of the Pan American Health Organization – World Health Organization.

Svedberg P (1987). *Undernutrition in Sub-Saharan Africa: A critical assessment of the evidence.* World Institute for Development Economics Research, Working Paper No. 15: Helsinki: UNU/WIDER.

Thanassoulis E, Dyson RG (1992). Estimating preferred target input-output levels using data envelopment analysis. *European Journal of Operational Research*, 56: 80-97.

Thomas D, Strauss J, Henriques MH (1990). Child survival, height for age and household characteristics in Brazil. *Journal of Development Economics*, 33: 197-234.

Twaddle AC (1979). The concept of health status. In: Jaco E (eds), *Patients, physicians and illness*. New York: The Free Press.

UNAIDS (Joint United Nations Programme on HIV/AIDS) (2002). *Report on the global HIV/AIDS epidemic*. Geneva: UNAIDS.

UNDP (United Nations Development Program) (1996). *Human Development Report*. New York: Oxford University Press.

UNDP (United Nations Development Program) (1999). *Human Development Report*. New York: Oxford University Press.

UNDP (United Nations Development Program) (2000). *Human Development Report*. New York: Oxford University Press.

UNICEF (United Nations Children's Fund). *The State of the World's Children 2000.* New York: UNICEF.

United Nations Statistics Division website. http://milleniumindicators.un.org.unsd. Accessed on 13 November 2002.

Vagero D, Lundberg O (1989). Health inequalities in Britain and Sweden. *Lancet*, ii: 35-36.

Valdmanis V (1992). Sensitivity analysis for DEA models: An empirical example using public vs. NFP hospitals. *Journal of Public Economics*, 48: 185-205.

van de Walle D (1995). The distribution of subsidies through public health services in Indonesia, 1978-87. In van de Walle D, Nead K (eds). *Public spending and the poor: Theory and evidence*. USA: The Johns Hopkins University Press for the World Bank.

van Doorslaer E, Wagstaff A, van de Burg D *et al* (2000). Equity in the delivery of health care in Europe and the United States. *Journal of Health Economics*, 19: 553-583.

van Rensburg ACJ, Fourie A, Pretorius E (1992). *Health Care in South Africa: Structure and dynamics*. Pretoria: Academica.

Varian H (1974). Equity, envy and efficiency. *Journal of economic theory*, 9: 63-91.

Wagstaff A (2000). Socioeconomic inequalities in child mortality: Comparison across nine developing countries. *Bulletin of the World Health Organization*, 78(1): 19-29.

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

Wagstaff A, van Doorslaer E (1993). Equity in the finance and delivery of health care: Concepts and definitions. In: van Doorslaer E, Wagstaff A, Rutten F (eds). *Equity in the finance and delivery of health care: An International Perspective*. Oxford: Oxford University Press.

Wagstaff A, van Doorslaer E (2000). Measuring and testing inequity in the delivery of health care. *The Journal of Human Resources*, XXXV(4): 716-733.

Wagstaff A, van Doorslaer E, Paci P (1989). Equity in the finance and delivery of health care: Some tentative cross-country comparisons. *Oxford Review of Economic Policy*. 5: 89-112.

Waldman RJ (1992). Income distribution and infant mortality. *The quarterly Journal of economics*, 107(4): 1283-1302.

Whitehead M (1988). The health divide. In Townsend P, Davidson M, Whitehead M (eds), *Inequalities in health*. London: Penguin.

Whitehead M (1993). Is it fair? Evaluating the equity implications of the NHS reforms, in Robinson R, Le Grand J (eds). *Evaluating the NHS reforms*. UK: Kings Fund Institute.

Whitehead M, Drever F (1999). Narrowing social inequalities in health? Analysis of trends in mortality among babies of lone mothers. *Electronic British Medical Journal*, 1-5.

Whitehead M, Scott-Samuel A, Dahlgren G (1998). Setting targets to address inequalities in health. *The Lancet*, 351: 1279-1282.

WHO (World Health Organization) (1978). *Primary Health Care: Report of the International Conference on Primary Health Care, Alma-Ata, USSR*. Geneva: WHO.

WHO (World Health Organization) (1999). *World Health Report 1999*. Geneva: WHO.

WHO (World Health Organization) Expert Committee on Nutrition (1995). *Physical status, uses and interpretation of anthropometry*. WHO Technical Report Series No. 854. Geneva: WHO.

WHO Working Group (1986): Use and interpretation of anthropometric indicators on nutritional status. *Bulletin of the World Health Organization*, 64(6): 929-941WHO (World Health Organization) (1999). *The World Health Report 1999*. Geneva: WHO.

Wilkinson RG (1996). *Unhealthy societies: The afflictions of inequality*. USA: Routledge.

Williams BT (1990). Assessing the health impact of urbanization. *World Health Statistics Quarterly.*

Wolfson M, Kaplan G, Lynch J, Ross N, Backlund E, (1999). Relation between income inequality and mortality: Empirical demonstration. *British Medica Journal*, 319: 953-957.

World Bank (1981). *World Development Report*. Washington: The World Bank.

World Bank (1998a). *African Development Indicators*. Washington DC: The World Bank.

World Bank (1998b). *World development indicators 1998*. Washington DC: The World Bank.

World Bank (1999). *World Development Report*. Washington DC: The World Bank.

Wouters A (1993). The cost and efficiency of public and private health care facilities in Ogun State, Nigeria. *Health Economics*, 2(1): 31-42.

Zwi A (2001). Injuries, inequalities and health: From policy vacuum to policy action. In: Whitehead M, Evans T, Diderichsen F and Bhuiya A (eds). *Challenging inequities in health: From ethics to action.* New York: Oxford University Press.