Equity and efficiency in health and health care for HIV-positive adults in South Africa

Thesis presented for the degree of
DOCTOR OF PHILOSOPHY

In the

Health Economics Unit
School of Public Health and Family Medicine
Faculty of Health Sciences
University of Cape Town

May 2007

Candidate

Susan Cleary

Supervised by

Prof Di McIntyre (Health Economics Unit, University of Cape Town, South Africa)
Prof Gavin Mooney (Social and Public Health Economics Research Group, Curtin University, Australia)

Contents

Index of tables

Index of figures

Abstract

This dissertation presents a framework for assessing equity and efficiency in health and health care for HIV-positive adults in South Africa, which is tested in the extensive analysis of empirical data on the costs and consequences of alternative HIV-treatment strategies in the public health care system.

The framework is built through asking three key questions. The first question – what is the good (value or benefit) of health care – considers what ought to be in the evaluative space of distributive justice in relation to this dissertation and in health economics more generally. The second question considers the factors that might constitute claims on this good, including personal responsibility, need, the social context as well as the impact of allocations of the good on the health of society and the social fabric. The final question – how should the good be distributed – examines alternative social choice rules for distributing the good and develops an approach grounded in procedural justice that legitimizes the choice of one rule over another.

To apply this framework, patient and population-level costs and consequences associated with alternative HIV-treatment interventions are analysed in Markov models. These are extensively validated and uncertainty is assessed through probabilistic and multi-way sensitivity analyses. Results of these analyses are key inputs into mathematical programming algorithms that allow an assessment of the implications of choosing one social choice rule over another in terms of gains in the good and the proportion of need that can be met through one or more treatment strategy across a range of budgets.

In discussing and concluding, these empirical results are reintegrated into the conceptual framework where the notion of claims on the good and a decision-making approach grounded in procedural justice is further developed. It is argued that the proper implementation of this framework will result in allocations of the good that are fair even if this is at a level of less than universal access to the most effective treatment strategy.

I would also like to acknowledge the support of Michael Thiede and Stephane Luchini who gave helpful comments on chapters and drafts, and other colleagues at the Health Economics Unit who took on some of my teaching responsibilities while I was on sabbatical.

A final word of acknowledgement goes to my supervisors, Di McIntyre and Gavin Mooney, who have been instrumental in the development of the conceptual and theoretical aspects of this dissertation and have supervised the extensive re-analysis of data from research projects into the format presented below.

Abbreviations

ART: Antiretroviral treatment

ARVs: Antiretroviral medicines

CHC: Community health centre

CMV: Cytomegalovirus

DHS: Demographic and Health Survey

DOTS: Directly observed treatment, short course

HAART: Highly active antiretroviral therapy, synonymous with ART

HIV/AIDS: Human Immunodeficiency Virus / Acquired Immune Deficiency Syndrome

IPD: Hospital inpatient day

MAC: Mycobacterium avium complex

MSF: Medecins sans Frontieres

NGO: Non-governmental organisation

NDoH: National Department of Health

OI: Opportunistic infection

OPD: Hospital outpatient department

PBMA: Programme budgeting and marginal analysis

PCP: Pneumocystis carnii pneumonia

PDE: Patient day equivalent

PMTCT: Prevention of mother to child transmission

PSA: Probabilistic sensitivity analysis

STI or STD: Sexually transmitted infection or sexually transmitted disease

TB: Tuberculosis

UNAIDS: The Joint United Nations Programme on HIV/AIDS

US$: United States dollars

WHO: World Health Organisation

Index of key terms

First-line ART: A treatment strategy for HIV that includes first-line antiretrovirals only and the ongoing treatment and prophylaxis of opportunistic and HIV-related infections and events.

First and second-line ART: A treatment strategy for HIV that includes first and second-line antiretrovirals and the ongoing treatment and prophylaxis of opportunistic and HIV-related infections and events.

HIV-treatment: Any treatment strategy for HIV; in this dissertation it includes No-ART (see below), first-line ART and first and second-line ART.

Independent programmes and mutually exclusive interventions: An independent programme such as HIV-treatment consists of a set of mutually exclusive interventions which in this case includes No-ART (see below), first-line ART and first and second-line ART. Because these are mutually exclusive, a patient can only receive one of these treatment strategies at a time.

Markov model: A state transition model that is particularly appropriate for simulating long-term diseases. A Markov model consists of a number of Markov states. The probability of moving between these states over "time" is determined by transition probabilities. The model is run over time until all patients are collected in an absorbing state. If this state represents death, then the model estimates life expectancy.

Monte Carlo simulation: First-order Monte Carlo simulation is a way of solving a Markov model. Each "patient" is sent through the model one at a time. Because different patients can take different paths (a number of transitions are possible from each Markov state) one is able to build up a profile of potential outcomes and to assess the variability in results. Through second-order Monte Carlo simulation, uncertainty in underlying data can also be assessed. If a range is attached to each transition probability instead of a single value, one is able to capture any data uncertainty in the final outcomes.

No-ART: Treatment and prophylaxis of opportunistic and HIV-related infections and events without ARVs.

Probabilistic sensitivity analysis: First and second-order Monte Carlo simulation

Chapter 1: Introduction

South Africa's HIV-epidemic is one of the worst in the world. A relatively large population coupled with high prevalence has meant that South Africa has only recently been surpassed by India as the country with the highest number of HIV-infected people. According to ASSA2003 projections[1], the country has experienced a rapidly growing HIV epidemic since the mid 1980s, with a peak in incidence at 781,000 new infections in 1998. In 2005, there were approximately 6 million HIV-infected people and over 400,000 HIV-related deaths. At least 1.8 million people have died of HIV-related causes during the twenty years since the start of the epidemic. Because of this, life-expectancy has declined from a high of 62.9 years in 1989 to 49.2 years in 2005 and HIV-related deaths currently exceed deaths from all other single causes.

Antiretroviral treatment (ART) has been shown to be effective in reducing morbidity and mortality in patients infected with HIV in developing countries (ART-LINC and ART-CC 2006) and the feasibility of providing access to this treatment has been demonstrated in pilot projects (Coetzee, Hildebrand et al. 2004). However, in South Africa, access to ART is currently constrained, particularly for those dependent on the public health system. The challenge over the next ten years will be to increase coverage rapidly in order ultimately to provide treatment opportunities to all of today's 6 million HIV-infected people.

Providing treatment for HIV/AIDS is a classic example of resource allocation in the face of scarcity. Although dealing with scarcity is the key rationale for the existence of the economics discipline, the allocation of resources to HIV-treatment is particularly charged for a number of reasons. Firstly, without treatment HIV-positive people will die a premature death. If no effective treatment were available, these deaths would be a tragedy, but because effective treatment is feasible, ongoing HIV-related deaths have been argued to be a moral outrage (Nattrass 2004). Secondly, HIV/AIDS is a new and complex disease. Because infection is associated with immune decline, HIV-treatment is not restricted to controlling the spread of the virus in the body, but also requires knowledge and capacity in the treatment of opportunistic and HIV-related infections and

[1] The Actuarial Society of South Africa (ASSA) is the pre-eminent body undertaking demographic modelling of the South African HIV-epidemic. The demographic data in this thesis have been extracted from the most recent version of the ASSA suite of models, the ASSA2003lite AIDS and Demographic Model (release 24 November 2005), as downloaded in December 2005 from www.assa.org.za.

events, many of which were previously rare. If HIV mainly affects prime-age adults[2] it can be expected to have a significant impact on the demand for health care in these age groups and for the country in total (Over 2004). Without a commensurate increase in health care supply, providing HIV-treatment can be expected to crowd out existing health care interventions. The allocation of resources to HIV-treatment is therefore not about changing the scale at which the HIV-treatment programme is operating, but about the creation of a new health care programme with associated training of health personnel, investments in infrastructure and medical equipment and development of drug procurement and delivery systems. Thirdly, treatment as currently proposed is relatively costly and the majority of this cost is associated with recurrent expenditure on medicines that need to be taken for the duration of a patient's lifetime. Although not a cure, treatment can be particularly effective and it is conceivable that patients could be maintained in care for years if not decades; if scale-up is successful, millions of South Africans will be enrolled in HIV-treatment programmes in the near future.

In recent years, the previous dearth of developing country health economics research into HIV-treatment has been replaced by a growing number of analyses of treatment resource needs (Bertozzi, Gutierrez et al. 2004; Gutierrez, Johns et al. 2004) and the cost-effectiveness of alternative strategies (Yazdanpanah, Losina et al. 2005; Bachman 2006; Badri, Cleary et al. 2006; Cleary, McIntyre et al. 2006; Goldie, Yazdanpanah et al. 2006). While these developments are to be welcomed, national strategic planning around HIV-treatment is not currently undertaken using an economics approach. If economic input is taken into account, this is normally the *ex post* consideration of the costs of scaling up the proposed strategy. While these costing studies are useful in assisting with the setting of annual treatment budgets, they provide no insight into the relative efficiency of alternative courses of action. On the other hand, while cost-effectiveness analyses can theoretically assist in the choice of efficient strategies, as normally presented, they provide no insight into the costs of scaling up and are relatively unhelpful to policymakers who are also concerned with affordability.

There have also been a growing number of commentaries published about equity in the distribution of health and health care to HIV-positive people (Daniels 2004; Macklin 2004; Rosen, Sanne et al. 2004; WHO and UNAIDS 2004; Bennett and Chanfreau 2005; Rosen, Sanne et al. 2005). While some of these have proposed a framework for equitable resource allocation, none have provided a method that enables the consideration of efficiency, equity or the costs of

[2] Defined as those aged 15 to 49.

scaling up. Whitehead (1992) argues that health inequity relates to differences in health status that are avoidable and unfair. Any definition of avoidable here must be a function of the availability of resources.

The contribution of this thesis is to bring an explicitly economics perspective to concerns about the distribution of health and health care to HIV-positive South African adults dependent on the public health care system. This is done through the creation of a framework that firstly considers what "the good" of health care could be to HIV-positive people, where the good could be defined in terms of utility within a welfarist perspective, health within an extra-welfarist perspective or capabilities and functionings, following Sen (1992). Having debated and defined this objective function, the framework secondly questions what claims HIV-positive people might have on this good. The concept of a claim is based on Broome (1991) who argues that claims are the object of fairness. If an HIV-positive person has a claim on health care, this is stronger than a preference or desire because a claim includes the notion of there being a duty on the part of society to provide this care. In considering these claims, the thesis broadens its scope from a focus on health care resources and health outcomes to consider the social context of the majority of HIV-positive South Africans; the impact of HIV-treatment on the macro-economy, households and social cohesiveness; and the impact of ART on the health of society which involves a balancing of the potential positive and negative treatment externalities.

The third section of the framework asks how the good might be fairly distributed. A variety of social choice rules are considered including those that aim to maximize the good, equalize the good or to trade-off between maximization and equalization. The impact of choosing one of these over another is illustrated through the introduction of a mathematical programming approach which allows the simultaneous consideration of efficiency, equity and the total costs of scaling up a variety of HIV-treatment interventions. Given however that a judgment of equity or fairness in the distribution of health and health care is subjective, this thesis refrains from promoting the specific views of the author about the appropriateness of one or other social choice rule and instead argues that the proposed framework has the potential to enhance explicitness in decision-making. Drawing on insights from procedural justice, it is argued that if this framework were to be used to elicit the distributive preferences of South African citizens with regard to distributions of the good, the resulting consequences of this process would be fair.

The remainder of this thesis has the following structure. Chapter 2 describes South African HIV-treatment policy and current clinical guidelines. While a full analysis of policy development processes is beyond the scope of the thesis, this background information is required to clarify concepts that will be discussed in later chapters.

In Chapter 3, theoretical and empirical literature is reviewed with a view to outlining the foundations of the decision-making framework. Chapter 4 reviews methodological literature and describes the approach that will be used in the analysis of empirical data.

Chapters 5-7 contain empirical results. Chapter 5 reviews and collates unit cost data from primary sources and from grey and published literature. Chapter 6 provides patient-level data on the lifetime costs and outcomes of alternative HIV-treatment strategies and interventions. Because HIV-treatment can be a long-term health care intervention, Markov modelling has been used to calculate these results. Models have been extensively justified and validated to reduce errors and enhance generalizability. This validation has included an extensive review of similar analyses in the empirical literature. Uncertainty has been assessed through first and second order Monte Carlo simulation (probabilistic sensitivity analysis) and simple sensitivity analyses.

Chapter 7 presents a mathematical programming approach to assessing technical efficiency, equity, the equity/efficiency trade-off and the costs of scaling-up HIV-treatment. This technique allows the decision-maker to consider the total health gains and the proportion of the population in need that can be treated within a range of possible HIV-treatment budget constraints.

Finally, Chapter 8 reviews and discusses the elements that have been covered in earlier chapters and proposes the use of procedural justice to assist in increasing legitimacy in HIV-treatment allocations. Chapter 9 concludes.

The new contributions of this thesis are empirical, analytical and conceptual. Empirical contributions include a wide review of unit cost data and one of the first primary-data driven cost-effectiveness analyses of HIV-treatment to be conducted in the developing world. Analytical contributions include the development of Markov models, the use of probabilistic sensitivity analysis in solving these models and in assessing uncertainty, and the development of mathematical programming to assess technical efficiency, equity and the equity/efficiency trade-off in HIV-treatment. The conceptual contribution relates to the development of a framework for

considering allocations of the good that has application to HIV-treatment and to health economics more generally. It is argued that the use of this method will increase the explicitness of HIV-treatment priority setting which in itself will contribute to greater fairness in resource allocation.

Chapter 2: South African HIV/AIDS policy and treatment guidelines

This chapter briefly outlines the government policy and HIV-treatment guidelines that were in place between 1994 and the end of 2006. These include the 1994 National AIDS Plan, the 2000-2005 National Strategic Plan, guidelines for the treatment of opportunistic and HIV-related infections, the "Operational Plan for Comprehensive HIV and AIDS Care", and antiretroviral treatment guidelines. At the time of writing, the government has also been in the process of finalizing a new National Strategic Plan that will set the framework for the implementation of HIV-related interventions across multiple sectors between 2007 and 2011. Because this document has not been finalized, it has not been included in this dissertation. The scope is limited to providing background information to allow for discussion of the economics of these policies and guidelines in later chapters.

Prior to 1994, South Africa was ruled by a minority white government under the National Party. From 1990, a series of discussions were held between the government and the African National Congress (ANC), which led to a negotiated settlement and the first democratic elections (Terreblanche 2002). The ANC won the elections, and Nelson Mandela became president of the country. As part of the negotiated settlement, jobs of civil servants were protected during the first five years of democratic rule, and a quasi-federal structure was established consisting of a national government and nine provincial governments. The former is responsible for collecting and distributing revenue, setting broad policy frameworks, and defining norms and standards for service delivery. The provincial governments are responsible for implementing policy and delivering services (Schneider and Stein 2001).

After 1994, the new ANC government adopted a National AIDS Plan for the country, which had been developed collaboratively between the ANC and the Department of Health prior to the elections. The plan embraced the sexual rights of women as a cross cutting theme and gave people living with AIDS a key role in AIDS policy development. It envisaged a coordinated network of AIDS policymakers at the national level and in the nine provinces within a multi-sectoral structure with implementing units in the ministries of health, education, welfare and defence (Schneider and Stein 2001). The plan was ambitious, but by 1998, AIDS policy implementation was argued to be characterised by a slow and hesitant start which lacked

coherence and continuity (Schneider and Stein 2001). Given the extent of restructuring that was required in the post-apartheid period, it appears that the AIDS plan had overestimated the implementation capacity of government. In May 2000, the government released the "HIV/AIDS/STD (sexually transmitted disease) Strategic Plan for South Africa 2000-2005" (Department of Health 2000a). Priority areas included prevention, treatment, legal and human rights and monitoring, research and evaluation. In order to improve treatment services provided through public health facilities, the following priorities were identified:

- Developing guidelines for treatment and care
- Ensuring an uninterrupted supply of appropriate drugs for the treatment of opportunistic and HIV-related infections
- Building capacity of health professionals to provide HIV-related care
- Improving prevention and treatment of tuberculosis (TB) and other opportunistic infections

An additional treatment-related objective was to investigate the use of antiretrovirals (ARVs) for the prevention of mother to child transmission (PMTCT) and to conduct research into the cost-effectiveness of non-ART treatment and prophylaxis (Department of Health 2000). At this time, there was no government policy with respect to the provision of ARVs for treatment of HIV.

In October 2000, the National Department of Health released its "Recommendations for the Prevention and Treatment of Opportunistic and HIV Related Diseases in Adults" (Department of Health 2000b) which contains guidance for the management of HIV-infected patients in public facilities. Services offered at the primary care level include HIV-testing, treatment and prophylaxis of opportunistic and HIV-related infections, STD treatment, palliative care, and TB treatment. Individuals who present with difficult clinical problems or who are seriously ill should be referred to hospitals.

During the presidency of Thabo Mbeki (1999 onwards) government involvement in HIV/AIDS policy and debate became increasingly fraught. It has been argued that the politicisation of the discourse about HIV/AIDS has sapped energy and attention away from the problems of policy development and implementation (Van Niekerk 2001). Government has been criticised for its engagement with dissident scientists who argue that HIV does not cause AIDS; the perceived obstructive attitude of national government towards antiretroviral treatment and prevention of

mother to child transmission interventions has led to numerous clashes and court cases between government and treatment advocacy non governmental organisations (NGOs) such as the Treatment Action Campaign. Given these events, a number of commentators have expressed concern at the perceived lack of government leadership with respect to HIV/AIDS (Campbell and MacPhail 2002; Cleary and Ross 2002; Benatar 2004; Nattrass 2004) where government leadership could be defined as personal public identification with the HIV/AIDS cause by influential politicians, as well as a willingness to mobilise funds and speed up implementation (Schneider and Stein 2001).

Nevertheless, in July 2002, a Joint Health and Treasury Task Team was established to investigate issues relating to the financing of an enhanced response to HIV, with a particular focus on antiretroviral treatment. After reviewing the report of the Task Team in August 2003, Cabinet instructed the Department of Health to develop a detailed operational plan on ART (Department of Health 2003).

This plan set a target of establishing at least one ART service point in every health district by the end of the first year of implementation. By 2009, the aim was to deliver "equitable access" by providing service points within each local municipal area. The provision of treatment was to be located within broader primary health care provision, in an attempt to avoid a vertical approach to ART. A vertical approach refers to the delivery of health services through largely free-standing programmes (Oliveira-Cruz, Kurowski et al. 2003). While a vertical approach might increase the speed of provision of ART services, this approach can divert resources from other public health services. It is hoped that if ART is located within the district health system, other primary health care services might be strengthened (Loewenson and McCoy 2004; Stewart and Loveday 2005).

According to the ART Operational Plan, the provision of HIV-treatment is to be guided by the following principles:

Quality of care:

- Care should be of the highest available quality, including proper diagnosis, counselling, treatment of opportunistic infections, other preventive and supportive strategies such as nutritional advice and supplements, traditional and complementary medicines, and ARVs.
- Attention should be given to the proper use of ART in order to minimize toxicities, adverse events and the development of drug resistance.

- Care and treatment protocols will be based on international best practice. Accreditation procedures will help to ensure that facilities that are approved to deliver ART are of good quality and observe the highest standards of care. Extensive training and certification of health professionals to deliver ART will be carried out.
- A monitoring and evaluation system will be developed to allow the ongoing scrutiny of quality of care issues.

"Universal"[3] access to care and treatment:

- The operational plan attempts to realize equitable implementation by achieving a balance between areas that can readily implement the programme and those that need additional investment to upgrade their health capacity.

Strengthening the national health system:

- Upgrading staffing by improving HR capacity and incentives to recruit and retain health professionals in historically underserved areas
- Upgrading facilities by refurbishing and building new facilities
- Developing patient health information systems
- Upgrading drug procurement and distribution systems
- Upgrading management systems
- Consolidating the National Health Laboratory Service (NHLS)

Providing a comprehensive continuum of care and treatment:

- This continuum includes ongoing medical services to provide treatment for HIV-related and opportunistic infections, antiretroviral treatment, an extensive nutrition intervention, traditional healing, counselling, adherence support groups, community mobilisation efforts to reduce stigma and discrimination, patient transport, home and community based care, and palliative care.

Sustainability:

[3] While the term "universal access" is frequently used in HIV policy debates, this is a misuse. Universal access to HIV-treatment means that the entire population would have access irrespective of whether they needed it. For the purposes of this thesis, it is assumed that universal access means equal access for HIV-positive people who meet policy-prescribed medical criteria.

- Once people enter a "comprehensive treatment" programme, this must be sustained. The drugs, tests and human and physical infrastructure required to sustain treatment are costly. Despite the potential financial burden of a programme of this nature, the majority of the finances will come from the fiscus, with lesser support from donor resources.

In 2004, the National Department of Health released the national ART guidelines (Department of Health 2004) which established the following criteria for initiating ART in adults:

Medical criteria:

- CD4[4] count <200 cells/μl irrespective of World Health Organisation (WHO) stage (see appendix for details of the staging system (WHO 1993))

OR

- WHO Stage IV (also known as AIDS) irrespective of CD4 count

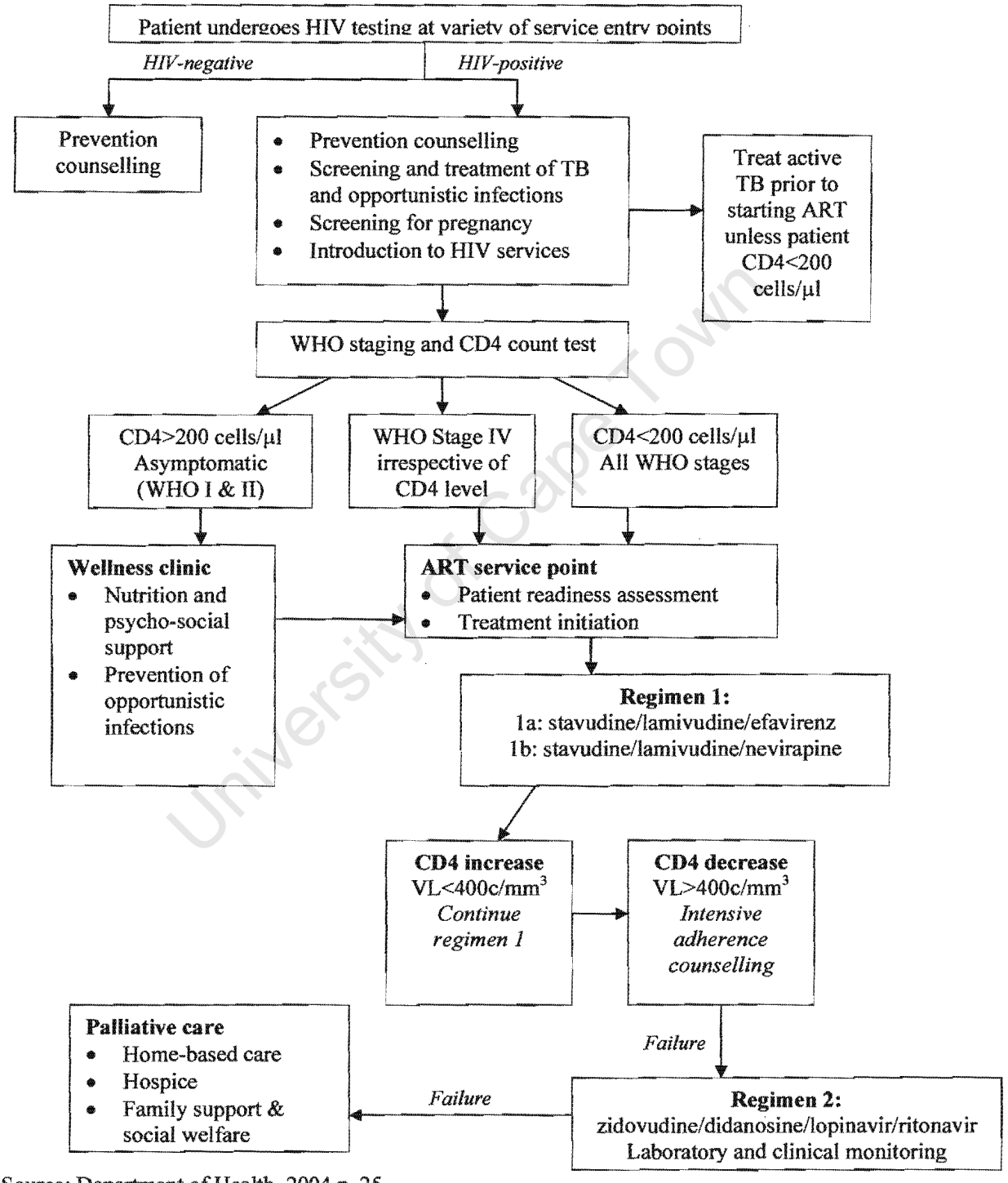
Psycho-social considerations (not exclusion criteria):

- Demonstrated reliability (e.g. having attended three or more scheduled visits to an HIV clinic)
- No active alcohol or substance abuse
- No untreated active depression
- Willingness to disclose HIV-status to at least one friend or family member or to join a support group
- Acceptance of HIV status, understanding of consequences of HIV status and the role of ART
- Ability to attend services on a regular basis (transport could be arranged for those in rural areas)

According to the Western Cape "Antiretroviral Treatment Protocol" (Provincial Government of the Western Cape 2004), if patients meet these clinical and psycho-social criteria, they are referred from a service entry point (for example the local primary care clinic or TB clinic) to the

[4] The CD4 is a particular type of immune cell in the human body and the level of the CD4 gives an indication of the status of the immune system. A low CD4 count indicates that the immune system has declined with the implication that the risk of the development of symptomatic disease and the risk of death is increased.

Figure 1: Management of a patient on Antiretroviral Treatment



Source: Department of Health, 2004 p. 25

To summarize, this chapter has outlined the current policy framework that guides the public sector response to HIV-treatment. By March 2005, the government had succeeded in its aim of opening one ART site in each health district (Stewart and Loveday 2005) and by the end of 2006, approximately 200,000 adults and children had been started on ART in the public sector (Mark Blecher, National Treasury, personal communication). This treatment programme has therefore become one of the largest in the world within a short space of time. Despite this success, later chapters will show that the percentage of need that is being met is still low. Unless technology changes, increasing access to ART will continue to dominate the health care landscape in South Africa, given that over 5 million people are currently infected and overall prevalence in the country continues to grow.

Chapter 3: Developing a framework for distributing health and health care

1 Introduction

This chapter proposes a framework for social decision-making regarding the allocation of the benefits of health care, referred to as "the good" of health care, to HIV-positive adults in South Africa who are dependent on the public health care system. The problem concerns the allocation of these resources within an exogenously fixed government budget earmarked for HIV-treatment, and does not address the wider problem of allocating health care resources between HIV and other disease areas. This framework makes important conceptual contributions both to the economics of HIV-treatment and health economics in general in that it debates and defines the good, considers the determinants of the claims that HIV-positive people might have on this good, debates a number of social choice rules for distributing the good and finally argues that if decision-making takes place within a procedurally just approach, the resulting distributions will be fair. Three questions are posed in the development of this framework:

- What is the good that should be distributed?
- To whom should the good be distributed? What are the personal characteristics that are the bases for claims on the good?
- How (i.e. on what basis and to what level or what quantity) should the good be distributed?

Answering the question "what is the good" requires an identification of the value or benefit of health care to HIV-positive people, as discussed in Section 2. Section 3 considers how personal characteristics might allow one to consider people to be equal or unequal in terms of their claims on the good. Section 4 examines a number of different social choice rules that could serve as the values or the principles on which distributions are based. However, given that reasonable people could disagree on the choice of such a rule as well as a number of issues regarding the claims of HIV-positive people on the good, the section also outlines the concept of a procedurally just process in decision-making. Section 5 summarises.

2 What is the good?

At one level, the question posed in this section is "what is the good (value or benefit) of health care"? But at another level, the question is Sen's (1982) more fundamental one about what ought to be in the evaluative space in distributive justice where "principles of distributive justice are normative principles designed to allocate goods in limited supply relative to demand" (Lamont 2003). In this view, the good is not health care per se - it would also be unusual to say that health care in itself has sufficient moral weight to place it into the evaluative space of distributive justice. In the following sections, utility or preferences, capabilities and functionings, and health are each examined as potential candidates for the good. However, in order for any of these to be a coherent measure of the good for use in social choice, it is necessary for particular types of information to be available.

2.1. Information about the good in social choice

Arrow was the first to prove that it would be impossible to create a rule to guide distributive decisions for society if only limited information was available about the good[5]. Put formally, Arrow showed that if it was desirable to have, for each logically possible configuration of individual preferences over feasible social states, an ordering of social preferences that satisfied collective rationality and independence from irrelevant alternatives, then one would have to violate the weak Pareto principle or non-dictatorship (Arrow 1973).

Imagine an isolated individual. His or her personal skills, qualities and the physical environment jointly limit the range of possible actions. Assume alternative actions are mutually exclusive. Thus the basic question for the individual is the choice of actions. The value system is the rule an individual uses to choose between mutually exclusive actions. This value system is also assumed to have a particular structure – namely being derivable from an ordering. The assumption that an ordering exists over choices means that consistency assumptions are imposed on the value system. The first is completeness, which means that for a choice between two actions or states, one is preferred or there is indifference. The second assumption is that of transitivity, where for

[5] Arrow discussed his findings in terms of preferences or utility. In this section, utility is replaced by the term "the good" because this discussion is relevant to all candidates for the evaluative space that are considered in this chapter.

three alternatives x, y and z, if x is preferred or indifferent to y and y is preferred or indifferent to z then x is preferred or indifferent to z. Finally, it is assumed that the choice is determined by the ordering – in other words, if there is an alternative that is preferred to another, this one is chosen – this is known as the condition of reflexivity. To summarise, a complete, reflexive, transitive relation is called an order. The individual is also assumed to play a central role in *social choice*. As before, preferences over social states can also be ordered.

The problem of social choice is to find a rule for aggregating or allocating the good over alternative social states. This process can be conceptualised as one of forming a constitution – a social choice rule which associates a social function to each possible set of individual orderings. The problem of distributive justice is therefore that of constructing a constitution or social choice rule, where each rule must obey the following reasonable conditions:

- Collective rationality: for any given set of orderings, the social choice rule is derivable from an ordering. Others replaced this with the principle of unrestricted domain which means that the constitution or social choice rule must make social prescriptions for any conceivable preference profile of society (Sen 1982; Roemer 1996). One implication is that one cannot assume that preferences will exclude sadistic preferences such as racism (Elster 1992).
- Pareto principle: if alternative x is preferred to y by every individual according to his ordering, then the social ordering also ranks x above y.
- Independence of irrelevant alternatives: the social choice depends only on the orderings of individuals with respect to the alternatives in that environment. For instance, when choosing among candidates in a presidential election, the choice is only among candidates who are actually running for office – other non-running candidates are irrelevant alternatives. Alternatively, the social ranking of x and y should not change because some individuals who once preferred y to z now prefer z to y (Elster 1992).
- Non-dictatorship: there is no individual whose preferences are automatically society's preferences without regard to the preferences of all other individuals.

Consider the majority voting system as a way of choosing between alternatives. Here, the social choice rule states that any set containing the majority of voters is the one that chooses the social state. Imagine there are three alternatives, x, y and z. One-third of the voters prefers x to y and y to z, one-third prefers y to z and z to x and one-third prefers z to x and x to y. Then x will be preferred

to y by a majority, y to z by a majority and z to x by a majority. Generally, there can be no social choice rule simultaneously satisfying the conditions of collective rationality, the Pareto principle, independence of irrelevant alternatives and non-dictatorship unless there is more information about the good (Roemer 1996).

Depending on the amount of information that is available, infinitely many or only one potential function would be representative. For instance, if the function only needs to represent the ordering of preferences on the good, then there are a multitude of acceptable versions – namely all transformations of the function as long as the transformations are strictly increasing. However, if the function must directly represent the preference ordering and if there is more information about the states represented in that ordering, then only one possible function would be representative.

The level of information required by a function can be described by the number of acceptable transformations of the function that are allowable. If ordinal information is available, one can only speak about the order of preferences on the good – for example x is preferred to y but one cannot speak about the intensity of the preferences. To do this, one would need cardinal information. Comparability allows one to compare the good between people.

- The good is cardinally measurable and fully comparable if and only if the group of transformations is limited to the identity transformation[6].
- The good is ordinally measurable and noncomparable if the function can be transformed by the group of strictly increasing functions. There are many possible representations of the function.
- The good is cardinally measurable and noncomparable if the function can be transformed by any increasing affine transformation[7].
- The good is ordinally measurable and fully comparable if the function can be transformed by any increasing transformation. This is similar to ordinally measurable and noncomparable except that the functions of all people are transformed in the same manner so as to preserve the interpersonal ordering of states.

[6] The identity transformation takes the form $f = I$

[7] Affine transformations are defined to be of the form $f = (\alpha^1 + \beta^1, \ldots, \alpha^H + \beta^H)$ where β^H are any real numbers and α^H are any positive real numbers.

- The good is cardinally measurable and unit comparable if it is meaningful to talk about differences between states for all citizens, but the absolute levels of the good for individuals is unknown[8]. In other words the good has interval but not ratio scale properties.

To summarize, if information on the good is ordinal and noncomparable, the set of transformations is as large as possible, but if more information is needed, the number of transformations becomes restricted. If full information is available, only one representation is possible. The type of information that is available about the good has implications for the types of social choice rules that can be used to make distributive judgements between states. To summarise:

- As shown by Arrow, it is not possible to construct a social choice rule if there is only ordinal information about the good.
- Similarly, there is no social choice rule if it is meaningful to talk about the intensity of the good (cardinal information) but not to compare between people (Sen 1982).
- A social choice rule that seeks to maximise or equalise the good between people is possible if the good is cardinally measurable and fully comparable.

The key message from this section is that the less information there is about the good, the fewer coherent social choice rules are possible.

2.2. Utility

A common approach in economics is to place utility or preferences in the evaluative space. Here it is assumed that people seek to maximise their utility, which is determined by the bundle of goods and services[9] that they possess. To do so, they purchase their ideal bundle based on their taste for alternative goods and services, on the prices of these alternatives and on their budget constraint (Rice 2003). Under the revealed preferences approach (also known as ordinal non-comparable utility), a person is said to prefer some bundle of goods x to another bundle y if she

[8] Transformations take the shape of $f = (\alpha + \beta^1, \ldots, \alpha + \beta^H)$ where β^H are any real numbers and $\alpha > 0$. Note that this set of transformations is therefore smaller than for affine transformations.
[9] Goods include tangible items such as medicines while services are intangible such as the treatment advice of a health professional (Lipsey, Courant et al. 1993). Thus health care includes goods and services.

chooses *x* from a set of options which includes *y*. Utility is not a measure of any particular quality, but simply represents a choice (Sugden 1993). An advantage of this approach is that it does not require any understanding of the motivations for choice or the psyche of the individual. As a result however it is then not feasible to attribute any value to the choice of *x* over *y* (Sugden 1993) or to use utility as a measure of the good in distributive justice.

The classical utilitarians viewed utility as a mental state that could be measured and compared between people - it was meaningful to state that health care could confer utility (e.g. happiness) on a person or that one person had a higher level of utility than another (Viner 1973). This allowed for the construction of coherent social choice rules.

However, even with additional information, placing utility in the evaluative space in distributive justice has a number of shortcomings. One is that social welfare is a function of personal utilities – this is also known as welfarism. This means that in making a choice between social states, no information is used other than the utility information related to those states; it does not matter what lies behind these choices (Sen 1982). An illustration from Sen (1982 p. 339-340) below is used to clarify what this means.

Consider a set of three social states *x*, *y* and *z*. These states have the following utility levels for persons 1 and 2:

	x	*y*	*z*
Person 1's utility	4	7	7
Person 2's utility	10	8	8
Sum Total	**14**	**15**	**15**

Imagine that the following non-utility information is also available: in *x* person 1 is hungry and 2 is eating as much as he likes. In *y* person 2 has given some food to 1. While 2 is worse off than in *x*, 1 has more utility and the sum total is higher. If one argues that the best social state is either one in which utility is maximised or utility is closest to being equal for persons 1 and 2, state *y* would be preferred to state *x*. Consider *z* – here 1 is as hungry as in *x*, and 2 is eating as much as he likes. However, person 1 is a sadist and is allowed to torture 2 – 2 suffers but his utility loss is less than the gain to 1. Welfarism would argue that if *y* is preferred to *x*, then so too is *z*. Because welfarism encourages an exclusive focus on consequences, other considerations are irrelevant

unless they have an impact on personal utility (Richardson and McKie 2005). A related issue is that focusing on utility can discriminate against people (such as person 2, the torture victim) who are deprived but have adapted to their position in life and have managed to retain a relatively high level of utility. If someone's desires are muted, their happiness from small improvements in their situation may be disproportionate to the benefit judged from an alternative perspective (Sen 1992).

Sen (1999) also argues that placing utility into the evaluative space is problematic because this assumes that choice is motivated solely by self-interest. While self-interest is an important motivation, it is also common for action to have a social component that is beyond pure self-interest – it is counterpreferential. Sen distinguishes between sympathy and commitment. Feeling sympathy about the suffering of others involves self-interest. However, it might also be the case that one is willing to make utility-diminishing sacrifices for others that go beyond sympathy. Sen argues that this is motivated by commitment. For example, if being concerned about the plight of AIDS orphans makes a person sufficiently unhappy that he/she chooses to spend weekends volunteering in an AIDS orphanage, this is an example of sympathy, of acting on self-interest. If however the presence of AIDS orphans leads to a determination to change a system that is felt to be unjust (or more generally this determination is not fully explained by diminished personal utility) then this is an example of commitment. Thus committed behaviour involves self-sacrifice in a way that self-interested behaviour would not [10].

2.3. Capabilities and functionings

As an alternative to utility, Sen (1992) argues that "the good" should be viewed in terms of capabilities and functionings, which "reflect a person's freedom to choose between alternative lives" (Sen 1992 p. 83). Functionings are the valuable activities and states that make up a person's well-being, such as having a healthy body and having a good job. Capabilities are the alternative combinations of functionings that are achievable – while one may not pursue all possible functionings, the freedom to pursue alternative paths is also important. For example, an HIV-positive person can have the same level of health (functioning) as an HIV-negative person, but less freedom to pursue alternative paths. A benefit of the functioning and capability approach as opposed to utility is that it offers a way of judging personal advantage that is not focussed on

[10] Committed behaviour should not be construed as being solely positive – certain committed behaviour can be harmful to the greater good in the same way that self-interested behaviour can be harmful.

the way in which people transform health care, for example, into utility and therefore does not penalise deprived people who can maintain their utility levels through small mercies. Sen's approach therefore focuses attention towards whether all people receive the same benefit (capability) from an intervention, irrespective of their capacity to benefit, thereby acknowledging that some people would require additional inputs to reach a given capability level.

2.4. Health

While capabilities and functionings could be defined more generally as the good in distributive justice, to do so risks having capabilities encompassing all human activity. When focusing on the distribution of health care resources, Culyer (2001) has argued for an "extra-welfarist" perspective which places health into the evaluative space. He argues that the distribution of health care is important because it serves a significant end – the individual's health. Health is a constituent of functionings - good health provides people with the freedom to pursue the states and activities that they value while poor health can severely compromise a person's freedom to choose between alternative life paths (Culyer and Wagstaff 1993). The distribution of health care that postpones death or decreases disease and other negative influences on the quality of life is of significance if society has agreed that all members should have the opportunity to flourish. Although health care provides other benefits, such as providing information, reassurance and a variety of other services that are marginally connected with health, Culyer argues that health should be in the evaluative space as opposed to these other outputs of health care because "health is *important* in ways that the other needs served by health care are not" (2001 p. 276). The good for HIV-positive people in South Africa will therefore be argued to be health.

2.4.1. Defining health

Focusing on health as the good for HIV-positive South Africans requires a definition of health that is appropriate to the topic and lends itself to measurement. Unless one can measure health, one cannot assess the outcomes of alternative health care allocations. Evans and Stoddart (1994) argue that most activity in the health care system reflects a view that health is an absence of disease or injury. Under this definition, the appropriate policy response to ill-health is to provide health care only if effective care exists. An advantage of this definition is that it lends itself to measurement and quantification for example through survival or death and incidence of

diagnosable conditions. However, the World Health Organisation has defined health as "a state of complete physical, mental, and social well-being, and not merely the absence of disease or injury" (in Evans and Stoddart 1994 p. 28). This definition of health, similar to the capabilities approach, implies that health is affected by all human activity. The determinants of health would therefore include a range of factors such as the targeted use of health care services, genetic endowments, environmental sanitation, adequacy and quality of nutrition, housing, stress, supportiveness of the social environment, self esteem, sense of personal adequacy and self control. While it is difficult to argue against the validity of the World Health Organisation definition, the practicality of it for the purposes of health policy is problematic, especially if health has to be measured.

Evans and Stoddart (1990) therefore propose an alternative framework. This differentiates between disease as labelled by the health care sector, health and function as experienced by the individual and the World Health Organisation definition of health which they rename wellbeing. For example, two HIV-positive people with a CD4 level of 231 cells/μl might have the same disease and severity of disease as labelled by the health care system, but might experience significantly different levels of health and function. Thus health in this thesis is defined narrowly but from the perspective of the individual as "the absence of illness or injury, of distressing symptoms or impaired capacity" (Evans and Stoddart 1994 p. 47). While the disease as labelled by the health care system will have a significant impact on the individual's perception of her health, the extent to which a disease will cause illness will differ from person to person. Similarly, illness will have a negative impact on wellbeing, but will not be the only factor that is influential.

2.4.2. Measuring health

Following on the definition of health proposed above, this section outlines the theoretical and methodological approaches and challenges in measuring health. Such measurement is complex firstly because health is culturally determined – it is perceived differently by different people (Mooney 2002). Also, health is multidimensional (i.e. it is not solely about being alive, but is also about the quality of that life) and in order to avoid impossibility results (see above) measurement must enable a comparison to be made between people and a distinction to be made between different intensities of health – a cardinal scale is required. If health is measured on an ordinal scale, it is only meaningful to rank health states as better to worse. An example of an ordinal scale

for HIV is the WHO staging system. With this it is possible to say that WHO Stage III is better (i.e. healthier) than WHO Stage IV but it is not possible to say how much better. When a cardinal scale is used, it is possible to make this judgement. If an interval scale is used (one type of cardinal scale) it would be possible to say that the move from say WHO Stage I to WHO Stage II was valued the same as the movement from WHO Stage II to III, but it would not make sense to say that WHO Stage I was three-times better than WHO Stage III. However, if a ratio scale was used (another type of cardinal scale) such statements could be made validly. We have here an indication of the stringent informational requirements of ratio-scale cardinal health state measurement.

In health economics, one method for measuring health on a ratio scale is the Quality Adjusted Life Year (QALY). The QALY model suggests that one year of life in perfect health has a value of one while life years in less than perfect health have values of less than one. Generally it is assumed that death has a value of zero and that certain health states that are considered to be worse than death have negative values. In keeping with ratio scale properties, moving from 1 QALY to 2 QALYs followed by death is the same as moving from 11 QALYs to 12 QALYs followed by death and 12 QALYs are six-times better than 2 QALYs.

Constructing the quality of life component of the QALY requires measurement of health-related quality of life (HRQoL). This is often done by administering a questionnaire to individuals in a health state of interest to the evaluation (unless the subjects are children, or emotionally or cognitively impaired individuals in which case proxy respondents such as family members could be used (Torrance and Feeny 1989)) or to the general public. This includes questions about an individual's functioning in "domains" that are considered important in assessing health states, including "physical/mobility function, emotional/psychological function, sensory function, cognitive function, pain, dexterity, and self care."(Torrance and Feeny 1989 p. 559-560). A person's perception of her level of functioning in each domain provides a subjective description of this person's health state. The measurement of HRQoL therefore amounts to creating ordinal rankings over different domains that collectively make up a health state.

One shortcoming of using patients' preferences to rank domains is that if patients have adapted to their conditions and have a relatively optimistic view on their health state, they might overstate their level of health and function. A similar point was made about using utility information in the case of the perennially deprived. If the general public valued a health state differently from

patients, might one be tempted to use their values instead? In answering this, one needs to bear in mind the entirely subjective nature of health state valuation – it is a matter of personal opinion. To judge a health state as over or under valued is fraught given the multidimensional nature of health. It would therefore be difficult to justify the use of non-patient responses to rank domains given that only patients have direct experience of the health state

The next step in HRQoL measurement is to transform these ordinal rankings into a composite cardinal value for each health state. While ordinal ranking of domains is usually done by patients, the cardinal valuation of health states is often a separate process, and can be done by the general public, health care professionals, policymakers or patients (Torrance and Feeny 1989). It is usually argued that it is most appropriate to use the values of the general public if the QALY data are being used in resource allocation decisions for society (Brazier and Fitzpatrick 2002; Dowie 2002; Feeny 2002).

There are three common methods of valuation: the time trade-off (TTO), the standard gamble (SG) and the visual analogue scale (VAS). The time trade-off method values health states by considering the trade-off between length of life and quality of life – this trade-off is inherent to the QALY concept. Using this method, the value of a health state can be estimated by assessing the number of years the person is indifferent between living in good health versus living for a longer period with a lower health-related quality of life (Torrance and Feeny 1989). A shortcoming of this method is that it assumes neutral time preference when it might be more likely that people value benefits that accrue sooner higher than those that accrue later (Broome 1993). However, others have argued that respondents might build this discounting into their trade-offs with the implication that the time preference of respondents is adequately captured (Martin, Glasziou et al. 2000).

The standard gamble method is grounded in von Neumann Morgenstern (VNM) expected utility theory (Torrance and Feeny 1989; Broome 1993). Proponents of this method often argue that it is a gold standard because it allows an assessment of health-related utility (Torrance and Feeny 1989). Although this thesis interprets the good as health and not as health-related utility[11], the ideas underlying this approach will nevertheless by outlined. Von Neumann and Morgenstern

[11] In essence, this means that the QALY is viewed as an objective measure of a person's health status, instead of a subjective measure of the utility associated with that health state.

(1944) developed a theory of utility under uncertainty to describe how a rational person *ought* to behave when taking decisions under uncertain outcomes. Its axioms are:

- Preferences exist and are transitive – for any pair of risky prospects, y is indifferent or preferred to x and for any three risky prospects, if y is indifferent or preferred to x and x is indifferent or preferred to z then y is indifferent or preferred to z.
- Independence – an individual should be indifferent between a two-stage risky prospect and its probabilistically equivalent one-stage counterpart.
- Continuity of preferences – if there are three outcomes such that x is preferred to y which is preferred to z, there is some probability p at which an individual is indifferent between outcome y with certainty and receiving the risky prospect of outcome x with probability p and outcome z with probability $1 - p$.

The standard gamble method is a direct application of the continuity of preferences axiom (Torrance and Feeny 1989). A paired comparison between two alternatives is described to the respondent. Choice A is to accept the respondent's current health state while choice B involves a gamble on a treatment with a probability of death or of perfect health. The probability is varied until the respondent becomes indifferent between choices A and B. Following VNM axioms, the value of the health state is equal to the indifference probability (Torrance and Feeny 1989). A shortcoming of this method is that it assumes that an individual is risk neutral. If a person is risk-averse or risk-loving, health state valuations will be biased (Torrance and Feeny 1989; Broome 1993).

According to Torrance and Feeny (1989) even if individuals were risk neutral, QALYs would not be equivalent to expected (health-related) utility unless:

- The two attributes of quality and quantity were mutually utility-independent (preferences for gambles on the one are independent of the amount of the other).
- The trade-off of quantity for quality exhibited the constant proportional trade-off property (the proportion of remaining life that one would trade off for a specified quality improvement is independent of the amount of remaining life).
- The single-attribute utility function for additional healthy life years were linear with time (for a fixed quality level one's utilities are directly proportional to longevity – risk neutrality with respect to time).

Given that the above assumptions are unlikely to hold, the QALY model does not hold under expected utility theory (Broome 1993; Brouwer and Koopmanschap 2000). This assertion has been validated in empirical tests (Doctor, Bleichrodt et al. 2004). However, the QALY model might hold under non-expected utility theories - such as rank-dependent utility and prospect theory. Rank-dependent utility deviates from expected utility because it allows probability weighting and prospect theory deviates in that it allows probability weighting and loss aversion. In brief, probability weighting relaxes the assumption of risk neutrality, while loss aversion allows greater sensitivity to losses than to gains. Doctor, Bleichrodt et al. (2004) performed an empirical test of the QALY model under expected utility, rank dependent utility and prospect theory. Their findings suggest that subjects do not behave according to expected utility theory, but that the QALY model holds under either rank-dependent utility or prospect theory. The implication is that probability weighting and loss aversion should be taken into account when valuing health states. However, to the best of the candidate's knowledge, this method has yet to be used in large population surveys to establish health state valuations, which means that it is not yet available for use in empirical research.

The final method for valuing health states is the visual analogue scale (VAS). Here a respondent ranks a health state by drawing a line on a scale such that the intervals between different health states correspond to the respondent's preferences for these states. There are two measurement biases associated with the VAS. Context bias reflects the fact that the VAS score for a state depends on how many better or worse states were evaluated by the respondent in the same sitting. The value of a state might be depressed if it is valued at the same time as a number of better states. End aversion comes about because respondents are reluctant to use the extreme categories on the scale. On the other hand, VAS has benefits in that it is generally argued to be easier to administer and easier for subjects to understand (Torrance, Feeny et al. 2001).

The final key assumption that is required to operationalise the QALY approach in resource allocation is inter-personal comparability, which is normally interpreted as meaning that a QALY gained is of equal value to everyone (Williams 1996). When resource allocation concerns the entire range of diseases and interventions in the health sector, the acceptability of this assumption is questionable because widely divergent groups of people are compared against each other solely in terms of their health gain. However, the assumption is more acceptable if the comparison is restricted to HIV-positive adults who are dependent on the public health system, as is the case in

this thesis. It can also be argued that it is egalitarian to treat each HIV-positive person's health gain equally (Williams 1996).

To summarize, the construction of a QALY requires a number of steps. In the first step, patients create ordinal rankings on their current health state across a number of domains such as pain, usual activities and anxiety. In a separate step, a population sample transforms these ordinal rankings into a cardinal measure of each health state with ratio scale properties, using a time trade-off, standard gamble or visual analogue scale method. Resultant values are multiplied by the number of life years spent in each health state to calculate QALYs. Finally, for the purposes of resource allocation it is usually assumed that a QALY is of equal value to everyone.

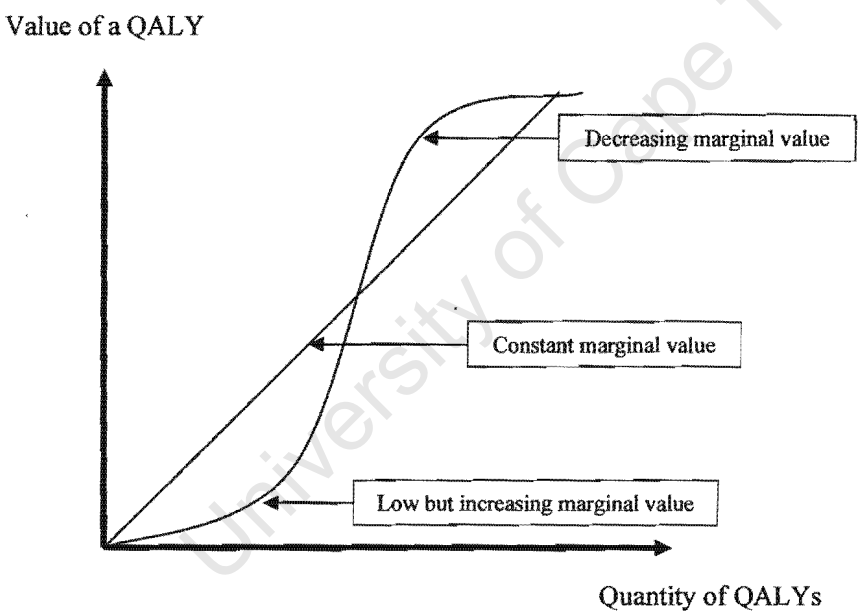
Before concluding, it is necessary to ask how good the QALY is as a measure of health – does it measure what it is supposed to measure, is it reliable and reproducible by different people and can one relate it to some of the variables that are relevant to the policy question?

In this dissertation, the QALY is intended to measure health as defined from the point of view of the HIV-positive person as the absence of illness or injury, of distressing symptoms or impaired capacity. This definition has a number of similarities to the health domains included in HRQoL questionnaires such as pain and anxiety. In addition, these questionnaires are answered by patients in keeping with the definition of health being from the point of view of the individual. However, it should be borne in mind that HRQoL is frequently assessed through generic questionnaires with specific domains that might not be appropriate to all settings and that different cultures might have different conceptualisations of health that may not be adequately reflected. In addition, while health is defined from the perspective of the individual, the construction of a cardinal measure on each health state is done by a population sample. This is however defensible given that the QALY is used in society level resource allocation decisions. But what is less defensible are the assumptions that the QALY is a cardinal measure of health with ratio scale properties and that the QALY model does not normally allow for increasing or decreasing marginal value of health gains. This means that a QALY gained in the context of limited health gains has the same marginal value as a QALY gained after greater health gains.

Figure 2 illustrates the potential relationship between the value of a QALY and the quantity of QALYs for one individual. The assumption used in the QALY model is that the marginal value of a QALY is constant, as illustrated by the straight-line. In other words, no matter how many

QALYs have been gained, each additional gain has equal value. However, empirical tests have suggested that this function might be concave or become increasingly concave with increased survival time (Martin, Glasziou et al. 2000). However, it is hypothesised here that this function might be S-shaped. If this is the case, QALYs have a low but increasing marginal value at a low quantity. Increasing marginal value continues until a certain level of QALYs is attained, after which point the curve inflects and becomes concave to indicate that each additional QALY has a decreasing marginal value. This reflects what Gafni and Torrance (1984) have called the quantity effect, where the relative desirability for additional units of a QALY would initially increase, but later decrease once a certain amount had been achieved.

Figure 2: The marginal value of a QALY - increasing, decreasing or constant



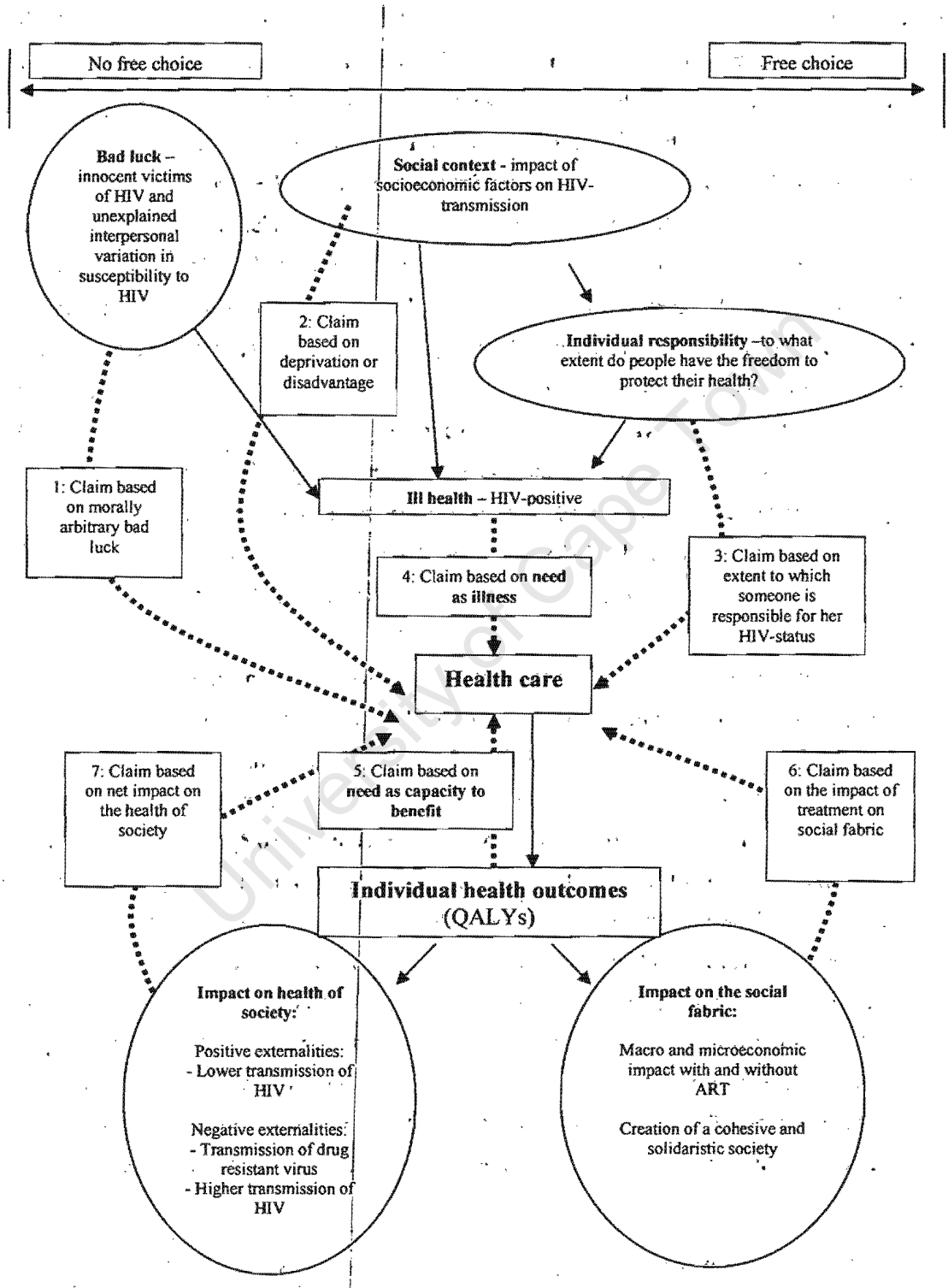
Is it possible to relate the QALY to some of the variables that are relevant to the policy question? Dowie (2002) has posed a similar question in terms of decision validity – does the QALY measure what is necessary for the decision context? Does it embrace the full range of outcomes that could result from all alternative treatment interventions under consideration? Here the answer is a tentative yes, if QALYs are used cautiously. The QALY model, as in other models, is a simplification of reality that should be used in a way that encourages thought and draws attention to areas of disagreement between the model and reality. Finally, the extent to which QALY health

state values are reproducible will be assessed in later sections by reviewing the empirical literature on HIV/AIDS health state values.

3 To whom should the good be distributed?

This section covers the various claims that HIV-positive people might have on the good. The notion of a claim is based on Broome (1991) who argues that claims (as opposed to other reasons) are the object of fairness. If an HIV-positive person has a claim to health care, this is stronger than merely having a reason for wanting health care because a claim includes the notion of there being an obligation on others to provide care. A number of criteria could be the constituents of claims, including the social context in which most HIV-positive people live, whether or not people have personal responsibility in their HIV-status, need for health care, the impact of HIV-treatment on the broader health of society, and the impact of HIV/AIDS and treatment on the social fabric. These claims on the good are summarised in Figure 3, which has been adapted from Olsen, Richardson et al. (2003). Different claims will be discussed in the following sections.

Figure 3: A framework for considering to whom the good should be distributed. Solid arrows show the causes of illness and consequences of health care. Dotted arrows show claims on the good.



4.1. Consequentialist theories

Theories of justice are consequentialist if they recommend that policies should be evaluated in terms of their consequences. A number of theories are consequentialist in nature, including utilitarianism, which focuses on maximising outcomes, and egalitarian theories that propose equality in the space of outcomes. For example, Stinnott-Armstrong (2003) argues that Sen's theory of capabilities is consequentialist in the space of capabilities, although it should be noted that capabilities also include procedural aspects in that the focus is on access to capabilities as opposed to what is done with the capability.

4.1.1. Health maximisation and economic evaluation

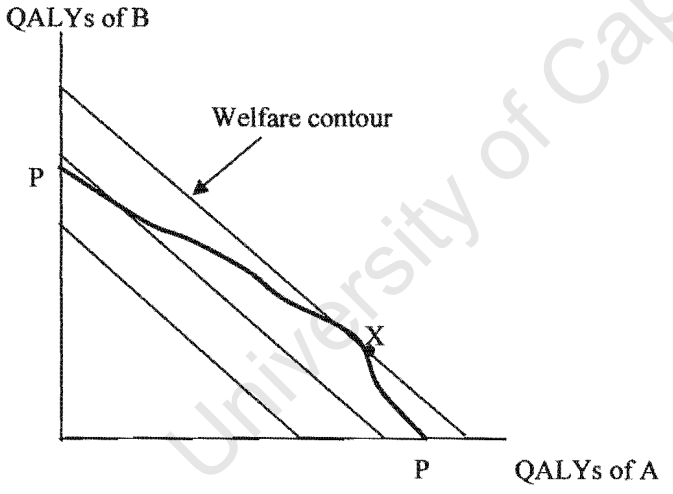
Health maximisation as a social choice rule is concerned with finding the allocation of health care resources that maximises health. Conceptually, this approach is similar to classical utilitarianism. Utilitarians argue that society should seek to achieve the greatest good for the greatest number (Viner 1973; Roemer 1996). There are many different brands of utilitarianism, but they share three features: consequentialism (policies are evaluated by their consequences), welfarism (consequences are evaluated in terms of the **individual** utility that they generate) and sum-ranking (the social choice rule maximises the sum total of individual utilities). As explained, ordinal utilitarianism differs from classical utilitarianism in that the former rejects sum ranking because it only allows ordinal non-comparable information on preferences, whereas the latter allows preferences to be cardinal and interpersonally comparable (Rice 2003).

While utilitarianism is traditionally associated with maximising utility, some health economists have argued for an extra-welfarist perspective which considers the non-utility aspects of welfare (Rice 2003). In health economics, these non-utility aspects are often argued to relate solely to health outcomes (Culyer and Wagstaff 1993). So, the extra-welfarist perspective maintains certain elements of utilitarianism such as sum-ranking and consequentialism, but replaces individual utility with individual health outcome.

Following Williams and Cookson (2000), health maximisation can be presented in terms of an optimisation problem where a social choice rule is used to choose between options in an opportunity set. The opportunity set can be presented as a health or QALY possibilities frontier,

which presents the different distributions of health between people. Any options on and within the frontier are technically feasible (can be achieved with existent technology) and are possible within the budget constraint. Assume that the frontier is continuous, concave to the origin and monotonically decreasing from left to right. Consider two sets of people, A and B. The axes in Figure 4 correspond to the QALYs of groups A and B respectively. If we fix the QALYs of A at some arbitrary point, the QALYs of B will depend on the value of A at that point, the supply of resources and the technical feasibility of transforming resources into QALYs. If the same process is followed for the QALYs of B, a locus called the QALY possibility frontier can be mapped out as depicted by PP. This illustrates the maximum level of QALYs any group can enjoy, given the levels of other groups – in other words it presents the interpersonal distribution of QALYs (De V Graaff 1973; Williams and Cookson 2000).

Figure 4: Health possibilities frontier and health-related social welfare functions



Source: Williams and Cookson, 2000 p. 1874

In Figure 4, welfare contours are constructed at 45° to the origin to illustrate that QALYs are of equal social value to everyone – only the total amount is important. These contours represent the locus of indifference between all pareto optimal states. A pareto optimal state is defined as a state where no-one can be made better off without making someone else worse-off. At the point of tangency between the welfare contour and PP at X, QALYs are maximised for that budget (De V Graaff 1973; Williams and Cookson 2000) .

The application of the health maximisation social choice rule in health economics is undertaken through two different types of economic evaluation – cost-effectiveness analysis and cost-utility analysis[13]. These differ according to the way in which the benefits of health care are measured, but take a similar approach in their assessment of costs. The following sections include an explanation of the concept of cost that is appropriate in economic evaluation followed by a discussion of cost-effectiveness and cost-utility analysis.

4.1.1.1. Cost

The notion of cost used in economic evaluation is that of opportunity cost, which itself rests on the principles of scarcity and choice. Scarcity exists because societies do not have enough resources to fulfil every want, irrespective of the wealth of the society. Because of this, choices need to be made between the activities that are undertaken and those that are not. Thus the cost of a programme in economic evaluation is defined as its opportunity cost, which is equivalent to the benefits foregone from the best use of resources in an alternative health care programme (Donaldson 1990). It is also argued that if the benefit of a programme is expressed in terms of health outcomes such as life years or QALYs, then only the opportunity cost of health care resources can be considered (Gerard and Mooney 1993; Posnett and Jan 1996; Mooney and Jan 1997; Donaldson 1998; Currie, Donaldson et al. 1999). While some methodological texts recommend a societal perspective in costing (Luce, Manning et al. 1996; Brouwer, Rutten et al. 2001), this would require the inclusion of non-health care resources (e.g. time costs associated with seeking health care). Because these need to be compared with the potential gains or losses not only in health benefits but also in non-health benefits (e.g. diminished utility because of spending leisure time seeking health care), their inclusion in an economic evaluation that expresses outcomes in terms of health benefits would be inconsistent.

4.1.1.2. Cost-Effectiveness Analysis (CEA)

If it has been decided that a particular health care programme is to be delivered or a particular disease is to be treated, the cost-effectiveness analysis tries to identify the best way of doing so (Donaldson 1990). To be more precise, the "best" method is defined in terms of technical

[13] The other key form of economic evaluation is the cost benefit analysis. This form of analysis places a monetary value on health gains, so that both costs and outcomes are commensurate. However, owing to a number of difficulties in valuing health in these terms, the cost benefit analysis is infrequently used in health economics.

efficiency: the production of a given quantity of output with the least cost combination of inputs (Birch and Gafni 1992). Cost-effectiveness analysis is therefore argued to be a method for determining the least cost way of achieving a given output, whether the same level of output can be achieved with less of one input or the best way of spending a given budget for a group of patients (Donaldson, Currie et al. 2002).

CEA measures the outcomes of a health care programme in natural units, such as life years saved, viral suppression or CD4 gain. It follows that CEA can only compare between programmes that produce directly comparable outcomes (Birch and Gafni 1992) and that CEA does not involve comparisons between patients with different diseases (Donaldson, Currie et al. 2002). Moreover, because outcomes in CEA are unidimensional, if improvements in quality of life are an additional impact from antiretroviral treatment, a CEA that focused on improved life expectancy would understate the health-related benefit of the intervention.

The most common output from a CEA is an incremental cost-effectiveness ratio (ICER). This is calculated by dividing the additional costs of an intervention by the additional outcomes with the implication that lower ratios are preferable to higher ones. Given that the CEA only focuses on alternative approaches to achieve a given objective (e.g. to treat a particular disease) the incremental costs and benefits usually relate to the introduction of a new treatment in comparison to the status quo. In some unusual cases, a new treatment can be more effective and less costly than the status quo. In this case one could say that the new programme is technically efficient: in comparison to the status quo, the new intervention could produce a given level of output with fewer resources. More frequently, however, the new treatment is both more effective and more costly than the status quo. In this case, a judgement based on technical efficiency cannot be made and the analyst cannot comment on whether the new programme is cost-effective or not. To fund the new programme for the same number of patients would require diverting health care resources from an alternative disease or health care programme (Birch and Gafni 1992; Donaldson, Currie et al. 2002; Sendi, Gafni et al. 2002). Whether this is worth doing is a question of allocative efficiency, which will be discussed under the section on cost-utility analysis.

4.1.1.3. Cost-Utility Analysis (CUA)

Theoretically, the cost-utility[14] analysis is also able to assess technical efficiency, although the same caveats regarding technical efficiency judgements for programmes with higher costs and effects still applies as outlined above. The key difference between the CUA and CEA is that the CUA uses a broader outcome than the CEA. This is normally the QALY, but it is also possible to use the disability adjusted life year (DALY)[15]. If QALYs are accepted as an adequate measure of the benefit of health care[16] and if it is accepted that they can successfully combine length of life and health-related quality of life in one index, one can argue that the CUA has advantages over the CEA (Mooney 2003). This is because assessment can be made of treatment for a disease that has an impact on length of life and health-related quality of life in comparison to an alternative. In this context, the CUA would be better placed to assess technical efficiency than the CEA because it provides a better measure of the health-related benefit of the intervention.

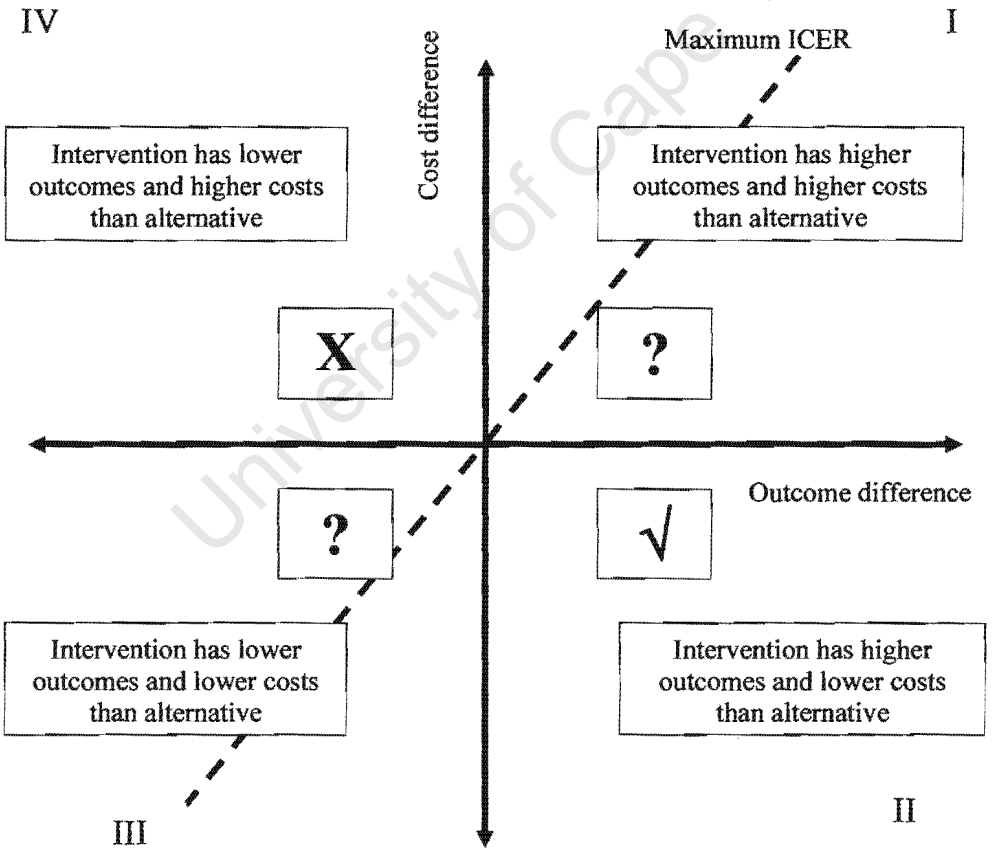
However, it is also argued that the CUA can assess the efficiency of allocations between competing health care programmes (allocative efficiency). In this broader role, CUA aims to maximise health outcomes across programmes or interventions within the health care budget constraint (Wagstaff 1991). The implication is that the CUA no longer restricts itself to evaluating goals that have already been defined and that comparison is now made between different groups of patients, health care programmes and diseases. In this formulation of the CUA, it is once again assumed that the QALY is an adequate measure of the benefit of health care and that society views a QALY as of equal value no matter who gains. It is also assumed that there are no other objectives in health care other than health maximisation (Birch and Gafni 1992; Mooney 2003).

In practice, two different approaches have been suggested to assess allocative efficiency in CUA: the threshold ICER and QALY league table. The threshold ICER reflects the maximum amount that society would be willing to pay to purchase a QALY – any intervention with an ICER below this level would be implemented and interventions with ICERs above this level would not (Sendi, Gafni et al. 2002). This can be illustrated graphically in Figure 4 through the use of the cost-effectiveness plane (Briggs 1995; Briggs 2001). The plane is divided into four quadrants. The horizontal axis shows the difference in effect between two interventions (QALYs gained) while

[14] Although this thesis is interpreting the QALY as a measure of health and not health-related utility, the conventional term cost-utility analysis has been maintained for reasons of clarity.
[15] Whether the QALY or DALY is used does not alter the point made in this section, so the discussion will continue to focus on the QALY.
[16] For more details on the adequacy of QALYs as a proxy for the good see section 2.

the vertical axis shows the incremental costs. If the results of the cost-effectiveness analysis show the difference in costs and effects in quadrants II or IV, absolute dominance exists. In the case of II, the new treatment is more effective and less expensive, and should be implemented. In the case of IV, the new treatment is more expensive but less effective, and should not be implemented. If results indicate quadrants I or III, a trade-off is required. In the case of III, the new treatment is less costly, but is also less effective, and in the case of I, the new treatment is more expensive and more effective. In this case, the "standard" CUA decision-making rule is to implement the intervention as long as its ICER is less than a maximum acceptable cost-effectiveness ratio, as indicated by the slope of the dotted line (Briggs 1995; Briggs 2001).

Figure 5: The cost-effectiveness plane



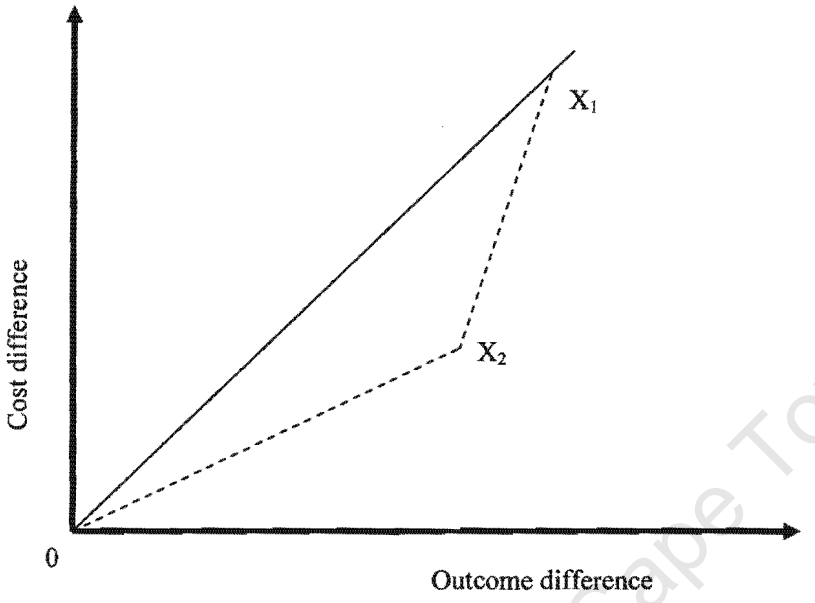
Source: adapted from Briggs (1995)

Despite the central role that is increasingly being given to the threshold value of the ICER in CUA, little attention has been given to determining the value of this threshold in practice. Theoretically, the threshold ICER represents the opportunity cost of resources at the margin which means that it is equal to the ICER of the last programme selected before the budget is exhausted. The implication is that a QALY league table of all existing programmes needs to be produced in order to calculate the threshold ICER (Gafni and Birch 2006).

In the QALY league table approach, the question that is addressed is "what additional QALYs can be bought by allocating additional health service resources to the listed existing programmes?" or to a new programme (Gerard and Mooney 1993 p. 60). The underlying rationale is to produce each health care programme until the level of the ICER is the same across all interventions and the decision to introduce a new intervention therefore involves a comparison of the ICER of that intervention against other existing interventions. Note that this implies that one cannot assess whether existing allocations are efficient.

A league table is constructed from a number of individual CUA studies. The validity of recommendations based on a league table will therefore be constrained by the availability of these studies, the quality of the research and whether they have been conducted in a setting that is appropriate to the decision context that the league table seeks to address. Even if good quality studies are available, these will be context specific – the ICER will depend on the cost structures and the epidemiology of the particular setting in which the study was conducted (Gerard and Mooney 1993). Given however the paucity of CUA studies, one danger is that analysts will make assumptions of constant returns to scale when constructing the league table (Birch and Gafni 1992; Ubel, DeKay et al. 1996; Nord, Pinto et al. 1999; Sendi, Gafni et al. 2002). Consider quadrant I of the cost-effectiveness plane depicted in Figure 6. If there are constant returns to scale, the ICER of intervention X would be $0X_1$. However, it is possible that intervention X would have a lower ICER at lower levels of scale (for example if provision were through urban centres) as shown by $0X_2$. As the intervention expands to rural areas, costs might increase, as exhibited by X_2X_1. In other words, the ICER is dependent on the size of the intervention that is being considered and the context. The implication is that each local or provincial government would require a league table for its specific context which would differ from league tables for other provinces or for the country as a whole.

Figure 6: Returns to scale on the cost-effectiveness plane



Source: Sendi, Gafni et al. 2002 p. 25

A further challenge for the threshold ICER or QALY league table approach is that the ICER should be calculated relative to the highest valued alternative use because the level of the ICER is dependent on the cost to QALY ratio of the comparator (Birch and Gafni 1992; Gerard and Mooney 1993; Ubel, DeKay et al. 1996; Nord, Pinto et al. 1999). For example, if the ICER for ART were calculated relative to No-ART and No-ART had a high cost per QALY ratio, then the ICER of ART would appear relatively more favourable than otherwise.

Given the many pitfalls of the QALY league table or threshold ICER approach, Birch and Donaldson (1987) argued that one theoretically correct approach to solving the health maximisation problem would be through the use of linear or integer programming. Following Stinnett and Paltiel (1996) the health maximisation problem could be structured as follows:

$$\max_{x_i \ldots x_n} \sum_{i=1}^{n} x_i E_i$$

Equation 1

Subject to:

$$0 \leq x_i \leq 1$$

$$\sum_{i=1}^{n} x_i c_i \leq C$$

$$\sum_{i=1}^{n} x_i \leq 1$$

Where:

C is the present value of all health care resources over the planning period

i is an index describing all possible mutually exclusive interventions available $(i = 1,...,n)$

c_i is the present value of the cost of providing intervention i over the planning period

E_i is the present value of the QALYs of intervention i over the planning period

x_i is a decision variable – if $x_i=0$, intervention i is not implemented, if $x_i=1$ intervention i is implemented

The objective function in Equation 1 seeks to maximise QALYs, subject to a number of constraints. The first set of constraints restricts each intervention's implementation level to be between 0 and 100 per cent. In other words, interventions are assumed to be divisible. The second constraint is on the budget in order to ensure that total costs do not exceed available resources. The final constraint ensures that the sum of the portions of mutually exclusive interventions implemented cannot exceed 100 per cent.

Using linear programming to assess allocative efficiency across health care programmes in CUA would require data on the costs and consequences of all mutually exclusive interventions within each independent health care programme. However, the approach can be used more modestly in assessing technical efficiency in the CUA when it has been decided that a particular treatment programme will be pursued and when a health care budget for this independent programme has been allocated. This approach to assessing technical efficiency has a number of strengths. Firstly, it solves the health maximisation problem within one independent programme in a theoretically correct manner (as a maximisation problem within a budget constraint). Secondly, because it uses data on the full costs of each intervention, the costs of scaling up are also available for

consideration. Thirdly, it draws attention to the total health gain that can be achieved within alternative health care budget constraints.

While this is an improvement on the standard decision rules of CEA/CUA, the shortcomings of assuming constant returns to scale and complete divisibility are still problematic. Although the latter will be relaxed in the next section, data on increasing or decreasing returns to scale are rarely available. In addition, this approach is not able to quantify the opportunity cost of the health care budget allocated to one independent programme over another; an assessment of allocative efficiency would require costs and benefits of all independent programmes included in the total health care budget. However, bringing clarity to the health gains associated with alternative budgets within one programme should assist in setting evidence informed budgets for that programme.

4.1.2. Equal health

In contrast to the health maximisation approach, an equal health approach proposes that a fair distribution of HIV-treatment would be one that led to the same health gain for all HIV-positive people at a given level of need. Culyer and Wagstaff (1993) argue in favour of the equal health approach firstly because health is necessary for flourishing and fairness requires an equal opportunity to flourish. However, equal opportunity to flourish could be compatible with equal health or with equal access to health care. To move from opportunity to flourish to equal health, Culyer and Wagstaff (1993) argue that differences in health that are unrelated to own free choice (such as genetic predispositions and the socioeconomic environment) are inequitable (i.e. unfair). Secondly, they argue that differences in health that are related to personal responsibility are also inequitable. The principle of equal health is however qualified by a side condition that greater equality cannot be achieved by reducing the health of some as a deliberate act of policy. In addition, they suggest that the apparent severity of health equality can be tempered through trading-off equal health against health maximisation.

The health maximisation approach in Figure 4 can be restated to encompass equality by making the welfare contours convex to the origin and symmetrical around a 45° line. The greater the curvature in the welfare contour, the greater the weight given to equal health over health maximisation. In the extreme case of strict equality in health, the welfare contours would be L-shaped implying a zero trade-off between maximisation and equalisation. This approach can be

operationalised for one independent programme (e.g. HIV treatment) by replacing the linear programming approach from Equation 1 with an integer programming specification (Birch and Gafni 1992; Stinnett and Paltiel 1996):

$$\max_{x_i \ldots x_n} \sum_{i=1}^{n} x_i E_i \qquad \textbf{Equation 2}$$

Subject to:

$$x_i = 0..or...x_i = 1$$

$$\sum_{i=1}^{n} x_i c_i \leq C$$

$$\sum_{i=1}^{n} x_i \leq 1$$

Where:

C is the present value of all health care resources over the planning period

i is an index describing all mutually exclusive interventions available $(i = 1,...,n)$

c_i is the present value of the cost of providing intervention i over the planning period

E_i is the present value of the QALYs of intervention i over the planning period

x_i is a decision variable – where $x_i=0$, intervention i is not implemented, if $x_i=1$ intervention i is fully implemented

Although the objective function in Equation 2 is the same as in Equation 1, the value of this function is further restricted by an assumption of complete indivisibility as shown in the first set of constraints. What this means is that each intervention's implementation level can be either 0 or 100 per cent. Because the integer programming approach of Equation 2 simply adds additional constraints to the linear programming approach of Equation 1, the optimal objective function value obtained from Equation 2 is lower than Equation 1, although the difference would be negatively related to the size of the budget. This difference shows the foregone health benefit from an equality constraint (Stinnett and Paltiel 1996).

4.1.3. Trading off equality and maximisation

The health maximisation social choice rule is often criticised for paying too little attention to who the QALY recipient is. At the other extreme, equal health is criticised for paying too much attention to the worst off. According to Elster (1992), a "common sense"[17] alternative is to maximise total welfare subject to a floor constraint. In the terminology of this thesis, this would imply maximising QALYs after providing a decent minimum level of QALYs to all. The key issue in operationalising this approach is to decide on the decent minimum, but as will be shown, this is function of the health care budget that is available to the HIV-treatment programme. At lower budgets, the decent minimum will be different from what would be possible at higher budget constraints.

Empirical evidence supports this commonsense approach (Miller 1992). When presented with four alternative principles for distributing income (maximising the minimum, maximising the average, maximising the average subject to a floor constraint and maximising the average subject to a range constraint[18]) maximising subject to a floor constraint was first choice for between two-thirds and three-quarters of respondents. The second choice, by a large margin, was to maximise the average (i.e. QALY maximisation if the evaluation was in the space of QALYs instead of incomes). Maximising the minimum which corresponds to an equal health approach was chosen by between zero and four per cent of respondents.

The "decent minimum" can be operationalised through specifying an intermediate position between Equation 1 and 2 in Equation 3. In this specification, QALYs are maximised subject to some form of "decent minimum" being offered to all patients within an independent programme (i.e. HIV treatment). The sum of mutually exclusive interventions must be equal to one, but patients do not have to receive the same form of treatment, as indicated in the last line of constraints relating to Equation 3.

[17] Elster (1992) defines the commonsense conception of justice as the principles of justice held by those who have given serious thought to the matter but who are not professional philosophers. He argues that these views might be held by lawyers, economists and politicians.
[18] In other words, the gap between top and bottom incomes should not exceed a certain amount.

$$\max_{x_i \ldots x_n} \sum_{i=1}^{n} x_i E_i$$ **Equation 3**

Subject to:

$$0 \leq x_i \leq 1$$

$$\sum_{i=1}^{n} x_i c_i \leq C$$

$$\sum_{i=1}^{n} x_i = 1$$

Where:

C is the present value of all health care resources over the planning period

i is an index describing all possible mutually exclusive interventions available $(i = 1,\ldots,n)$

c_i is the present value of the cost of providing intervention i over the planning period

E_i is the present value of the QALYs of intervention i over the planning period

x_i is a decision variable – if $x_i=0$, intervention i is not implemented, if $x_i=1$ intervention i is implemented

4.2. Procedural justice

Procedural justice focuses on the fairness of the process through which a distribution is achieved. In contrast to consequentialism which judges fairness in terms of outcomes, non-consequentialism judges fairness through the procedures that lead to the distribution of outcomes irrespective of the outcomes (Daniels 2004). Rawls (1971) has defined a number of different types of procedural justice. Perfect procedural justice was defined with the example of the division of a cake. If it is agreed that the fairest outcome is for everyone to get a piece of the same size, then one possible procedure would be that whoever cuts the cake gets the piece that remains after everyone else has chosen. The key constituent of this form of procedural justice is that there is an independent criterion for defining a fair division (i.e. equality) and a procedure that is guaranteed to lead to it. This differs from imperfect procedural justice. Here there is an

independent criterion for the right outcome, but currently there is no procedure that will lead to this outcome. For example, in the case of criminal trials, it might be agreed that the fair outcome is that those who are guilty are found guilty and vice versa for those who are innocent. The problem is that the criminal justice system is an imperfect procedure that cannot guarantee this outcome. By contrast, in pure procedural justice there is no independent criterion for the right result (Daniels 2004). In the terminology of this thesis, this means that reasonable people will disagree on which of the social choice rules should be used to guide distributions.

4.2.1. Fair process

Daniels (2004) argues that fair process allows greater legitimacy in patient selection when there is no clear agreement about substantive principles that determine fair distributions. For example, some might argue that personal responsibility for HIV-status should limit claims on the good whilst others would argue that the social context of HIV-infection implies an additional duty to provide health and health care for the poor. Some might favour a health maximisation social choice rule while others would be in favour of a decent minimum.

While these controversies might not be irresolvable, Daniels (2004) argues that they pose an area of substantial disagreement that could be solved through fair process.

The central requirements of fair process are:

- Publicity: the process must be transparent and involve publicly available rationales for the priorities that are set. This has the added benefit of encouraging good governance.
- Relevance: stakeholders who are affected by the decisions should agree that they rest on reasons, principles and evidence that they view as relevant to making fair decisions about priorities. This has the added benefit of ensuring stakeholders that their voice has been heard.
- Revisability and appeals: decisions can be revisited and revised in light of new evidence and arguments. This appeals process provides protection to those who have legitimate reasons for being an exception to adopted policies.
- Enforcement or regulation: a mechanism is in place to ensure that the previous three conditions are met.

4.2.2. Communitarian claims

An alternative approach to procedural justice, called communitarian claims, has been proposed by Mooney and colleagues (Mooney and Jan 1997; Mooney 1998; Mooney, Jan et al. 2002; Mooney 2005). This rests in communitarianism which places importance on the value of community and social relations. The aim is to create a good society with a balance between individual and community rights. It is also argued that individualistic notions of self discount the role that community plays in forming and sustaining individual identities – traditional individualism with its rhetoric of personal liberty and rights tends to impede compromise, mutual understanding and the discovery of common ground. Although communitarians are respectful of the individual, they are concerned about the overemphasis on individuals and individual rights (Black and Mooney 2002). Buchanan (1989) argues that communitarians emphasise that a genuine community is more than just an association of individuals. Members of a community have common ends, and these ends are conceived of and valued as common ends. "Each member thinks of furthering the community's ends as a gain for *us*, not as a gain for herself which happens to be accompanied by similar gains for other individuals constituting the group" (p. 857).

Therefore, under this approach, claims are not just the claims of the community, but are communitarian claims, suggesting that there is an intrinsic value in community and that a duty is owed by the community to an individual by virtue of her being a member of that community. In addition, there is value to the community in arbitrating over the process of setting criteria (Mooney 2005). This approach also emphasises that the way in which distributions of the good between HIV-positive people is achieved might have an impact on the extent to which South Africa moves towards greater or lesser social solidarity.

In order to operationalise communitarian claims Mooney, Jan et al. (2002) recommend dividing the population into groups which could determine the constituents of claims. For instance, "social class, existing health status, capacity to benefit … may be seen as forming bases for differential claims" (p. 1660). The second task is to establish the relative weights attributable to these claims. The establishment of the claims and also the differential weights would be done by the community. However, Mooney does not suggest that the community should replace the policy-maker in considering the benefits and the costs of different technologies, but that the community should be consulted about the principles on which policy-makers should act. In other words, the community sets the value basis for decision-making in the health sector (Mooney 2005).

However, Mooney (1998) concedes that the degree of homogeneity of the society and the extent of social solidarity would be important in ensuring that claims are recognised. "An atomistic, individualistic society will be slow to recognise that the community does have a duty with respect to meeting such claims. The more embedded individuals are in a community and the greater the recognition of such embeddedness, the greater will be the strength of the communitarian claims in that community" (p.1176). Although Mooney recognises that communitarianism is only good if the community on which it is based is good, he argues that the prospect for a good community seems greater when individuals are embedded in the community.

5. Summary

The ultimate aim of this chapter has been to propose a framework for decision-making around allocations of health and health care to HIV-positive adults in South Africa. This thesis is therefore located within the field of distributive justice, where "principles of distributive justice are normative principles designed to allocate goods in limited supply relative to demand" (Lamont 2003). Distributive justice is a broad field, but of particular relevance in this thesis is consequentialism and procedural justice. In consequentialism, normative properties only depend on consequences (Sinnott-Armstrong 2003). By contrast, procedural justice focuses on the fairness of the process through which distribution is achieved (Rawls 1971).

In constructing this framework, three questions were posed. The first question – what is the good - was partly about defining the outcomes of health care and partly about defining the evaluative space of distributive justice as applied to HIV-treatment in South Africa. The good of health care was argued to be health, which was defined as an absence of illness from the point of view of the individual. Placing health in the evaluative space was defended because health is a fundamental constituent of capabilities - without good health a person is severely constrained in leading a life that she or he might have reason to value.

Having defined health as the good, the next step is to find a way of measuring health that is an accurate reflection of this definition, is able to capture some of the variables of importance to the policy question and is reproducible. Given the multidimensional and culturally specific nature of health, even if defined narrowly from the point of view of the individual, it is obvious that there is

not a perfect match between the QALY and health. However, the use of the QALY is nevertheless defended in this thesis. Firstly, because there is no comparison proposed between HIV and other diseases, the interpersonal comparability assumptions are less restrictive than if the QALY were to be used to assess allocative efficiency across diseases and interventions. Secondly, the QALY is preferred to a unidimensional outcome such as the life year because there is evidence to suggest that different HIV-treatment interventions can have an impact both in terms of length and quality of life. QALYs therefore capture more of the variables that are relevant to the policy question than un-weighted life years.

The second section in this chapter considered a number of characteristics that could be used as the basis for constructing claims on the good. These included the personal characteristics of HIV-positive people, the need for health care, and the impact of HIV-treatment on the health of society and the social fabric.

The third section examined a number of consequential social choice rules that could assist in distributive decisions. The three social choice rules were health maximisation, equal health and the decent minimum. While this is not a complete list of all possibilities, these have been chosen because they reflect a broad spectrum of the possible approaches that can be taken to distribution. In the literature, health maximisation is generally associated with allocative efficiency and practical implementation normally involves a threshold ICER or QALY league table approach. Neither of these approaches will be pursued given their theoretical shortcomings. As an alternative, technical efficiency will be assessed as a health maximisation problem operationalised through linear programming across a number of different HIV-treatment strategies. The second social choice rule that was discussed was equal health. Costs and consequences associated with this approach in HIV-treatment will be assessed using integer programming. The final consequentialist rule – the decent minimum - proposes an equity-efficiency trade-off where all patients in need receive HIV-treatment, but some patients receive less effective forms of treatment. Again, this will be assessed through linear programming as detailed in Equation 3. (p. 47)

While linear or integer programming is a useful approach to assessing the costs and consequences of alternative forms of HIV-treatment, an ultimate decision about which approach ought to be adopted is a question of values; some people will favour health maximisation while others will insist on an equal opportunity for health for everyone in need. Similarly, there is likely to be

Finally, because this thesis acknowledges that reasonable people will disagree about the relevance of using one social choice rule over another, Chapter 8 provides guidance on the implementation of a fair process in HIV-treatment priority setting.

Chapter 4: Methodology

1 Introduction

This chapter reviews relevant methodological literature and describes methods for calculating the costs and outcomes of three mutually exclusive HIV-treatment interventions: the treatment and prophylaxis of opportunistic infections and events (hereafter No-ART), the treatment and prophylaxis of opportunistic infections and events with first-line antiretroviral agents (hereafter first-line ART) and the treatment and prophylaxis of opportunistic infections and events with first and second-line ARVs (hereafter first and second-line ART). Methods for establishing these costs and outcomes are described at the patient-level and at the population-level. In the former, costs and outcomes refer to those calculated over the lifetime of one patient, while in the latter, these are the full costs and outcomes of scaling up each intervention to a defined population in need over a given planning period. The building blocks in these calculations include the utilisation of HIV-related services, their associated full economic unit costs, death, loss to follow-up, treatment failure[19] and health-related quality of life data. Because ART is a long term intervention for which final outcomes are unavailable from primary data, methods for extrapolating these building blocks, validating modelled results and for assessing uncertainty have been reviewed and developed.

The building blocks in this analysis (utilisation of services, unit costs, health-related quality of life and clinical outcomes) have been estimated in a cohort of patients enrolled in public sector HIV treatment pilots. The results from this setting will allow the calculation of a base case scenario. However, because these might not be generalizable to other settings, a generalized scenario will also be constructed by comparing the results from the base case scenario to those in the empirical literature. Where differences are found between these, adjustments will be made to unit costs, utilisation and outcomes.

[19] Death or survival, loss to follow-up and treatment failure will collectively be called clinical outcomes data to distinguish these from final health outcomes such as QALYs and life years.

2. Methodology for patient-level analyses

2.1. Study population and setting

Patients included in this study live in an area known as Khayelitsha which is located on the outskirts of Cape Town in the Western Cape Province of South Africa. Many live in "shacks" made from wood, cardboard and tin in unsanitary environments with limited access to toilet facilities, tapped water or electricity. The level of unemployment in the area is approximately 46 per cent (Nattrass 2002). In April 2000, three HIV clinics were opened within Nolungile, Michael Mapongwana and Site B community health centres (CHCs) in the area known as Khayelitsha. These clinics provided treatment and prophylaxis for HIV-related and opportunistic infections and events, counselling and support groups for HIV-positive people. Prophylactic agents included trimethoprim-sulphamethoxazole and influenza vaccines for all patients and fluconazole for patients with previous cryptococcal meningitis. Acute infections were managed at the clinics but severely ill patients were referred to secondary and tertiary hospitals. Patients suspected of having TB were referred to TB facilities.

In May 2001, the service was extended to include ART for patients with CD4 less than 200 cells/μl at any WHO stage or with WHO stage IV and any CD4 level. ART patients continued to receive treatment and prophylaxis for acute infections and appropriate referrals. Details of early clinical outcomes in this cohort are available from published sources (Perspectives and Practice in Antiretroviral Treatment 2003; Coetzee, Hildebrand et al. 2004; Goemaere, Louis et al. 2004). This site was chosen for this research as it was the first public-sector based ART treatment pilot in South Africa and experience from this has informed the development of national ART guidelines. It provides access to the long-term prospective follow-up of a large number of patients in a poor setting with a community HIV prevalence that is high relative to other areas in the Western Cape.

While utilisation, clinical outcomes and HRQoL data were prospectively measured on the Khayelitsha cohort, in order to calculate costs that were widely generalizable, primary cost data were combined with appropriate secondary data, either from published or grey literature. For ART services, costing was undertaken at the three Khayelitsha CHCs and the TC Newman district hospital. The costs of No-ART services were also calculated in these facilities, and

additional costing was undertaken at Crossroads, Guguletu and Browns Farm CHCs; Wellington, Mbekweni and Phola Park clinics and the GF Jooste outpatient department (OPD). No-ART visit costs were also derived from secondary cost analyses undertaken in Guguletu CHC, Guguletu and Nomzamo clinics and Groote Schuur OPD (Govender, McIntyre et al. 2000). For TB treatment, primary costing was undertaken at Nyanga CHC/clinic, and was supplemented with secondary data from Sinanovic, Floyd et al (2003). Nyanga TB clinic was chosen because its patients have the highest burden of HIV-related tuberculosis in the Guguletu/Nyanga health district[20] (Soraya Elloker, District Manager, personal communication). For tertiary level inpatient care, primary inpatient costing was undertaken at Tygerberg Academic Hospital and was supplemented with secondary data from Groote Schuur Hospital (Govender, McIntyre et al. 2000). The cost of secondary/district level inpatient care at GF Jooste Hospital and Carnation step-down facility was derived from Haile (2000) and Smith De Scherif, Schoeman et al. (2005). Because of the additional HIV-related burden of disease placed on the medicine wards at GF Jooste, the Carnation step-down facility was established in 2003. This provides additional care for medicine patients (many of whom are HIV-positive) at a lower clinical staffing intensity.

Table 1 contains the full list of facilities included in cost analyses, type of service, the source of data (primary or secondary), HIV-prevalence (according to the 2003 Western Cape antenatal clinic survey) and the area where the facility is located.

[20] Sixty-seven per cent of Nyanga TB patients are HIV-positive.

Table 1: Settings for calculation of unit costs of HIV-related services

Type of facility	Name of facility	Type of service (and data source)	HIV-prev-alence	Area
Clinics	Chapel Street Clinic	TB (*)	11.6	Cape Town central
	Guguletu Clinic	Primary care (§) and TB (*)	28.1	Guguletu/Nyanga
	Phola Park Clinic	Primary care (primary data)	10.1	Paarl
	Mbekweni Clinic	Primary care (primary data)	10.1	Paarl
	Wellington Clinic	Primary care (primary data)	10.1	Paarl
	Nomzamo Clinic	Primary care (§)	9.3	South Peninsula
CHCs	Nyanga CHC/Clinic	TB (primary data and *)	28.1	Guguletu/Nyanga
	Guguletu CHC	Primary care (primary data and §)	28.1	Guguletu/Nyanga
	Crossroads CHC	Primary care (primary data)	28.1	Guguletu/Nyanga
	Brown's Farm CHC	Primary care (primary data)	28.1	Guguletu/Nyanga
	Nolungile CHC	Specialised HIV (primary data)	27.2	Khayelitsha
	Michael Mapongwana CHC	Specialised HIV (primary data)	27.2	Khayelitsha
	Khayelitsha (Site B) CHC	Specialised HIV (primary data)	27.2	Khayelitsha
Step-down facility	Carnation	Inpatient care (†)	28.1	Guguletu/Nyanga
Secondary hospitals	GF Jooste Hospital	Inpatient and outpatient care (†, ± and primary data)	28.1	Guguletu/Nyanga
	TC Newman	Specialised HIV (primary data)	10.1	Paarl
Tertiary hospitals	Groote Schuur Hospital	Inpatient and outpatient care (§)	11.6	Cape Town central
	Tygerberg Hospital	Inpatient care (primary data)	8.1	Tygerberg Western

Sources:
[§]Govender, McIntyre et al. (2000)
*Sinanovic, Floyd et al. (2000; 2003)
[±] Haile (2000)
[†] Smith De Scherif, Schoeman et al. (2005)

2.2. Scope of costs

Costing takes a public health sector perspective in the context of scaling up a large new health care programme over the long-run. Scaling up this programme requires medicines, laboratory investigations and other variable resources as well as long-run investments aimed at increasing the capacity of the health care system in order to provide care at the envisaged quantity while at the same time avoiding the crowding out of other priorities. The scope of costs therefore includes both variable and fixed costs given that both have an opportunity cost in this context. Variable costs are defined as the value of goods, services and inputs that change as an intervention is implemented, for example medicines, imaging and laboratory investigations. It also makes sense to think of certain costs as semi-variable. While these do vary with the level of activity, this

happens in a non-continuous manner. An example is a salaried doctor on a short-term contract. While this doctor might be able to see up to thirty patients per day, he or she would still earn the same salary if the patient load were reduced over the duration of his or her contract. However, because of being on a short-term contract, the doctor can be reallocated to a clinic with a higher patient load. On the other hand, a doctor on a long-term contract linked to a particular health facility could be regarded as a fixed cost. Fixed costs are resources that are held at a constant level, independent of the level of production, such as capital (furniture, equipment and buildings) and overheads (administration, security, cleaning and utilities) (Clewer and Perkins 1998). These are variable in the long-run.

Costs are categorised as total costs, average costs and marginal costs all of which are a function of the level of output. The total cost is the sum of the total fixed cost and the total variable cost. Average total cost, also known as the unit cost, is obtained by dividing total cost by total output; it is also possible to distinguish between average variable cost and average fixed cost. At any level of output, marginal cost is the increase in total cost that results from an increase in one unit of output. It relates to the rate of increase of the total cost curve – if total cost is increasing rapidly (decreasing returns to scale), marginal cost is high; if total cost is increasing gradually (increasing returns to scale), marginal cost is low; if the increase in total cost is constant for each unit of output (constant returns to scale), marginal cost is also constant. By definition, marginal cost intersects the average total cost curve at its lowest point; if there were constant returns to scale, marginal and average costs would be equal across the entire schedule. In the short-run, the fixed cost does not vary with the level of production so the value of the marginal cost curve is not affected by changes in the fixed cost. This changes in the long-run when all costs are variable (Clewer and Perkins 1998; Rice 2003).

The marginal cost is the appropriate cost statistic to use in economic evaluation (Jacobs and Baladi 1996). However in practice it is difficult to estimate the marginal cost, with the implication that analysts assume that the average cost is a suitable proxy for the marginal cost. If the analysis takes a short-run perspective, the average variable cost would be used, but taking a long run perspective (as is the case in this dissertation) means that the long-run average total cost is taken as the proxy for marginal costs. This assumption is only valid if the level of production coincides with the lowest point on the marginal cost curve or if there are constant returns to scale. While the marginal cost is the correct cost statistic to use, in the long run it is possible that the average total cost might be similar to the marginal cost.

Table 2 details the categories of costs to be included in this evaluation. Patient-specific costs are those that are consumed by the patient, such as medicines, laboratory investigations and imaging. Clinical staff costs relate to the time spent by medical officers and professional nurses providing HIV-treatment. Overhead costs are costs that are shared by more than one programme, such as clinic and hospital administration, security services, cleaning services and utilities such as electricity and water. Capital costs include the equipment, furniture, vehicles and buildings that are required to increase access to HIV-treatment. Costs under the category of other related resources include counselling, support groups, direct observation of TB treatment and nutrition. While these resources have a clear opportunity cost, can this opportunity cost be related to health outcomes? While generally under the remit of social services, the opportunity cost of these resources can be related to health outcomes because these services are believed to contribute to patient retention and adherence, both of which are directly related to the final health outcomes of the HIV-treatment programme.

Table 2: Scope of costs to be included in the analyses

Health care resources	
Patient-specific costs (medicines including antiretrovirals, laboratory investigations, imaging)	
	Clinical staff (medical officers and professional nurses)
	Overheads (non-clinical staff, running costs)
	Capital (equipment, furniture, vehicles, buildings)
Other related resources	
	Counselling and support groups
	Direct observation of TB treatment
	Nutritional support

The costs detailed in Table 2 are normally referred to as direct costs, which can be defined as "the value of all the goods, services, and other resources that are consumed in the provision of an intervention or in dealing with the side effects or other current and future consequences" (Luce, Manning et al. 1996 p. 179). However, this is an inconclusive list. Direct costs also include the

inputs to care incurred by patients and their families such as the transport costs associated with seeking care and the time associated with informal care that is provided by friends or family members. Although these resources have an opportunity cost, the foregone benefit is much wider than health outcomes, therefore their opportunity cost cannot be related to the outcome used in this thesis (Mooney and Jan 1997). Indirect costs – the secondary costs relating to paid and unpaid productive activities (Donaldson 1990) – would also be excluded for similar reasons. These arise because of reduced productive activities while a patient is seeking treatment or is convalescing in hospital or at home.

2.3. Health care utilisation

The health care services associated with HIV-treatment have been categorised into inpatient care, TB treatment and clinic visits. Data on the utilisation of these services have been collected as part of the Khayelitsha HIV treatment pilot using a before and after study design to calculate utilisation in the ART and No-ART groups. This means that ART patients were used as their own control – the pre-ART period was used to calculate No-ART utilisation and the post-baseline period was used to calculate ART utilisation. While a clinical trial comparing No-ART to ART would be the gold standard for measuring utilisation, obvious ethical limitations imply that the before and after study design is the only possible choice in this context. In a setting where access to ART is optimal, the before-and-after study design could result in a selection bias whereby patients with advanced disease who have not yet accessed ART do not adequately represent patients who never access ART. The delay between the launch of the service and the availability of ART together with the huge unmet demand for ART in this study ensured however that the pre-ART period was representative of patients who did not access ART at all. The exception is costs associated with death due to the survivor bias inherent in the design. No-ART patients do not die because in reality they are pre-ART patients. These costs will be addressed by the inclusion of transition costs associated with death (see below).

Utilisation of ART and No-ART clinic visits was established from 1,729 patients with 1,146 No-ART patient years and 2,229 ART patient years of follow-up over a median No-ART and ART follow-up of 0.63 years (IQR 0.33-1.32, max 4.35) and 1.03 years (IQR 0.68 – 1.70, max 4.08). While data on referrals for inpatient and tuberculosis care were collected as part of the Khayelitsha pilot, this required validation. Although the clinics collect data on patients referred for TB treatment or inpatient care, this data required substantial validation to ensure that events

were correctly recorded. This validation was performed on a sub-sample of 670 patients, with 501 patient-years for No-ART (1,342 inpatient days and 159 TB episodes) and 693 patient-years for ART patients (with 840 days in hospital and 86 TB episodes). These data on inpatient utilisation excluded patients who had died. Instead, the utilisation of inpatient services in the period prior to death was included by calculating a separate "cost of dying" from a sample of 83 patients who had been using services in the HIV clinics but had died before being able to start ART. These patients were followed up at hospitals to establish their utilisation of inpatient care in the 6-month period preceding death. The same procedure was followed for the 81 ART patients who died. For ART this is a useful approach to ensure that costs around death are not underestimated when data are extrapolated. Further details are provided in section 2.5 on Markov modelling.

2.4. Unit costs

Unit costs of health services are defined as the (long run) average total cost per ART or No-ART visit, per inpatient day at tertiary and secondary/district facilities and per tuberculosis case treated. Unit costs were calculated using the ingredients approach and the step-down method in combination (Creese and Parker 1994; Brouwer, Rutten et al. 2001). Each methodology will be discussed in separate sections below. All costs were inflated to September 2003 prices – the midpoint of the April 2003 - March 2004 financial year that is used in most public facilities. The consumer price index excluding mortgage bonds was used to inflate recurrent costs and equipment capital costs. The Bureau of Economic Research Building Price Index was used to inflate building capital costs. Where necessary, doctor costs were increased to reflect the recent government scarce skills allowance of 15 per cent. These costs were converted to US$ using an average 2003 exchange rate (US$1=7.56 Rands) (USA Federal Reserve Board, 2005).

2.4.1. Patient-specific costs and the ingredients method

The ingredients approach, also called micro-costing (Luce, Manning et al. 1996), involves identifying the specific resources concerned in delivering a service – such as medicines, laboratory investigations and imaging. These are measured from actual utilisation of these resources by a sample of patients. These "patient-specific" costs were estimated by collecting data on physical units of patient-specific resources (e.g. quantities of different types of medicines, laboratory investigations, imaging and procedures).

At the three HIV clinics, ARV prescriptions, prophylactic medicine prescriptions (fluconazole and trimethoprim-sulphamethoxazole) and numbers of X-rays are recorded in a database. Prescriptions for curative medicines and multivitamins were extracted from the records of 60 patients who had been on ART for at least one year[21]. This amounted to a sample of 757 visits for No-ART patients and 1,532 visits for ART patients. The type and frequency of safety and monitoring laboratory investigations was based on national protocols in order to increase the generalizability of results.

While the three Khayelitsha HIV clinics also provide care for those who are not yet on ART, most patients in South Africa would access ambulatory No-ART care within general primary care services. A number of additional cost analyses at Wellington, Phola Park and Mbekweni clinics, Crossroads, Guguletu, and Browns Farm CHCs, and TC Newman and Jooste OPDs were therefore undertaken. Because patients had not given informed consent to be included in a research project, ethical constraints meant that record reviews were not possible. Instead, medical officers were hired to work in each facility and provide services to patients. At the end of each clinical consultation, the medical officer recorded patient-specific resources and the HIV-status of the patient. This was an ethically acceptable method because it kept knowledge of a patient's HIV-status within the clinical relationship. Of the 810 patients treated by medical officers across facilities, 121 (15 per cent) had a confirmed HIV-positive diagnosis and were included in the calculation of costs.

The costs of HIV-related TB care were similarly established at the Nyanga Clinic/CHC. In this facility, the medical officer treated 47 patients, of whom 25 had an HIV-positive diagnosis. Utilisation of non-TB drugs and multivitamins was based on these 25 HIV-positive patients. Estimates of other patient-specific resources (TB drugs, X-rays, diagnostic tests) were based on South African national TB treatment protocols (The South African Tuberculosis Control Programme: Practical Guidelines 2000) and were compared to a published TB cost analysis (Sinanovic, Floyd et al. 2003). While the sample size used to calculate the cost of non-TB drugs dispensed to HIV-positive patients was small, the overall impact of any over or underestimation of these costs is likely to be negligible as they are a tiny proportion of total costs.

[21] In South Africa, patient record reviews are only allowed if patients have given informed consent to be included in research projects. All patients in the Khayelitsha pilot had provided such consent, enabling us to undertake this review.

At the tertiary hospital, clinical staff filled out resource-utilisation forms for every HIV-positive and HIV-suspected patient in the medicine, surgery, gynaecology, obstetrics, and oncology departments. Sixty-one patients with 243 inpatient days were included in the analysis. A secondary cost analysis from Groote Schuur tertiary hospital was also used (Govender, McIntyre et al. 2000).

At Jooste and Carnation, secondary patient-specific cost data were derived from the masters dissertation of a student in the Health Economics Unit (Haile 2000) and additional data were collected through a research collaboration (Smith De Scherif, Schoeman et al. 2005). While there has been no change in government policy with respect to the treatment of opportunistic and HIV-related infections at hospitals since 2000, the burden of HIV at Jooste has grown considerably over the last five years, implying that clinical practice might have changed because of resource constraints. However, this might not be the case in other settings so both sources of data on costs at Jooste have been included.

South African "National Antiretroviral Treatment Guidelines" (Department of Health 2004) specify that patients on first-line ART receive a nucleoside reverse transcriptase inhibitor (NRTI) backbone of stavudine and lamivudine in combination with a non-nucleoside reverse transcriptase inhibitor (NNRTI), which can be either efavirenz or nevirapine. The second-line regimen consists of an NRTI backbone of zidovudine and didanosine, in combination with lopinavir/ritonavir, a protease inhibitor. Prior to the publication of the national treatment guidelines, the first-line NRTI backbone in Khayelitsha consisted of lamivudine and zidovudine, but this was changed to be in line with national guidelines in 2004. To increase the generalisability of results, the costs of the nationally recommended regimens have been assumed.

The opportunity cost of patient-specific resources was assumed to be equivalent to market values. The following market values have been used:

- Medicine costs from provincial government tender prices
- Laboratory test costs from National Health Laboratory Services (NHLS) public sector prices. The NHLS is the only provider of laboratory services to the public health system.
- Imaging and procedure costs from the Uniform Patient Fee Schedule - UPFS (2005) - a schedule used in public sector hospitals to determine payments by private patients. The

private UPFS scale was used because other fee scales are subsidised, whereas the private scale is based on costs.

- ARV prices were those obtained by the government through the national tender process, and are inclusive of VAT and delivery costs to the various provincial depots (Gray 2005).

Mean patient-specific unit costs were calculated by multiplying physical units of resources consumed with their market values.

2.4.2. Overhead and capital costs using the step-down method

Unit costs include resources that cannot be directly linked to patient utilisation. These "overhead costs" include utilities (water, electricity), non-clinical staff (administrative, cleaning and security personnel) and non-patient specific stores and livestock. The costs of pharmacists were also included under overhead costs. Capital costs, on the other hand, are defined to include the costs of medical equipment, furniture, buildings and initial staff training.

An overhead cost per visit in primary care clinics, CHCs and the HIV clinics was calculated using the step-down method (Conteh and Walker 2004). Under the assumption that all patients utilise a similar amount of overheads during each visit, the step-down method specifies that an overhead cost per visit would be calculated by establishing overhead expenditure from routine data over a period of time. This would be divided by the total number of patient visits to the facility during the same time frame. An annual period was chosen to minimise any biases that might result from seasonal variations in visits and expenditure in facilities. The calculation is as follows:

Overhead cost per visit = annual overhead expenditure/annual visits

The methodology of calculating overhead costs for inpatients and hospital outpatient departments is similar. Because the hospitals included in data collection did not have cost-centre accounting systems, overhead expenditure was allocated directly to inpatients and outpatients using an allocation factor based on the patient day equivalent (PDE) (Barnum and Kutzin 1993; Conteh and Walker 2004).

The PDE is calculated as follows:

$PDE_{inpatients}$ = (annual inpatient days) + (annual outpatient visits × weighting factor)
$PDE_{outpatients}$ = (annual inpatient days × 1/weighting factor) + (annual outpatient visits)

The weighting factor was calculated as the average ratio of the cost per outpatient visit to the cost per inpatient day in the medicine and surgery departments at Groote Schuur hospital from April 2002 to January 2003 where a cost-centre accounting system allows this calculation to be made. This calculation estimated that an outpatient visit cost approximately 0.265 times an inpatient day. The standard assumption in South Africa is that an outpatient visit costs one-third of an inpatient day when there has been no attempt to calculate patient-specific resources separately which are normally higher per outpatient visit than per inpatient day. The use of 0.265 instead of 0.33 is justified here because patient-specific resources have been calculated separately and this ratio is therefore only being used to allocate overhead costs between inpatient and outpatient departments.

Therefore, in order to allocate overhead costs between inpatient days and outpatient visits, the following PDE was calculated:

$PDE_{inpatients}$ = (annual inpatient days) + (annual outpatient visits × 0.265)
PDEoutpatients = (annual inpatient days × 3.77) + (annual outpatient visits)

This allowed the calculation of the overhead cost per inpatient day and outpatient visit as follows:

Overhead cost per inpatient day = annual overhead expenditure / $PDE_{inpatients}$
Overhead cost per outpatient visit = annual overhead expenditure / $PDE_{outpatients}$

Capital costs are calculated by creating a list of all capital items, establishing the replacement value of each item, estimating the approximate working life of the item, and annuitizing using a real interest rate to arrive at an annual cost (Walker and Kumaranayake 2002). This cost was allocated to inpatient days or clinic visits using the step-down method. In South Africa, a real interest rate of 8 per cent is often used, which is the return on long term government bonds, in line with recommendations in the literature (Drummond and Jefferson 1996). This approach to calculating capital costs is a feasible (although time consuming) approach for clinics but less so

for hospitals. For this reason, many inpatient cost analyses have not included capital costs in South Africa (Karstaedt, Lee et al. 1996; Govender, McIntyre et al. 2000). However, a relatively sophisticated capital costing model has been developed by the national Department of Health (NDoH) to assist in budgeting for the upgrading of hospital facilities (Rod Bennet, personal communication). This model calculates capital costs based on the following inputs and assumptions:

Average area per bed [1]
Total number of beds [2]
Total area of beds [1] × [2]
Building cost per meter squared (based on quantity surveyor's estimates, inflated to the 2004 level, using the Bureau of Economic Research Building Cost Index, and including a mark-up for building fees)
Equipment cost estimated at 30 percent of the building cost for regional hospitals.

This model has been used to calculate capital costs in hospitals. However, it should be borne in mind that the equipment cost of 30 percent of building cost is a crude assumption. Recent work undertaken by staff at the Health Economics Unit suggests that the equipment capital costs in four district hospitals in Limpopo, KwaZulu-Natal and Mpumulanga varied between 10 per cent and 41 per cent. Nevertheless this is unlikely to have significantly influenced the final results; as will be demonstrated later, capital costs account for a very small percentage of the total costs of care.

2.4.3. Counselling staff costs

At the three Khayelitsha clinics and the TC Newman OPD, lay counsellors are employed to assist with patient counselling and education about HIV and ART. These counsellors are managed by Non Governmental Organisations (NGOs) under contract to the provincial government. The full costs of counselling from the point of view of the provincial government included the counsellor's salaries and the support and administration costs incurred by the NGOs. In Khayelitsha, interviews with the counselling coordinator revealed that counsellors spent approximately 50 per cent of their time counselling. The remainder was spent retrieving or filing patient folders, translating between Xhosa and English and dividing bulk-bought medicines into visit dosages. Of the counselling time, 80 per cent was spent with ART clients and the remainder with non-ART clients. The result was that 50 per cent of the full counselling cost was spread

equally across patient visits and that of the remaining cost, 80 per cent was spread across ART visits and 20 per cent across No-ART visits using routine visit headcounts. At TC Newman, the counselling costs were allocated directly to ART visits because this facility does not routinely manage No-ART patients.

2.4.4. Clinical staff costs

A final cost item of importance is the cost of clinical staff time. While the overall cost of clinical staff may not be a key cost driver (small in magnitude when compared to ARV costs) clinical staff shortages are likely to be a key constraint in attaining equal access to HIV-related care for all patients who are in need. At the HIV clinics, researchers used stopwatches to time clinical consultations for 54 ART visits and 94 No-ART visits, and interviewed staff about their average working hours. At all other clinics, CHCs, hospital OPDs and the TB clinic, the research doctor recorded the length of 147 clinical consultations.

This estimate of the time per clinical consultation was multiplied by the average clinical staff cost per minute in each facility (adjusted to reflect working hours and annual working days) to calculate the clinical staff cost per visit.

The clinical staff cost per minute was based on the following inputs:

- Data on employed clinical staff who are involved in providing treatment to patients in each facility (i.e. medical officers and professional nurses)
- Annual cost of employment (CoE) in 2003/04 for each cadre including pensions and other benefits and the recent policy of a medical officer scarce skills allowance (an additional 15 per cent on top of existing salary scales)

An average annual CoE in different clinical staff categories was then calculated as follows:

Average annual CoE = (employed staff of different cadres * relevant CoE) / number of clinical staff

From this annual cost, a cost per minute was calculated. There are 220.25 working days per annum (on average, there are 365.2 days per year, 104 days fall on weekends, 22 days are annual

holidays, 11 days are public holidays and it was assumed that there would be 8 days of sick leave). In addition, during an 8-hour working day, only part of the day would be spent seeing patients and the remainder would include various administrative and management tasks. If one is calculating clinical staff costs by timing consultations, one needs to include a share of these administrative costs in the cost per minute calculation. Interviews with clinic managers and medical officers indicated that clinical personnel spend approximately 6 out of 8 hours seeing patients each day. Based on these inputs, the clinical staff cost per minute was calculated as follows:

Clinical staff cost per minute = Average annual CoE / 220.25 / 6 hours / 60 minutes

At the inpatient level, it was more difficult to calculate an HIV-specific clinical staff cost owing to restrictions on researcher access to hospital wards. Therefore many HIV-related inpatient cost analyses have split clinical staff time using the step-down method (Kinghorn, Lee et al. 1996; Govender, McIntyre et al. 2000; Haile 2000; Hansen, Chapman et al. 2000; Guinness, Arthur et al. 2002). Clinical staff costs were estimated at Jooste and Carnation by establishing the number of full time equivalent doctors and nurses in each relevant ward and their average cost of employment. The resultant cost was split equally between inpatients in these wards. At Tygerberg, the annual expenditure on medical officers was established and allocated to inpatient days using the PDE method.

2.4.5. Measuring and valuing health related quality of life

HRQoL is a multidimensional concept, including many different aspects of health such as functional status, physical status, and emotional status. It is argued that HRQoL should be included in an economic evaluation if it is conceptually relevant to the choice being made – in other words, HRQoL might be important if it differs between two alternative treatments (Dowie 2002). In a related study of HRQoL on a subset of patients from the Khayelitsha cohort, Jelsma et al. (2005) conclude that "HRQoL can be greatly improved by HAART [Highly Active Antiretroviral Therapy], and that the possible side effects of the drugs seem to have a negligible impact on the wellbeing of the subjects" (p. 1). Their findings indicate that it is important to consider improvements in HRQoL as an outcome associated with ART.

In measuring HRQol, it is possible to use a generic or a condition-specific instrument. A generic instrument *intends* to cover the full range of health outcomes, whereas a condition-specific instrument attempts to pick up the range of outcomes associated with a particular condition (Dowie 2002). A review of different HRQoL descriptive or measurement techniques indicates that while there are a number of different instruments that have been developed to measure HRQoL, only a limited number of these can be used to calculate QALYs in an economic evaluation (Brazier and Fitzpatrick 2002). The most common are the Quality of Well-Being (QWB) scale, the Rosser Classification of illness states, the Health Utilities Index (HUI-II or HUI-III) and the EuroQol EQ-5D. Depending on the "domains" in each instrument, the number of feasible health states would differ. The QWB has 1,170, the Rosser has 29, HUI-II has 24,000, HUI-III has 972,000 and the EQ-5D has 243. The larger the number of health states, the longer the instrument takes to complete, and the more difficult valuation becomes. However, with fewer health states, the descriptive validity of the instrument might be lower. For instance, the EQ-5D, which has 243 health states, is argued to be less sensitive to changes in HRQoL than other instruments (Brazier and Fitzpatrick 2002). However, it has benefits in terms of empirical research because it takes much less time to administer than instruments with more health states.

For this thesis, secondary HRQoL data have been used (Jelsma, MacLean et al. 2005). These data were collected using the EQ-5D at treatment initiation (baseline), 3, 6 and 12 months on ART from patients enrolled in the Khayelitsha programme (n=95, 97, 98 and 83 respectively). No-ART HRQoL was assumed to be the same as baseline HRQoL in ART patients. The EQ-5D consists of questions across five domains: mobility, self-care, usual activities, pain or discomfort and anxiety or depression. Patients can self-report "no problems", "some problems" or "severe problems" in each domain. The EQ-5D also has a VAS section (EuroVAS), allowing patients to value their current health state. The 243 health states generated from answers on domains have been valued in a number of SG, TTO and VAS surveys (see www.euroqol.org).

While the use of these data is justifiable given that they were collected from a sub-sample of the same cohort of patients, controversies remain regarding how to transform these ordinal HRQoL rankings into a cardinal measure of a health state. There is no agreement in the literature as to which is preferable between TTO, SG or VAS methods. A further difficulty is that no local population valuation has been undertaken in South Africa.

Table 3 presents the mean scores from Khayelitsha HRQoL data using three different valuations: United Kingdom time trade-off (Dolan, Gudex et al. 1995), Harare time trade-off (Jelsma, Hansen et al. 2002) and Khayelitsha patients' own valuations using the EuroVAS. In general, the UK value set and the Khayelitsha EuroVAS give lower valuations of health states than the Harare value set. However, these differences would not necessarily matter because one is primarily interested in the change in quality of life between No-ART and ART. Here there is greater agreement, with the UK TTO values and the EuroVAS values both showing a similar percentage increase of 14 per cent and 15 per cent while the Harare TTO values show a percentage increase of 8 per cent.

The final column in the table shows SF36 HRQoL data from two Cape Town studies (O'Keefe and Wood 1996; Pitt, Badri et al. 2005) which have been converted to values using a United Kingdom standard gamble value set (Brazier, Roberts et al. 2002). These results show the greatest percentage increase from No-ART to ART out of presented data.

Table 3: Comparison of HRQoL values

Survey setting	Khayelitsha	Khayelitsha	Khayelitsha	Groote Schuur / Guguletu
Survey instrument	EQ-5D	EQ-5D	EuroVAS	SF36
Survey valuation	TTO, Dolan et al 1995	TTO, Jelsma et al 2002	Patients	SG, Brazier et al 2002
Health states:				
ART 0-3 months	0.71	0.77	0.62	
ART 3-6 months	0.81	0.83	0.71	0.81
ART 6-12 months	0.82	0.84	0.74	
ART >12 months	0.85	0.86	0.76	
No-ART	0.71	0.77	0.62	0.69
Percentage increase: No-ART to 3-6 months on ART	14%	8%	15%	17%

Sources:
Khayelitsha HRQoL measurement using EQ-5D and EuroVAS: Jelsma, MacLean et al. (2005)
Groote Schuur measurement using SF36: O'Keefe and Wood (1996)
Guguletu measurement using SF36: Pitt, Badri et al. (2005)
TTO values: Dolan, Gudex et al. (1995)
TTO values: Jelsma, Hansen et al. (2002)
SG values: Brazier, Roberts et al. (2002)

The second difficulty with valuing HRQoL data is that no local valuation set is available. In this instance, the EuroQol group recommends using the United Kingdom tariff set which has been derived from a general population survey (Dolan, Gudex et al. 1995). This would imply assuming that people in different countries valued health states similarly. While this is unlikely to be the case, the important issue is whether it is acceptable to make this assumption. To illustrate potential country differences, Table 4 contains a comparison of the scores on 8 randomly selected health states that were valued using the TTO method in general population samples in the United Kingdom and the United States and an urban sample from Zimbabwe. In the EQ-5D system, health states correspond to five domains and three levels of domains. Thus a health state denoted 12211 corresponds to "No problems walking about", "Some problems washing or dressing self", "Some problems with performing usual activities", "No pain or discomfort", and "Not anxious or depressed". This comparison shows that health state values were higher in the United States and Zimbabwe than in the United Kingdom.

Table 4: Comparison of TTO valuations of 8 selected health states in the United Kingdom, United States of America and Zimbabwe

	Health state valuations using TTO			Differences	
Health state description	United Kingdom (UK) ±	United States of America (USA) †	Zimbabwe §	UK - USA	UK - Zimbabwe
11121	0.80	0.83	0.83	-0.03	-0.03
11122	0.73	0.80	0.78	-0.08	-0.06
22121	0.62	0.74	0.68	-0.12	-0.06
21232	0.09	0.40	0.45	-0.31	-0.36
23321	0.15	0.38	0.41	-0.23	-0.26
22331	-0.03	0.31	0.31	-0.34	-0.34
23232	-0.13	0.20	0.22	-0.33	-0.35
33333	-0.59	-0.11	-0.15	-0.49	-0.44

Sources:
± Dolan, Gudex et al. (1995)
† Shaw, Johnson et al. (2005)
§ Jelsma, Hansen et al. (2002)

Some broad conclusions can be gained from this discussion:

- While the UK TTO valuations of EQ5D health states were lower than the Harare and US values, these were most similar to the Khayelitsha EUROVAS values
- The UK TTO values of EQ5D health states were similar to standard gamble values on SF36 health states

One could conclude from this that it would be appropriate to use either EuroVAS or UK TTO values. A further justification is that these were similar to SF36 data valued using the standard gamble method. This is not to imply that the standard gamble method is a gold standard, but rather that the use of the EUROVAS or UK TTO values on the EQ-5D health states seems justifiable given that a different instrument (SF-36) using a different valuation method (standard gamble) has produced highly comparable results. Primary data provide little guidance on whether it is more appropriate to use EUROVAS or UK TTO. The use of the latter will therefore be defended because using non-patient valuations is recommended for resource allocation decisions for society (Torrance and Feeny 1989).

2.5. Markov modelling

Because ART is a long-term intervention for which primary outcome data will be unavailable for the foreseeable future, it is necessary to extrapolate from available data to estimate life-expectancy and lifetime costs. This extrapolation has been undertaken via Markov modelling, which is suited to modelling long-term stochastic processes – random processes which evolve over time – and is particularly appropriate to modelling the progression of chronic diseases. A modelling approach also allows the synthesis of data from secondary sources and the extrapolation of primary data. A Markov model consists of a number of mutually exclusive and collectively exhaustive Markov states – each patient can only be in one state at a time (Kuntz and Weinstein 1997). At least one of these is an "absorbing state" from which no outgoing transitions are possible. Patients remain in each Markov state for an equal increment of time, called a Markov cycle, before being allowed the option of moving to a different state (or staying in the current state) as determined by a transition probability. States are distinguished from each other on the basis of different health characteristics (such as HRQoL) and different health care costs. While it seems contradictory to associate costs with a health state, these are kept separate in the calculation. For example, the costs associated with being in an acute-disease Markov state over a one-month period include health care resources required to treat this acute disease, while on the outcome side one is assumed to accrue one month of life-expectancy which is weighted by an appropriate HRQoL value.

At the end of each Markov cycle, patients move between states or stay in their current state as determined by transition probabilities. As mentioned above, Markov states are also associated with different risks of particular clinical events such as death, which are captured in the transition probabilities. Finally, when the model is run over a large number of cycles, long term costs and consequences of the disease under different interventions are calculated (Sonnenberg and Beck 1993; Briggs and Sculpher 1998). Sections 2.5.2 to 2.5.6 provide additional details about each element that is required to construct the Markov model.

One key shortcoming of a Markov model is that there is no memory of previous cycles – a patient's history is lost at the end of each cycle, with the implication that the probability of moving out of a state is independent of previous health states. This is known as the Markovian assumption. Because this is a limitation in certain circumstances, modellers make use of temporary states, called tunnel states, so that cost adjustments can be made or temporarily

different transition probabilities can be applied (Sonnenberg and Beck 1993; Briggs and Sculpher 1998).

2.5.1. Markov cycle length

There is no agreement about the appropriate cycle length for Markov models of HIV/AIDS in the literature – some use 1 month (Freedberg, Scharfstein et al. 1998; Freedberg, Losina et al. 2001; Schackman, Goldie et al. 2001; Richter, Hauber et al. 2002; Schackman, Freedberg et al. 2002; Badri, Cleary et al. 2006), some use 6 months (Sendi, Bucher et al. 1999) and others use 12 months (Miners, Sabin et al. 2001; Tebas, Henry et al. 2001). The methodological literature suggests that the cycle length needs to be appropriate for the timing of events in the disease or its life expectancy (Kuntz and Weinstein 1997). Shorter cycle lengths are required if events happen quickly – for example in malaria, a patient can be infected, cured and re-infected within a one month period which means that cycle lengths would need to be shorter than this to capture these distinct health states and events in the model. On the other hand, in HIV/AIDS, the time from initiating treatment to death can be many years. For this reason, a three-month cycle length has been chosen in this thesis because this is long enough to capture important changes in clinical outcomes (for example viral suppression).

2.5.2. Markov states

Markov models for HIV/AIDS commonly base Markov states on three different "markers" of HIV disease progression: the CD4 count, HIV RNA level (i.e. the number of copies of the virus in the body or viral load) and the WHO Stage (see Appendix A for details). Because CD4 cells are killed directly or indirectly by HIV, the number of these in the body is an important indicator of the progression of HIV. In addition, the CD4 cell count is used to qualify patients clinically for ART - in South Africa, patients are eligible for ART once their CD4 count is less than 200 cells/μl. CD4 counts are also used to evaluate response to ART (Hendriks, Satten et al. 1996). The viral load, on the other hand, is a more direct indicator of treatment success – if the viral load is undetectable, the treatment is working, but if it becomes detectable above a certain limit, the patient is assumed to be "failing" treatment and could be switched to a different regimen. However, the viral load is only an indirect predictor of mortality; the CD4 is a better predictor of disease progression than viral load (Sterling, Chaisson et al. 2001). The WHO stage can also be

an important predictor of mortality, especially in immune compromised patients. For instance, a patient with CD4<50 and WHO Stage IV (i.e. AIDS) would have a higher risk of death than a patient with the same CD4 level who has yet to develop AIDS.

Published Markov models of HIV/AIDS have used a wide combination of these markers and the choice of which to use is partly related to the modeller's discretion and access to data, and partly related to the study question that is being examined. In analyses of the cost-effectiveness of ART versus No-ART, it is fairly common to stratify Markov states in terms of the CD4 stratum and the WHO Stage (Sendi, Bucher et al. 1999). If the analysis is examining the cost-effectiveness of prophylaxis against various opportunistic infections, then Markov states would also have to be created for these diseases (Sendi, Craig et al. 1999). Some studies that examine the cost-effectiveness of initiating ART within different CD4 strata (for example, CD4>350 cells/μl, CD4 200-350 cells/μl or CD4<200 cells/μl) have only used the viral load to define the Markov states (Tebas, Henry et al. 2001) while others have used the CD4 level and WHO Stage (Badri, Cleary et al. 2006).

A more complex model that uses all three markers has been used to examine the cost-effectiveness of prophylaxis against various opportunistic infections (Freedberg, Scharfstein et al. 1998), the cost-effectiveness of ART versus No-ART (Freedberg, Losina et al. 2001) and the cost-effectiveness of different starting CD4 strata for ART (Schackman, Goldie et al. 2001; Schackman, Freedberg et al. 2002). This model classifies health states as acute, chronic or dead, and each state is stratified by the CD4 count (0-50, 51-100, 101-200, 201-300, 301-500, >500) and the viral load (>30 000, 10 001-30 000, 3001-10 000, 501-3000 and ≤500). For each combination of CD4 and viral load stratum, patients have a risk of developing an opportunistic infection (OI) and the probability of dying is dependent on the outcome of each acute OI event. Pneumocystis carnii pneumonia (PCP), toxoplasmosis, mycobacterium avium complex (MAC), disseminated fungal infections and cytomegalovirus infection (CMV) are distinct OIs in the model; remaining OIs were captured as "other OI". While a patient's history of opportunistic infections is an important determinant of outcomes, data shortages often require many published studies to relate mortality, morbidity and costs to the CD4 count (Richter, Hauber et al. 2002). In addition, it is likely that this approach does not estimate costs accurately. This is because it is assumed that all non-ARV costs occurring within a two-month window around each opportunistic infection are incurred in the Markov state relating to that OI. This is highly unlikely to be the case – primary data from inpatient costing for this dissertation indicates that patients are hospitalised

with a number of co-infections and it would be impossible to accurately allocate non patient-specific costs such as hospital overheads to each infection. In addition, the sample from which costs and transition probabilities were calculated would have to be very large.

For this thesis, the patient-level Markov models for No-ART, first-line ART and first and second-line ART have been developed in specialised software (TreeAge Pro 2005 Suite, MA, USA). In keeping with practice in the literature, the CD4 count has been used to stratify risk of death. This is operationalised by dividing the model into CD4 count strata of 50-199 cells/μl, and <50 cells/μl because these categories have been shown to be associated with different mortality rates in large cohort analyses (Hogg, Yip et al. 2001; Egger, May et al. 2002). Further heterogeneity between patients was observed to be related to the amount of time the patient had been on ART – mortality rates decreased steadily from baseline until the end of the follow-up period of 48 months. Similarly, the costs of health care are much higher in the first year because patients are relatively seriously ill and potentially require inpatient care, are undergoing frequent laboratory investigations as specified by South African "National Antiretroviral Treatment Guidelines" (2004) and require more frequent visits to the clinic.

To capture this heterogeneity adequately, it was necessary to divide the CD4-based Markov states into a number of temporary states, known as tunnel states. The use of these allowed the transition probabilities and costs to be varied as time on ART increased. During the first 6 month period, temporary states were created for each Markov cycle and for each CD4 category. After 6 months on ART, there was no longer any significant difference in terms of costs and outcomes within CD4 strata, and these states were merged. However, differences in mortality rates and costs in relation to duration on ART were still significant, necessitating the ongoing use of tunnel states between 6 and 48 months on treatment. The "Operational Plan for Comprehensive HIV and AIDS Care, Management and Treatment for South Africa" (Operational Plan for Comprehensive HIV and AIDS Care 2003) recommends first-line and second-line antiretroviral drug (ARV) regimens. After 6 months on ART, both ART models included a probability of failing the first-line regimen[22]. For the first and second-line ART model, if first-line treatment was failed, the patient transitioned to the second-line regimen – the inclusion of separate states for the second-line regimen is essential because of the higher costs of these ARVs. In the first-line ART model, the transition probability of failing first-line was added to the probability of dying with the implication that failing first-line increased the risk of death. See Figure 1.

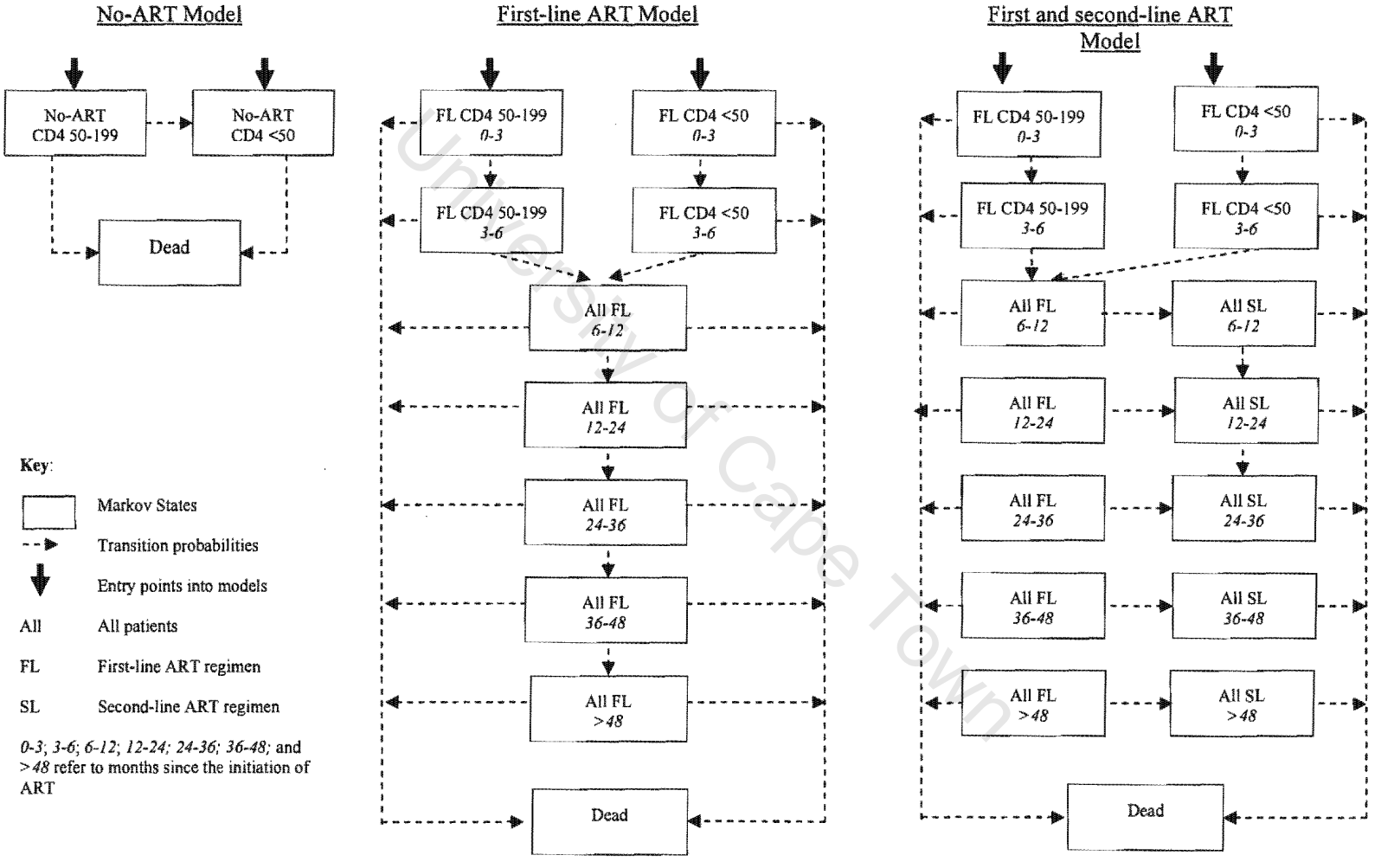
[22] No patients transitioned to second-line in the first six months.

There are a number of differences between the ART models in this thesis and the models in the literature. Most previous studies have modelled treatment effectiveness via CD4 count changes – if treatment is effective, patients are moved into higher CD4 strata, and vice versa. Khayelitsha data, on the other hand, have suggested that the CD4 count is an imperfect predictor of treatment success (Coetzee, Hildebrand et al. 2004). For this reason, overall life-expectancy has been based on a pragmatic extrapolation of the risk of death over time. In other words, effectiveness of ART is modelled as a risk of death according to time on treatment and first-line treatment failure is modelled as a risk of moving into a second-line state which is again dependent on overall time on treatment. A further difference between this ART model and others in the literature is that it only considers patients who enter care with CD4<200 cells/μl. While it is not the case that patients with CD4>200 cell/μl would never enter care, patients more frequently present for treatment and diagnosis once they are symptomatic and in terms of the government policy, ART can only be initiated with CD4<200 cells/μl. In other words, the greatest HIV-related treatment burden relates to the late stage of HIV disease.

The No-ART model was also stratified into CD4-based Markov states of CD4 50-199 cells/μl and CD4 <50 cells/μl, with death as the absorbing state. Here, progression of disease was modelled as a risk of transitioning to a lower CD4-based state with a higher risk of death. Because the life-expectancy of No-ART patients is much shorter than ART patients, median survival can be established directly from data – no extrapolation is required.

For all models, it is necessary to specify the point of entry into the model, which should be related to the aims and objectives of the analysis. Although patients can start ART in South Africa once their CD4 count drops below 200 cells/μl, in reality, many patients start much later, and it is possible that this will continue, partly in order to deal with the backlog of sick patients in the early stages of the rollout, and partly because patients present once they are relatively immune compromised. Although it is less effective and potentially also less cost-effective to start ART later, it was nevertheless assumed that patients would initiate one of the ART options in keeping with the experience in Khayelitsha – 63 per cent of patients were assumed to enter each model with CD4 50-199 cells/μl and the remainder with CD4 counts < 50 cells/μl (see Figure 1). The same starting distribution was modelled for the No-ART alternative so that each intervention is comparable in terms of the health states of patients at baseline.

Figure 7: Markov models for No-ART, first-line ART and first and second-line ART



2.5.3. Transition probabilities

Transition probabilities in a Markov model are required to specify all possible movements between Markov states. In Markov chains, transition probabilities are constant over time. If transition probabilities vary with time, the model is called a time-dependent Markov process (Sonnenberg and Beck 1993; Briggs and Sculpher 1998).

In general, transition probabilities have been estimated using Kaplan Meier product limit estimates which give the proportion of the cohort in a particular state (e.g. surviving or remaining on first-line) at a particular point in time. All probabilities need to be calculated per three-month period, but at times, the Kaplan Meier was calculated over a longer time frame. For example, if one knew that the proportion surviving at 6 months and 24 months were 0.8 and 0.2 respectively, then the probability of dying per three-month period between months 6 and 24 would be:

$$P_{dead} = 1 - \left[e^{\ln(KM_2 / KM_1)/(t_2 - t_1)\ cyclelength} \right] \qquad \textbf{Equation 4}$$

Therefore

$$P_{dead} = 1 - \left[e^{\ln(0.2/0.8)/(24-6)\ 3} \right] = 0.76$$

Where:
KM_2 is the proportion surviving at second time period
KM_1 is the proportion surviving at first time period
t_2 is the value of second time period in months
t_1 is the value of first time period in months
cycle length is the Markov cycle length for which transition probabilities are calculated

In the ART model, transition probabilities were estimated from Kaplan Meier product limit estimates of survival for 1,729 patients receiving ART in the first 48 months of the Khayelitsha

programme. There were 2,229 ART patient years of follow-up over a median period of 1.03 years (IQR 0.68 – 1.70, max 4.08). The probability of dying was calculated directly from primary data and specified separately for each three-month cycle over the first 6 months on treatment, per 6-month period for months 6-12 and per annual period in months 12-24, 24-36 and 36-48. This allowed an accurate specification of the decline in mortality over four years on ART. Patients who were lost to follow-up were conservatively treated statistically as deaths. Because life-expectancy might be higher in pilot settings, the potential overestimate of mortality would increase the generalizability of results to routine services. Non HIV-related (all cause) mortality was captured from South African life tables (Koch 2003) corrected for socio-economic status, age and gender. The probability of transitioning to the second-line Markov states was calculated from primary data separately for months 6-12, 12-24, 24-36 and 36-48. No patients switched to second-line between months 0 and 6.

Despite having up to four years of follow-up data on ART, there are still a number of uncertainties about the long-term costs and effectiveness of this intervention. After four-years of follow-up, over 70 per cent of patients were alive and remaining in care. To extrapolate these data in order to calculate lifetime costs and outcomes, a number of assumptions were made. In all cases, assumptions were relatively conservative. Firstly, the probability of dying after 48 months was assumed to be the average probability of dying between months 0 and 48. An alternative would have been to calculate an average over months 36 to 48 for example, which would have resulted in much higher life expectancy given the low rate of death once patients have been on ART for some time. Given the many uncertainties, it was considered better potentially to underestimate life-expectancy.

Secondly, the probability of switching to second-line was assumed to be the average probability as observed from primary data between months 0 and 48. Whether a patient is on first or second-line in the model does not have an impact on life-expectancy, but does have important implications for lifetime costs (as the second-line ARVs are more than twice as expensive as the first-line). Thirdly, there is currently very limited information about whether patients will die while on treatment, or whether they will cease treatment after virological failure on the second-line regimen. Given the unavailability of these data in the near future, it has been assumed that all patients would remain on treatment but that transitioning to death would incur an additional treatment cost in order to capture any inpatient care costs at the appropriate point in time.

No-ART outcomes were derived from the Cape Town AIDS Cohort, a local natural history cohort of ART-naïve patients[23] who presented at a specialised HIV-clinic established at the New Somerset secondary hospital in Cape Town between 1992 and 2000 (Post, Wood et al. 1996; Badri, Bekker et al. 2004). These HIV clinics served largely indigent patients who were referred from a wide range of public sector primary health care facilities (Badri, Wilson et al. 2002). Patients were followed up 3 to 6 monthly or when clinically indicated (Badri, Bekker et al. 2004). Given the likelihood that many patients in this cohort lived in similar circumstances to patients in Khayelitsha, the ethical limitations in following a matched No-ART cohort in Khayelitsha, and the lack of other No-ART cohorts, the use of these data to calculate No-ART life-expectancy is justified. One of the key shortcomings of using these data is that it is possible that current No-ART patients have worse prognosis than those in the Cape Town AIDS Cohort. This is because access to care might have worsened owing to the far greater number of patients that are in need. This potential shortcoming is addressed through sensitivity analysis.

These secondary data sources provide estimates of the rate of death for patients with CD4<200 cells/μl and CD4<50 cells/μl, but do not provide rates of death for patients with CD4 50-199 cells/μl or the rate of switching from higher to lower CD4 categories (i.e. 50-199 to <50 cells/μl). These probabilities have been calculated by using the hazard ratio for death for patients with CD4<50 cells/μl in Khayelitsha (2.28; p=0.021) (Coetzee, Hildebrand et al. 2004). The probability of dying in the CD4 50-199 category was calculated by dividing the probability of dying in CD4<200 cells/μl by 2.28. The probability of switching between CD4 50-199 cells/μl and CD4<50 cells/μl was then calculated to ensure that the overall survival matched survival for patients with CD4<200 cells/μl.

Sources of data for transition probabilities are summarised in Table 5.

[23] ART-naïve patients have never received ART.

Table 5: Sources of data and assumptions used in calculating transition probabilities

Input parameters		Data source and assumptions
Probabilities of transitioning to dead Markov states		
CD4<50 cells/ml, starting ART		
	0-3 months	Khayelitsha pilot
	3-6 months	Khayelitsha pilot
CD4 50-199 cells/ml, starting ART		
	0-3 months	Khayelitsha pilot
	3-6 months	Khayelitsha pilot
All patients on ART, irrespective of regimen		
	6-12 months	Khayelitsha pilot
	12-24 months	Khayelitsha pilot
	24-36 months	Khayelitsha pilot
	36-48 months	Khayelitsha pilot
	>48 months	Average mortality rate in 0-48 months
No ART, CD4 count <50 cells/ml		
	all quarters	Calculated from Cape Town AIDS Cohort
No ART, CD4 50-199 cells/ml		
	all quarters	Calculated from Cape Town AIDS Cohort
Probabilities of transitioning between alive Markov states		
First-line regimen to second-line regimen		
	0-6 months	N/A - no patients failed during this period
	6-48 months	Khayelitsha pilot
	>48 months	Average failure rate in 0-48 months
No-ART, CD4 50-199 cells/ml to CD4<50 cells/ml		
	all quarters	Calculated from Cape Town AIDS Cohort using death hazard ratio from Khayelitsha pilot

2.5.3.1. Non HIV-related mortality

Transition probabilities should include an all-cause mortality risk that captures death from non-HIV related causes (Briggs and Sculpher 1998). Non HIV-related mortality was derived from pre-HIV era South African life tables (Koch 2003) to avoid double-counting HIV-related deaths. Starting with a hypothetical cohort of 100,000 people at age zero, these life tables show the proportion of the cohort that remains alive until age 100, when everyone is assumed to die. Mortality is expressed separately for men and women in six different socio-economic groups. In order to calculate non HIV-related mortality, it was therefore necessary to establish the average age, sex, and socio-economic status of HIV patients.

These data were derived from a study of 749 people on ART in clinics in the Western Cape, located in Guguletu (Guguletu CHC and GF Jooste Hospital), Khayelitsha (Michael Mapongwana CHC), Hout Bay (Hout Bay clinic) and Paarl (TC Newman CHC) (Pienaar, Myer et al. 2006). Many of these patients are enrolled in routine care services, and it is therefore likely that they are fairly representative of the HIV population that is utilising care at this stage of the rollout in the Western Cape, but they are unlikely to be representative of the country's HIV population in general, in particular in terms of socioeconomic status, for two reasons:

- The Western Cape is a relatively wealthy province. However, 72 per cent of the sample was born outside of the Western Cape and 43 per cent was highly mobile, spending considerable time each year in other parts of the country.
- The sample was from clinics located primarily in urban areas so would not be representative of rural HIV-positive people. However, given that HIV-prevalence is highest in poor urban areas, the sample might be representative of the majority of HIV-positive South Africans[24]

While this sample might be overestimating the socio-economic status of HIV-positive South Africans who are dependent on public health services, until a nationally representative sero-status linked household survey collects these data, these Western Cape data will continue to be the best available.

In order to assess the socio-economic status of ART patients, an asset index was created following a methodology developed by Filmer and Pritchett (2001) based on asset questions in the 1998 adult Demographic and Health Survey (DHS) dataset. This methodology has been applied to the South African 1998 DHS data and the resultant asset index was found to be an internally coherent and robust indicator of poverty in South Africa that compared well to mainstream poverty indicators such as income and expenditure (Booysen 2001). The first step in this methodology is to recode asset variables into dichotomous variables, distinguishing between households that own the asset and those that do not (in other words, variables take the value of

[24] Data from a 2005 household survey for respondents aged 15-49 indicated highest prevalence in urban informal areas (25.8 per cent), followed by rural informal (17.3 per cent), and urban and rural formal areas at 13.9 per cent each. Urban informal areas have informal or temporary housing made out of tin, cardboard and wood. During Apartheid, squatters lived illegally in these areas.

Once the asset indices have been calculated, six socio-economic groups were constructed in the DHS and tabulated in order to ascertain the range of asset indices falling into each group. The next step was to map these results onto the ART sample using the same means, standard deviations and factor scores as had been calculated from the DHS sample. This allows one to estimate the socio-economic strata of ART users in comparison to households from the DHS. Non HIV-related mortality rates were then calculated based on the proportions of men and women in the different socio-economic groups and were transformed into three month transition probabilities using Equation 4.

2.5.4. Attaching weights to the Markov model

Weights are attached to each Markov state in the model to represent the cost and health outcomes associated with each state. For instance, if one is modelling life-expectancy using a three-month Markov cycle, a weight of 0.25 (representing three months in a year) would be attached to each state of the model in which the patient is alive and a weight of 0 would be attached to the dead state. If one were modelling QALYs, then the weight in each alive state would be multiplied by the appropriate HRQoL value. Similarly, the cost of being in each Markov state is computed for the length of the Markov cycle and is attached to the state. While it seems counterintuitive to attach costs to a health state, this is the terminology that is used in the modelling literature to describe the way in which costs can be associated with a particular state of health over the Markov cycle. When the model is run over a large number of cycles, each weight is summed across those cycles to give an estimate of lifetime QALYs and costs (Briggs and Sculpher 1998).

2.5.5. Adjustments to costs and outcomes

Since Markov models deal explicitly with time, they allow discounting of costs and outcomes at the point in time that these occur in the model. This is done by attaching a discount rate to each Markov state (Briggs and Sculpher 1998). This thesis follows common practice for developing countries in the literature, which is to use an annual rate of 3 per cent for both costs and effects (for example, see Goodman, Coleman et al. 1999). However, a zero rate is also used in sensitivity analysis.

Markov models assume that transitions between health states occur between cycles, and during the cycle patients remain in their current state. In reality, patients move between phases of a disease constantly. The error associated with using discrete as opposed to continuous time can be corrected during the first cycle of the model using one of three approaches. If no correction is performed, no rewards are accumulated during the first cycle of the model, thereby underestimating both costs and effects. If a full cycle correction is made, all patients receive full costs and effects during the cycle. If a half cycle correction is made, all patients receive half the costs and half the effects. In choosing which correction to make, one should consider the proportion of patients who die during the cycle. If the majority of patients die, then no cycle correction is appropriate, if half die, then a half cycle correction is appropriate but if most survive, then a full cycle correction should be made. The last was chosen because 90 per cent of patients live through the first 3 months in both ART and No-ART groups.

At times it is necessary to make adjustments to costs that are not incurred by all patients in a Markov state. These can be modelled as transition costs which relate to specific transitions between states. In this thesis, a cost of dying was specified as a transition cost incurred by any patient who died of HIV-related causes[25]. This approach ensures that the resources that are utilised at the time of death are not underestimated when primary data are extrapolated, which is particularly relevant for ART in this thesis given that data are only available for a maximum of four-years, during which time less than 30 per cent of the sample died. An example illustrates this point. Imagine calculating inpatient utilisation from 100 patients followed over 100 patient-years. During this time, 100 inpatient days were accrued by 10 patients who died resulting in an average inpatient utilisation of 1 per patient-year. However, if one were to apply these data to health states with higher rates of death, one would underestimate costs. A more robust approach would be to calculate separate costs for the patients who live, which would be attached to the alive Markov state, while the cost of dying would be associated with transitioning to the dead Markov state .

2.5.6. Evaluating the model

Two methods of evaluation are appropriate to Markov processes with time-dependent transition probabilities: cohort simulation and first-order Monte Carlo simulation. To understand the mechanics of cohort simulation, imagine that a hypothetical cohort of 10,000 patients enters the

[25] Patients dying of non-HIV related causes do not incur this cost.

model at time zero and is distributed into appropriate CD4-defined Markov states. After each cycle, the cohort will be distributed to different states as required by the transition probabilities attached to each state. When the model is run over sufficient cycles, lifetime costs and outcomes are calculated.

First-order Monte Carlo simulation works differently in that it follows the same large group of patients through the model one at a time. Each "patient" takes a different path through the model and when one considers the paths of a number of patients, one gains an overall profile of costs and outcomes. The cohort simulation method is more precise than Monte Carlo simulation because it gives an exact (expected value) solution whereas Monte Carlo simulation will never give the same result twice owing to the random nature of the simulation. The advantage of Monte Carlo simulation is that it gives an estimate of the likely variance associated with the structure of the model. Both evaluation methods will be used to calculate results.

2.6. Model validation and sensitivity analysis

2.6.1. Validating the model

Because models are a simplified version of reality, it is important that they describe the real world to an acceptable level. Systematic validation of a model can help to achieve this goal. Sendi, Craig et al. (1999) propose four sequential levels of model validity: technical validity, predictive validity, face validity and modelling process validity.

Technical validity involves identifying and correcting for modelling bugs such as unexpected model behaviour, redundant variables, programming errors and typing errors (Sendi, Craig et al. 1999). A useful method of debugging a model is to vary one or more parameter over its entire range and to examine any anomalies in model outputs. This process will be undertaken for the following transition probabilities in the models:

- Probabilities of dying (ART and No-ART model)
- Probability of moving from CD4 50-199 to <50 cells/μl (No-ART model)
- Probabilities of transitioning to second-line (ART model)

Predictive validity of a model can be tested by comparing intermediate and final outcomes from the model against outcomes from published or primary data sources (Sendi, Craig et al. 1999). This will be undertaken by comparing the proportion of the cohort between 0 and 48 months that is dead in the ART model in comparison to real data at these points in time. A similar exercise will be undertaken for the No-ART model.

Face validity assesses whether the model produces the output that one would expect (Sendi, Craig et al. 1999). For instance, one would expect that patients starting ART with CD4<50 cells/μl would have lower survival than patients starting with CD4 50-199 cells/μl. This assumption can be checked against the predictions from the model.

Modelling process validity is evaluated by comparing the results and conclusions from different groups of researchers who have independently addressed the same question. Although results will differ between settings, the conclusions should be the same in terms of relative cost-effectiveness (Sendi, Craig et al. 1999).

2.6.2. Accounting for uncertainty

Briggs (1995) defines four broad types of uncertainty in economic evaluation, which relate to:

- Data requirements of the study
- Generalizability of results
- Extrapolation of data
- Choice of analytic method

Uncertainty relating to data requirements is increasingly studied within the framework of the stochastic cost-effectiveness analysis, and this approach has been employed within this thesis. This required collecting patient-level data on resource use and health outcome. Usually the resource use is assumed to be stochastic, and deterministic unit costs are used as weights on the resource items (Briggs 2001). Additional details are provided in section 2.6.2.3.

Uncertainty relating to generalizability of results is concerned with the extent to which the results of this thesis are applicable to other settings. Generalizability has been assessed by comparing ART clinical outcomes to other published cohorts in low income countries (ART-LINC and ART-CC 2006; Etard, Ndiaye et al. 2006) and No-ART outcomes to a review of natural history data (Schneider, Zwahlen et al. 2004). Inpatient and visit utilisation data have been compared to a published South African cost-effectiveness analysis (Badri, Cleary et al. 2006) and to national guidelines for follow-up of patients on ART (National Antiretroviral Treatment Guidelines 2004). Unit costs have been constructed from a wide range of primary and secondary sources to enhance generalizability. Where disagreements between this study and published data have been found, adjustments will be made with the use of simple sensitivity analysis (see section 2.6.2.1).

Uncertainty relating to extrapolation is an important problem in the evaluation of ART. After 4 years of follow-up in Khayelitsha, over 70 per cent of patients were alive and remaining in care, implying that the calculation of final outcomes requires data to be extrapolated over many cycles in the model. This uncertainty will be addressed with the aid of simple sensitivity analysis (see section 2.6.2.1) and by calculating results under three different modelling time horizons: five years, ten years and simulation until 100 per cent of the cohort is dead. On the other hand, because life-expectancy is much shorter for No-ART patients (Post, Wood et al. 1996; Badri, Bekker et al. 2004), for them no extrapolation is required to estimate final outcomes.

The final source of uncertainty relates to analytical methods. These include methods for valuing and measuring HRQoL and methods for discounting costs and benefits. This study will consider uncertainty relating to HRQoL measurement and valuation by presenting outcomes as life years and QALYs and uncertainty relating to the discount rate by varying the rates.

Uncertainty can be assessed using sensitivity analysis. A number of different types are available – each is discussed below.

2.6.2.1. Simple sensitivity analysis

Simple sensitivity analysis is the most common form of sensitivity analysis, and involves varying one or more parameters over a plausible range, while other parameters keep their base-case values. This allows the analyst to establish the separate effect of variations in each parameter on

the results (Briggs, Sculpher et al. 1994), which is useful for establishing the technical validity of the model (Sendi, Craig et al. 1999) (see 2.6.1), and for uncovering the impact of analytical methods on results (Briggs 2001).

A multi-way simple sensitivity analysis involves varying two or more parameters at the same time. This can take the form of scenario analysis (Briggs, Sculpher et al. 1994). For instance, in the case of ART, it might be useful to explore the impact of a cost-saving scenario, where patients would be assumed to receive less frequent laboratory monitoring and less expensive ARVs. This scenario would calculate the potential costs and outcomes of a less resource intensive version of ART, as recommended by the World Health Organisation for resource poor settings (WHO 2002). Current South African guidelines recommend a much more resource intensive approach (Operational Plan for Comprehensive HIV and AIDS Care 2003; National Antiretroviral Treatment Guidelines 2004). Multi-way sensitivity analysis can also be used to enhance the generalizability of results. For instance, while Khayelitsha patients are referred for tertiary inpatient care, many people in South Africa would not have access to tertiary hospitals and different hospital costs should therefore be used to enhance generalizability. Similarly, levels of patient retention and adherence might be lower in routine settings. These adjustments will all be made through multi-way scenario analysis.

2.6.2.2. Threshold analysis

Threshold analysis is concerned with identifying the critical value of parameter(s) above or below which study conclusions would change (Briggs, Sculpher et al. 1994). However, for interventions that are more effective and more costly (such as ART) this approach is only useful if one can compare the ICER to the maximum acceptable threshold, which is an approach that will not be used in this thesis for reasons outlined in Chapter 3. Alternatively, threshold analysis could be used to find the combinations of parameters that make ART cost saving (more or as effective but less costly) in relation to No-ART. However, it is highly unlikely that ART would be cost saving, because once treatment has failed, viral rebound and deterioration of the immune system makes patients once again susceptible to the HIV-related and opportunistic infections that characterise the No-ART strategy. It is therefore most realistic to assume that patients will need the same inpatient and palliative care that is typical of No-ART. Unless ARVs are less costly, and results

are discounted at a high rate, ART will not be cost saving. For these reasons, threshold analysis will not be used.

2.6.2.3. Probabilistic sensitivity analysis

Probabilistic sensitivity analysis (PSA) is a technique that is able to summarize the uncertainty in a simulation as probability distributions for model outputs, such as incremental QALYs, incremental costs or the ICER (Briggs 1995; O'Hagan, McCabe et al. 2005). To do this, parameters in a model are defined as distributions. In this dissertation, PSA was undertaken by specifying distributions on all transition probabilities and utilisation estimates. The impact of parameter uncertainty is propagated through the model by means of Monte Carlo simulation – this involves running the model many times using randomly selected values from input distributions in order to derive probability. This type of Monte Carlo simulation is designed to capture second-order uncertainty – the variability in the parameter of interest as opposed to the variability in the underlying population (this is first-order uncertainty – see section 2.5.6) (Briggs 2001; O'Hagan, McCabe et al. 2005).

At this point, it is necessary to differentiate between frequentist and Bayesian interpretations of probability. The PSA approach is inherently Bayesian in nature (Briggs 1999; Briggs 2001; O'Hagan and Luce 2003; O'Hagan, McCabe et al. 2005). Under a frequentist approach, parameters are considered to have true values and not to vary. For a frequentist, only repeatable events have probabilities. However, when undertaking PSA, it is assumed that parameters are random variables, which can take a range of values defined by the chosen distribution. For a Bayesian, probability describes uncertainty, which can be owing to intrinsic unpredictability, random variability, or because of imperfect knowledge (O'Hagan and Luce 2003). Bayesians also distinguish between prior distributions, likelihood distributions and posterior distributions. The prior distribution is based on what is known about a problem prior to viewing the new or current data. However, it is more helpful to think of the prior as summarising all external evidence about a quantity of interest (Spiegelhalter, Myles et al. 1999). Information from the prior distribution is then synthesised with the current data (as expressed in the likelihood distribution) to produce a posterior distribution (O'Hagan and Luce 2003). This synthesis is undertaken via Bayes' Theorem. If the prior information is less strong than the current information (as indicated by a higher variance) Bayes' theorem puts less weight on the prior than the likelihood distribution.

When the current data are strong relative to prior information, then the prior is unlikely to make a big impact on the posterior, and it is common to apply a "non-informative" prior. In this circumstance, the likelihood distribution is the same as the posterior. A key benefit of the Bayesian approach is its ability to make use of information in a transparent manner, in order to produce stronger results. This is because a posterior distribution will be a more precise estimate with lower variance than the likelihood distribution (O'Hagan and Luce 2003; O'Hagan, McCabe et al. 2005). However, in the case of this thesis, there is no alternative source of South African data on most of the parameters of interest, which justifies the assumption of a "non-informative" prior.

It is important to note that parameters might not be statistically independent, implying that it is necessary to think about correlation between parameters in the model. In other words, a joint probability distribution for all the parameter values is required. However, most analyses pragmatically assume independence (O'Hagan, McCabe et al. 2005), and this approach has been taken in this study.

Key differences between Bayesian and frequentist approaches are summarized in Briggs (1999). He identifies three main types of Bayesian methods based on the approach to prior information. Empirical Bayes bases prior information on previously available statistical information. The use of this information can be related to the frequentist approach to pooling available data. A second approach would come about if there were no prior information on the parameter of interest. This is the same as a frequentist approach when there are no data to pool. The final approach is called "subjective Bayes", which involves eliciting prior information from experts on the basis of their personal beliefs. To clarify, assume that the parameter of interest were the effectiveness of ART in South Africa, and that current data were derived from a large cohort study. Under an empirical Bayes approach, these current data could be updated with prior information from a smaller pilot study of ART in South Africa. However, if no such pilot had been undertaken, or if there were no published data, a non-informative prior could be assumed. Finally, a subjective Bayes approach would incorporate expert opinion as the prior. To summarize, the approach to PSA will be to assume an non-informative prior because data are derived from the first public sector pilot of ART in South Africa.

3. Methodology for population-level analyses

This section presents methodology for establishing the need for HIV-treatment, population-level costs and outcomes associated with alternative HIV-treatment interventions, methods for using linear programming to solve for HIV-treatment social choice rules and methods for estimating the socio-economic status of ART users in order to determine whether the service has the potential to preferentially benefit the poor. The section also considers the impact of the programme size on the costs of scaling up.

3.1. Estimating need

It has been assumed that a patient is in need of HIV-treatment when she or he develops AIDS (WHO Stage IV). The number of adults developing AIDS each year has been estimated from demographic modelling of the South African HIV-epidemic. The key data used in these models are from antenatal clinics, where blood samples of pregnant women are anonymously tested for HIV. However, these data are not directly generalisable to the population as a whole because for example pregnant women would not necessarily be representative of men, or of people who are not sexually active such as the youth and the elderly. On the other hand, there is evidence that HIV might affect the fertility of women, meaning that HIV-positive women are less likely to fall pregnant (UNAIDS and WHO 2005). Antenatal prevalence data therefore need to be adjusted to give estimates for the broader population through demographic modelling. In South Africa, a range of demographic models to achieve this have been released by the Actuarial Society of South Africa (ASSA). The latest model, ASSA2003, has been calibrated against the 2003 antenatal clinic surveys, as well as data from the 2002 Human Sciences Research Council (HSRC) national HIV survey (Shisana and Simbayi 2002) and the Reproductive Health Research Unit survey on the sexual behaviour and prevalence of HIV in youth in South Africa (Dorrington, Johnson et al. 2004).

It should be noted that it is necessary to use estimates of new need (new AIDS cases) as opposed to total AIDS cases in the modelling which means that need is slightly underestimated within the first year of the projection, but this underestimation should be negligible thereafter. While ASSA2003*lite* also provides estimates of total AIDS cases, using these data would be an

overestimate of need. For example, imagine that in year 1 and year 2 there are 100 and 110 total AIDS cases respectively. Year 2 cases are comprised of new AIDS cases and any patients surviving from year 1. If one assumes that need is 100 in year 1 and 110 in year 2, one is counting a proportion of year 1 need twice. On the other hand, the clinical criteria for ART initiation are CD4<200 cells/μl at any WHO Stage or an AIDS diagnosis at any CD4 level. The implication is that using new AIDS cases as an estimate of need will underestimate the number of people who clinically qualify for ART because many patients with CD4<200 cells/μl will be in other (non-AIDS) WHO Stages. Despite these shortcomings, these data are the best available and their use is therefore justified.

In an alternative scenario, need has been based on the Operational Plan's (2003) target of patients initiating ART. The plan does not distinguish between adults and children. Because this thesis is focusing only on adult HIV-treatment, the patient targets have been adjusted to exclude children by calculating the average proportion of new adult AIDS cases out of total new AIDS cases over the decision-making time frame from ASSA2003*lite* estimates.

The plan initially envisaged starting approximately 53,000 adults and children (49,290 adults) on ART by March 2004 and to reach 100 per cent of new AIDS cases by March 2009. This was ambitious given that the rollout was only announced in October 2003, the Operational Plan was only finalized in November 2003, National Antiretroviral Treatment guidelines were only available in 2004 and facilities had to be accredited and drug procurement systems had to be set up. In his State of the Nation Address delivered after the April 2004 general elections, President Mbeki stated that this initial patient target would be met by March 2005 (Mbeki 2004), which implies that the Operational Plan targets were to be lagged by one year. This alternative patient need scenario will therefore assume that patient uptake is 49,290 adults by March 2005, that 100 per cent of new adult AIDS cases are met by March 2010 and that 100 per cent of need continues to be met in subsequent years, based on ASSA2003*lite* estimates.

3.2. Population-level costs, QALYs and social choice rules

While methods for patient-level costs and outcomes focussed on the calculation of lifetime costs and individual health gains, at the population-level one is concerned with calculating the total costs and outcomes of scaling up each intervention to a defined population in need over a given

decision-making time frame. While the overall structure of the Markov models is the same as described at the patient-level, in order to estimate population-level results, new patients in need of treatment are assumed to enter the Markov models during each cycle of the planning period. Because the specialised software (TreeAge Pro 2005 Suite, MA, USA) used in the patient-level models cannot accomplish this, population-level models have been developed in Microsoft® Office Excel 2003. By inputting need into the model during each Markov cycle, one is able to calculate total and annual costs and total and annual health gains from each mutually exclusive HIV-treatment intervention.

These total costs and total outcomes over a specified planning period are then inputted into the mathematical programming algorithms outlined in Chapter 3. This information allows one to calculate the health gain that can be achieved and the percentage of patients that can be treated in total and in each treatment strategy over a range of budget constraints in each social choice rule.

3.3. Estimating the socioeconomic status of ART users

As described in section 2.5.3.1, the socioeconomic status of ART users has been based on 749 patients on ART in five clinics in the Western Cape (Pienaar, Myer et al. 2006). The key difference in methods in this section as opposed to above is that the socioeconomic status of the ART sample has been compared to the full DHS and to a sub-sample of the DHS consisting of adults living in informal settlements around Johannesburg and Cape Town, as constructed and described in Thiede, Palmer et al. (2005). This allows one to compare the socioeconomic status of ART users to first the population of the country as a whole and second a sample representative of the communities from which they originate, in order to assess whether ART has the potential preferentially to benefit the poor within a particular community.

3.4. Impact of programme size and implementation on costs

According to Luce and Manning (1996) a significant increase or decrease in the level of an intervention can lead to economies or diseconomies of scale, economies based on learning-by-doing, or economies of scope. Economies of scale (increasing returns to scale) means that the

long run average total cost decreases as scale increases; if there are constant returns to scale, output increases by the same proportion as the proportionate increase in all the inputs and unit costs are unchanged; if there are diseconomies of scale (decreasing returns), output increases by a smaller proportion than the increase in inputs (Clewer and Perkins 1998). Economies of scope occur when the cost of joint production is less than the cost of producing several related items separately (Clewer and Perkins 1998). Economies based on learning by doing refer to increased productivity as health professionals for example become more experienced at a particular activity (Luce, Manning et al. 1996).

Economies of scale at the micro level within HIV-treatment facilities could arise through spreading fixed costs across larger numbers of patients. It is assumed for the purposes of this thesis that these effects have been adequately reflected through the inclusion of a number of facilities operating at different levels of scale in cost calculations. However given that the scaling up of HIV-treatment might not be proportional to the current configuration of facilities in South Africa, one-way sensitivity analysis will also be used to assess the costs of scaling up in clinics and community health centres as opposed to hospital outpatient departments where overhead and fixed costs are higher.

At the macro level, a number of costs could be affected by the size of the proposed HIV-treatment programmes. For instance, given the potential for staff shortages, the large scale implementation of ART could lead to higher salaries being offered to health professionals in order to attract sufficient staff. Given the attention that has been given to the possibility of clinical personnel constraints (Kober and Van Damme 2004), the numbers of professional nurses and medical officers required to reach HIV-treatment targets will be calculated. Similarly, there is a need massively to increase the capacity in the national laboratory services to meet laboratory testing requirements. It is unclear whether this increase in capacity would decrease or increase unit costs. On the other hand, antiretroviral drug costs appear to be negatively related to the scale of programmes globally, the overall time since coming to market, the level of competition from generic manufacturers and political and social pressures on multinational pharmaceutical companies. These factors have lead to dramatic decreases in prices of the antiretrovirals that are currently prescribed in first-line regimens, but there has been less movement in the second-line drug prices (Luchini, Cisse et al. 2003). Many would therefore anticipate that additional price reductions could be realised. However it is also possible that future ART guidelines will include

Chapter 5: Unit Costs of HIV-related services

1 Introduction

The aim of this chapter is to pool primary and secondary data from a variety of sources to calculate generalizable unit costs of HIV-related services, where unit costs can be defined as the full health care resources required to produce a unit of service output. Cost analyses in this thesis have been conducted from a public health sector perspective and costs include direct health care resources and the resources of non-governmental organisations (NGOs) that are involved in supporting HIV-treatment. All resources have been allocated to the following units of service output:

- ART services at community health centres (CHCs) or hospital outpatient departments (OPDs) (henceforward called ART visits)
- Treatment of opportunistic and HIV-related infections and events at clinics, CHCs and hospital OPDs (henceforward called No-ART visits)
- Tuberculosis (TB) treatment for HIV-positive people
- Inpatient care for HIV-positive people at secondary and tertiary level hospitals
- ARV medicine and laboratory investigation costs per patient-month or per patient-quarter on ART[1]

2 Unit cost per visit

The breakdown of the cost per ART and No-ART visit is presented in Table 7 and Table 8 and summarized in Figure 9. In each table, sample sizes for costing purposes have been presented. Sample sizes for medicines, laboratory investigations and imaging refer to the number of *visits* that were used to calculate costs based on the ingredients method. The sample size for clinical staff refers to the number of clinical consultations that were *timed*. Overheads, capital and

[1] A patient-month is defined as one month on treatment for one patient and a patient-quarter is defined as three months on treatment for one patient.

counselling sample sizes were based on the *total visits* to the facility. In the case of hospital OPDs, these sample sizes refer to the number of outpatient and trauma visits, and do not include inpatient days. Sample sizes provide information about the quantity of data that has been collected and the size of each facility and will be used to calculate weighted average costs in later sections.

As shown in Table 7, costs at Jooste, TC Newman and Groote Schuur OPDs are higher than costs at clinics and CHCs. Within clinics and CHCs, costs derived from Govender, McIntyre et al. (2000) are higher than those estimated for this dissertation[2]. Costs of laboratory investigations and clinical staff were the main driver of these differences. A likely explanation of the differences is that since 2000, laboratory test costs have fallen for a number of key HIV-related investigations such as CD4 counts and viral load tests, and current staff establishments have a higher proportion of nurses and junior doctors. Within primary data, the costs of visits at clinics are lower than costs at CHCs. The main difference relates to lower overhead costs in the former in comparison with the latter.

[2] These cost analyses were conducted between 2003 and 2005.

Table7: Unit cost per No-ART visit across facilities (US$; 2004 prices)

Venue	Khaye-litsha CHCs	Cross-roads CHC	Browns Farm CHC	Guguletu CHC	Guguletu CHC §	Welling-ton clinic	Mbe-kweni clinic	Phola Park clinic	Guguletu clinic §	Nom-zamo clinic §	Jooste OPD	TC New-man OPD	Groote Schuur OPD §
Sample sizes													
Medicines	757	11	18	16	30	4	25	8	30	30	4	23	30
Laboratory and imaging	2,652	11	18	16	30	4	25	8	30	30	4	23	30
Clinical staff	94	11	18	16	30	4	25	8	30	30	4	23	30
Overheads and capital	18,546	49,044	35,703	194,660	194,660	56,009	66,990	31,850	49,420	49,420	24,503	198,192	459,374
Counsellors	18,546												
Costs													
Medicines	3.69	2.22	2.54	1.94	2.94	1.04	1.24	1.68	4.79	2.31	2.00	1.20	3.58
Laboratory investigations	1.88	1.75	2.14	1.78	7.39	0.46	0.93	0.76	7.93	10.51	9.19	11.84	8.78
Imaging			0.52	0.59				0.38	-	1.51	2.35	2.45	
Clinical staff	6.12	2.87	2.67	2.54	6.44	5.30	5.88	2.88	7.92	7.91	8.71	11.21	9.88
Counselling staff	0.15												
Overheads	5.65	12.48	5.93	8.17	5.98	2.35	1.44	2.67	3.66	3.03	14.32	6.27	12.18
Buildings	1.00	0.39	0.24	0.29	2.61	0.31	0.31	0.31	2.79	0.77	4.51	0.69	1.33
Equipment	0.42	0.05	0.09	0.06	0.56	0.05	0.01	0.02	0.30	0.22	3.82	0.13	0.43
Staff training	0.03												
Cost per visit	**18.92**	**19.77**	**14.14**	**15.36**	**25.93**	**9.50**	**9.80**	**8.69**	**27.39**	**26.26**	**44.90**	**33.80**	**36.18**

Source: [§]Govender, McIntyre et al. (2000)

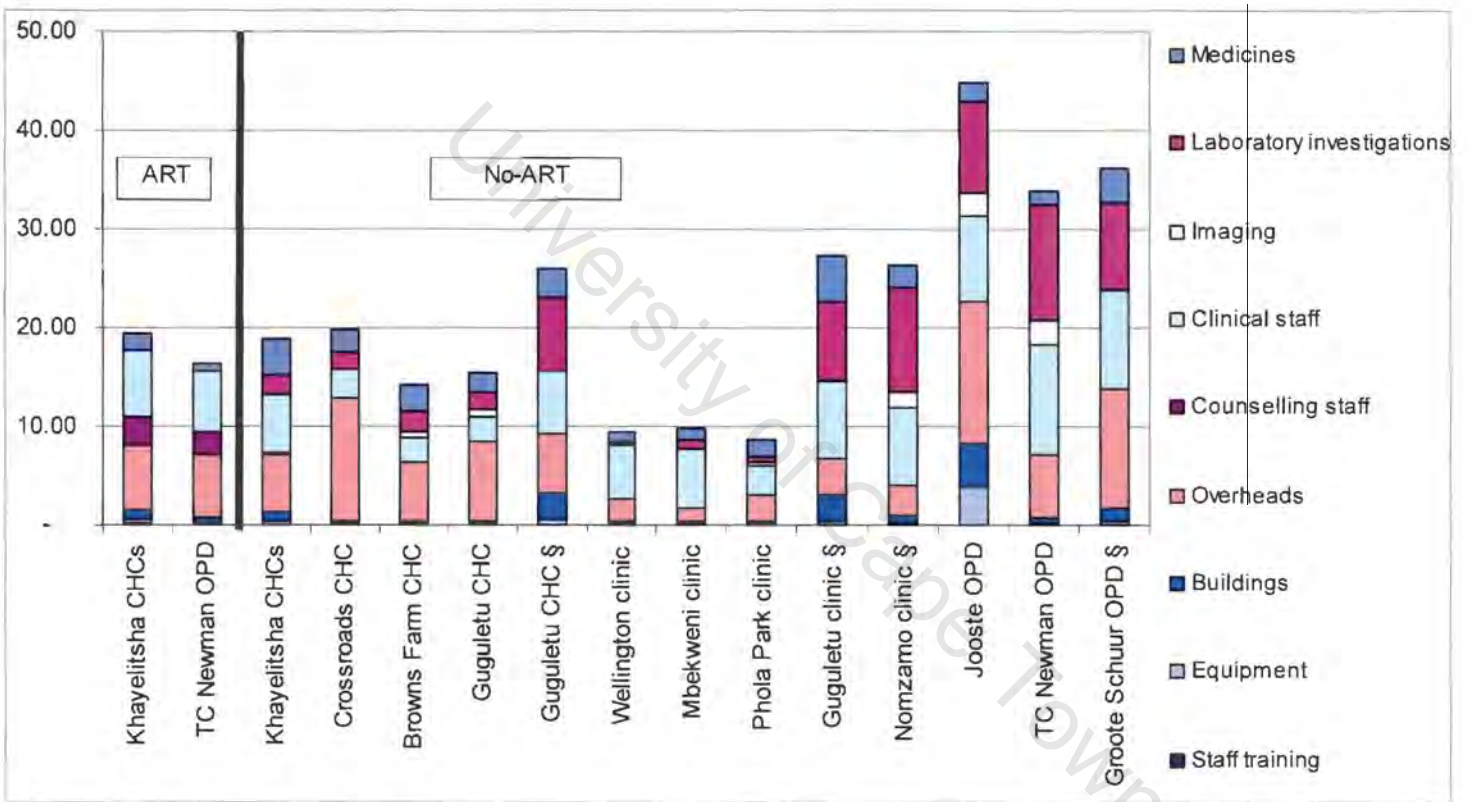
Table 8: Unit cost per ART visit across facilities (US$; 2004 prices)

Venue	Khayelitsha CHCs	TC Newman OPD
Sample sizes		
Medicines	1,532	12
Laboratory and imaging	ART protocol	ART protocol
Clinical staff	54	12
Overheads and capital	18,546	198,192
Counsellors	18,546	3,419
Costs		
Medicines	1.65	0.68
Laboratory investigations	N/A	N/A
Imaging		-
Clinical staff	6.64	6.20
Counselling staff	3.04	2.34
Overheads	6.38	6.27
Buildings	1.00	0.69
Equipment	0.42	0.13
Staff training	0.21	
Cost per visit	**19.33**	**16.31**

Table 8 contains a comparison of ART visit costs at the three Khayelitsha CHCs and TC Newman OPD. ARV medicines and laboratory investigations have been calculated separately and are not included in ART visit costs (see section 5). The key cost differences between the two settings related to non-ARV medicines and capital. Overhead costs were similar despite TC Newman being a hospital OPD.

Figure 9 contains a graphical breakdown of costs across facilities to aid comparison. It is clear from this figure that costs for clinics and CHCs from Govender, McIntyre et al. (2000) are considerably higher.

Figure 9: Unit cost per ART and No-ART visit across facilities (US$; 2004 prices)

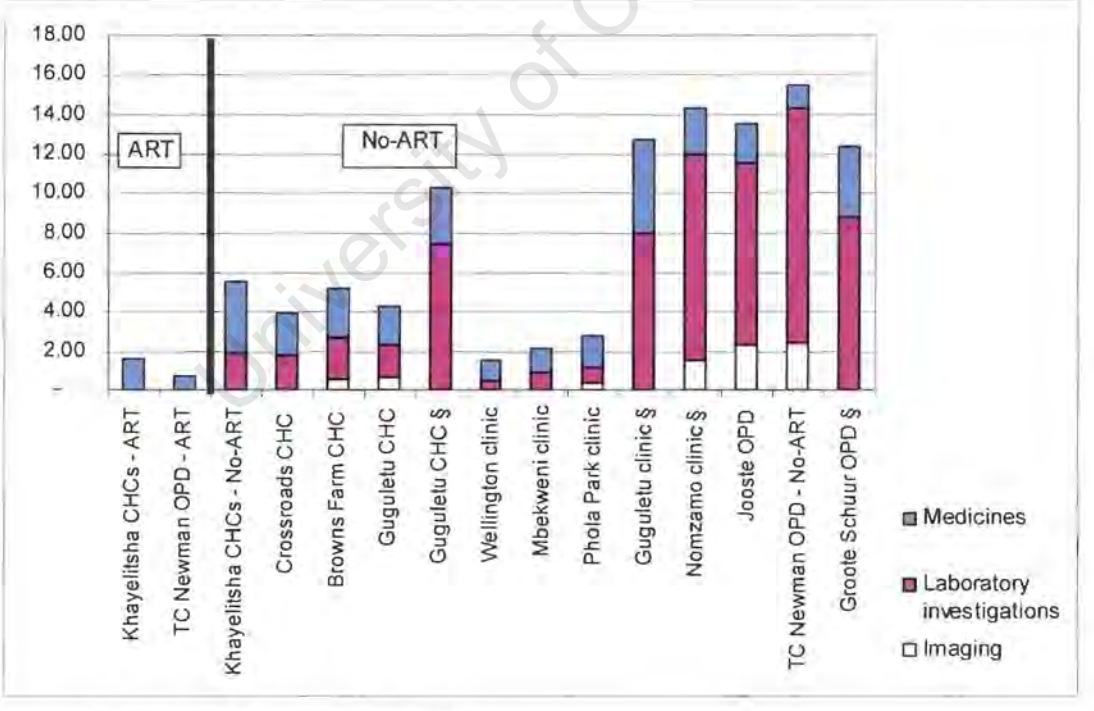


Source: § Govender, McIntyre et al. (2000)

2.1. Patient-specific cost per visit

Patient-specific costs are defined as costs that can be directly related to usage by patients, including medicines, laboratory investigations and imaging. ART visit costs do not include ARVs and laboratory investigations as it is more accurate to calculate these per patient-quarter. A breakdown of patient-specific visit costs is presented in Figure 10. The average medicine cost per ART visit in the Khayelitsha CHCs was US$1.65, and US$0.68 at TC Newman OPD. For No-ART services, there was a wide variation in medicine and laboratory investigation costs between facilities. Clinic and CHC patient-specific costs derived from Govender, McIntyre et al. (2000) were higher than primary data costs in similar facilities. If these costs are excluded, costs are highest in the hospital OPDs. No-ART costs were lowest in the nurse-driven clinic services of Wellington, Mbekweni and Phola Park.

Figure 10: Patient-specific cost per ART and No-ART visit across facilities (US$; 2004 prices)

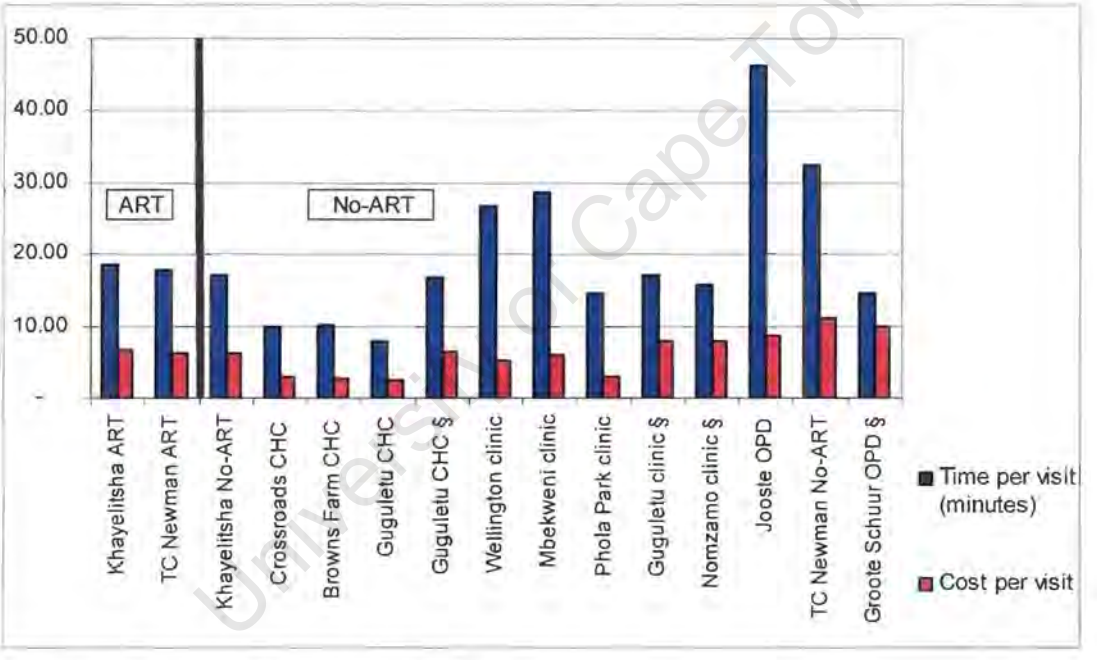


Source: § Govender, McIntyre et al. (2000)

2.2. Clinical staff cost per visit

Figure 11 shows the average time in minutes spent on clinical consultations across facilities and the resulting cost per visit. The time per visit ranged between 8 minutes at the Guguletu CHC to just over 46 minutes at the Jooste OPD. The cost per visit ranged between US$1.89 at the Browns Farm CHC to US$9.88 at the Groote Schuur OPD. The cost was not directly related to the time per visit because of different facility staff establishments - Wellington and Mbekweni had higher visit times, but their cost per visit was relatively low as these facilities are staffed only by nurses.

Figure 11: Average time (in minutes) and cost (US$; 2004 prices) per visit across facilities



Source: § Govender, McIntyre et al. (2000)

Table 9 presents a breakdown of clinical staff in selected facilities (these data were not available from secondary sources or from the Jooste OPD). Staff establishments differed widely between facilities. There were no medical officers at the clinics, and most facilities did not have a pharmacist. On average, the ratio of medical officers to professional nurses was 1.4 to 1 in ART services and 0.5 to 1 in No-ART services. ART services also had a much lower clinical staff to visit ratio, at 1 medical officer or professional nurse serving 2,450 patient visits versus 1 to 6,953

in No-ART services. This indicates that services provided within ART clinics (which include some No-ART services) were more staff intensive than general HIV (No-ART) services.

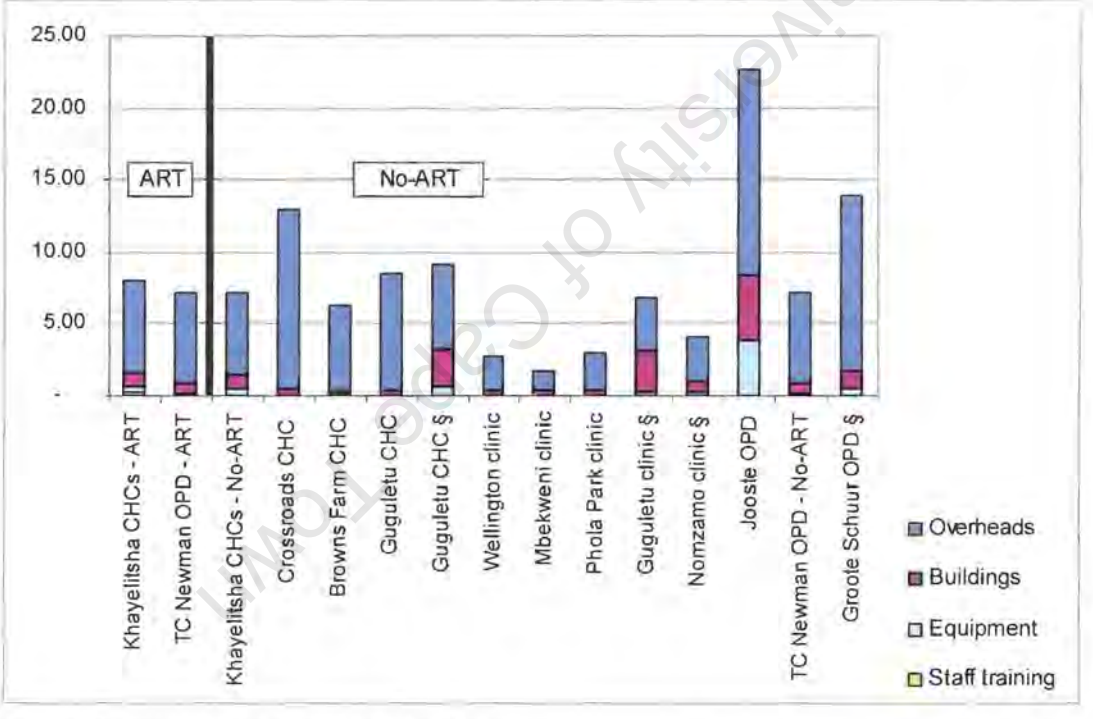
Table 9: Breakdown of clinical staff in selected facilities

		Khayelitsha CHCs	TC Newman OPD	Average ART	Crossroads CHC	Browns Farm CHC	Guguletu CHC	Wellington clinic	Mbekweni clinic	Phola Park clinic	Average No-ART
Headcount		**18,546**	**3,419**		**49,044**	**35,703**	**194,660**	**56,009**	**66,990**	**31,850**	
Clinic manager	Salary level 10						1.0				0.4
Senior medical officer	Salary level 11	3.0	0.6	2.6	2.0		9.0				4.1
Medical officer	Salary level 10	1.0	1.3	1.0			2.0				0.9
	Salary level 9				1.0	1.0					0.2
Total medical officers		4.0	1.9	3.7	3.0	1.0	12.0	0.0	0.0	0.0	5.6
Professional nurses	Salary level 8	2.0	0.6	1.8	7.0	4.0	17.0		1.0		8.2
	Salary level 7	1.0			1.0		5.0	5.0	5.0	4.0	3.7
Total professional nurses		**3.0**	**0.6**	2.6	8.0	4.0	22.0	5.0	6.0	4.0	**12.0**
Nurses	Salary level 6				3.0	1.0	11.0				4.9
	Salary level 5				5.0	2.0	16.0	2.0	3.0	3.0	8.0
	Salary level 4		0.6	0.1	1.0						0.1
	Salary level 3				1.0		1.0				0.5
	Salary level 2				2.0		4.0				1.8
Total other nurses		0.0	0.6	0.1	12.0	3.0	32.0	2.0	3.0	3.0	**15.3**
Pharmacist	Salary level 9		0.6	0.1			2.0				0.8
Total clinical personnel		7.0	3.7	6.5	23.0	8.0	68.0	7.0	9.0	7.0	**33.8**
Ratio of medical officers to professional nurses		1.3	3.2	1.4	0.4	0.3	0.5	0.0	0.0	0.0	0.5
# visits per medical officer		4,637	1,799	**4,195**	16,348	35,703	16,222	-	-	-	**11,331**
# visits per professional nurse		6,182	5,698	**6,107**	6,131	8,926	8,848	11,202	11,165	7,963	**8,865**
# visits per medical officer + professional nurse		2,649	1,368	**2,450**	4,459	7,141	5,725	11,202	11,165	7,963	**6,953**

2.3. Overhead and capital cost per visit

Overhead costs include all resources that cannot be linked to patient-utilization such as security, administrative and cleaning personnel, and water and electricity. Capital costs include buildings, equipment and initial staff training. A breakdown of overhead and capital costs is presented in Figure 12. These ranged between US$22.66 at the Jooste OPD and US$1.76 at Mbekweni. Costs are generally higher in the OPDs than in the CHCs, which in turn have higher costs than the clinics.

Figure 12: Overhead and capital cost per ART and No-ART visit across facilities (US$; 2004 prices)



Source: § Govender, McIntyre et al. (2000)

2.4. Average weighted unit cost per ART and No-ART visit

The ultimate aim of this section on visit costs is to calculate and justify a representative unit cost for ART and No-ART visits. It has been shown that laboratory costs in CHCs and clinics derived

from Govender, McIntyre et al. (2000) are higher than current costs because the costs of key laboratory investigations for HIV/AIDS have fallen dramatically. Given that their inclusion might overestimate the unit cost per visit, they have been excluded from the average cost calculations. On the other hand, the unit cost from the Groote Schuur (tertiary) OPD is in line with current calculations, and given that the mix of staffing in tertiary facilities has not changed dramatically over time, this cost is included in the calculation.

For No-ART, the calculation of weighted average costs has been done in the following manner. Firstly, the costs of clinics and CHCs have been grouped separately from OPDs. Secondly, within each group, the following methods have been used to calculate weighted average clinic/CHC and OPD costs:

- Patient-specific costs (medicines, imaging and laboratory investigations) have been weighted according to the costing sample size[3].
- Using a similar method, clinical personnel costs have been weighted according to the number of visits that were timed in each setting.
- Capital and overhead costs have been weighted according to the proportion of annual visits in *x* facility out of total annual visits from all facilities included in data collection.

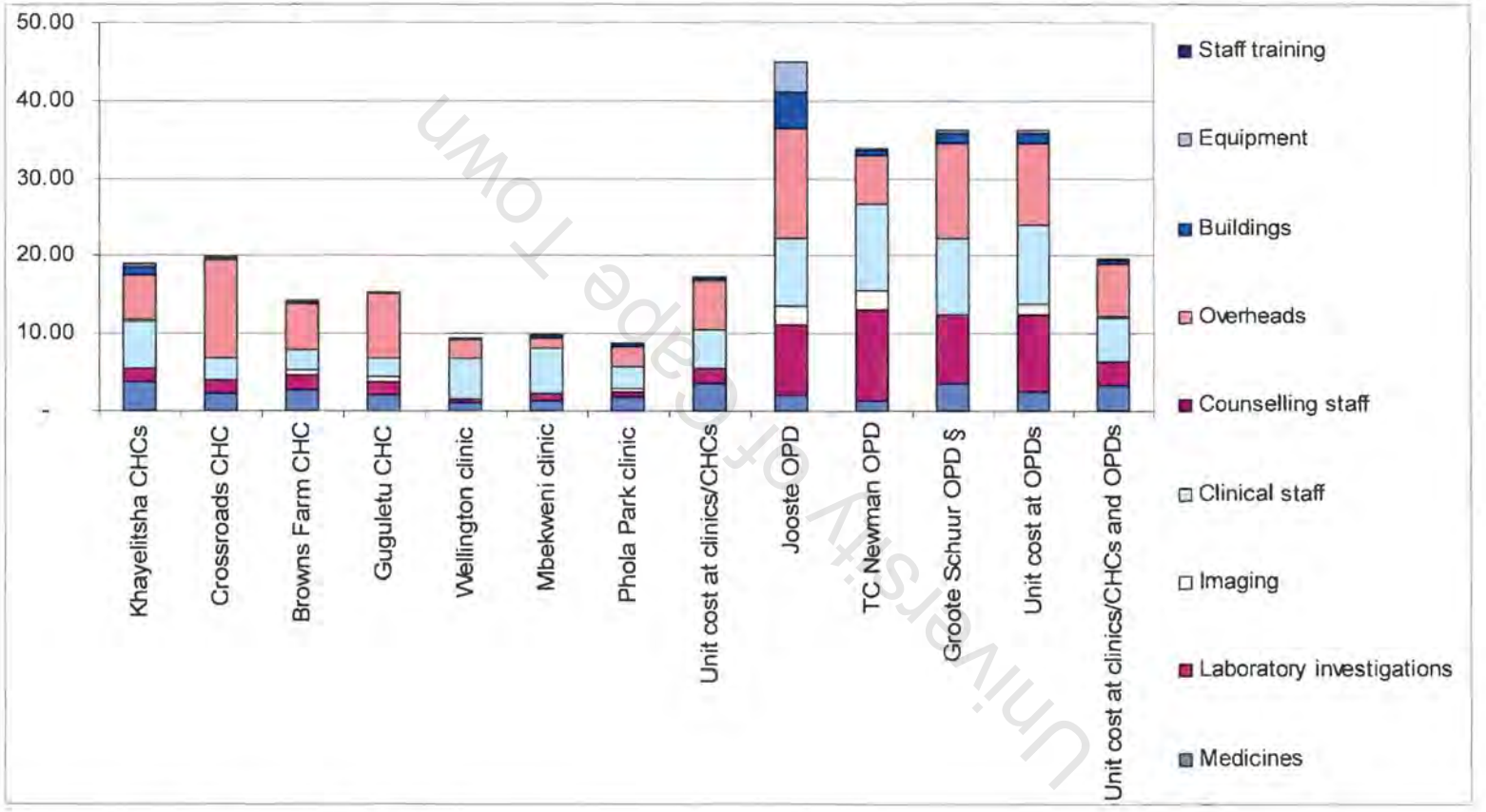
Once clinic/CHC and OPD average costs have been estimated, the third step is to calculate a final average weighted cost across all facility types, based on the proportion of visits that happen at clinics/CHCs versus OPDs across the country[4]. Data on the number of hospital outpatient department visits and clinic and CHC visits were extracted from the National Department of Health's IHPF model which was updated to 2002/03[5]. These data indicated that across the country 88 per cent of visits were to clinics and CHCs and 12 per cent to hospital OPDs. This proportion is similar to that found in an appraisal of 21 District Health Expenditure Review reports, which indicated that 16 per cent of visits took place at hospital OPDs (Cleary, Okorafor et al. 2005). It is assumed that any expansion in No-ART services maintains the current service configuration in South Africa. While this is reasonable in the short-term, in the longer run it is

[3] If the total sample of visits for calculating patient-specific costs was 100, and one clinic contributed 80 visits to the sample, the average patient-specific cost from this setting would be weighted at 80/100=0.8.
[4] This calculation includes all diseases and services. Data are not routinely collected for No-ART care.
[5] Attempts were unsuccessful to access a more recent version of the model.

likely that the country's stock of CHC infrastructure will be upgraded and increased to meet the burden of HIV-treatment (Department of Health 2007). A scenario will therefore be constructed where it is assumed that all No-ART visits occur at clinics and CHCs.

Figure 13 presents the breakdown of costs across the facilities that have been included in the final No-ART unit cost calculations, as well as the average weighted unit cost across clinics and CHCs, and across OPDs.

Figure 13: Average weighted unit cost per No-ART visit (US$; 2004 prices)



Source: § Govender, McIntyre et al. (2000)

For ART services, average weighted costs have been calculated using the same methods as for No-ART services. Figure 14 shows the final unit cost per ART visit.

Figure 14: Average weighted unit cost per ART visit (US$; 2004 prices)

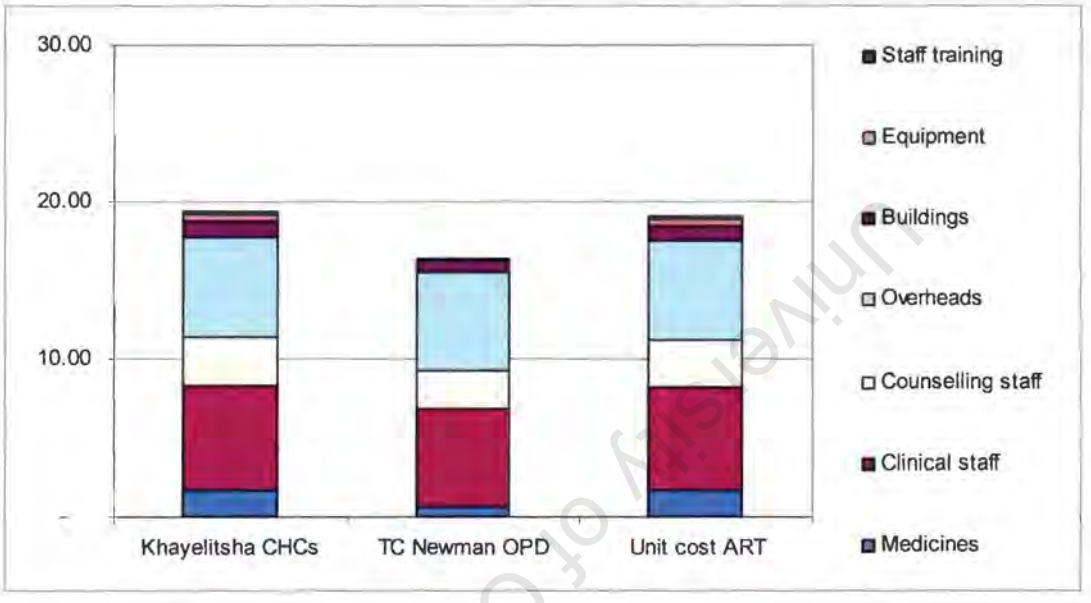


Table 10 compares the final unit cost for ART and No-ART visits. Clinical staff, counselling and capital costs are higher for ART visits while medicine costs are higher for No-ART. Bear in mind that ARV medicine and laboratory costs have been calculated separately for ART services.

Table 10: Comparison of unit cost per ART and No-ART visit (US$; 2004 prices)

	ART	%	No-ART	%
Medicines	1.64	8.6%	3.38	17.2%
Laboratory investigations		0.0%	2.89	14.7%
Imaging		0.0%	0.15	0.8%
Clinical staff	6.56	34.5%	5.70	29.0%
Counselling staff	2.93	15.4%	0.13	0.7%
Overheads	6.36	33.4%	6.79	34.6%
Buildings	0.95	5.0%	0.45	2.3%
Equipment	0.37	2.0%	0.11	0.6%
Staff training	0.21	1.1%	0.02	0.1%
Cost per visit	**19.02**	100.0%	**19.62**	100.0%

3 Unit cost of tuberculosis treatment

Tuberculosis (TB) is one of the leading causes of death for HIV-positive patients (Badri, Wilson et al. 2002). In South Africa, TB treatment is offered through a vertical service. If a patient is suspected of having TB, he/she is referred from general primary care services to TB services. At this point, TB diagnosis is confirmed and treatment is initiated. TB treatment consists of four or five doctor-based clinic visits for treatment initiation and monitoring, along with structured ongoing directly observed treatment (DOT) support at a clinic, by community treatment supporters or at the workplace. Treatment has to be sustained for 6 months for new patients and 7 months for retreatment patients. Standard protocols exist for laboratory investigation, imaging, and medicine regimens depending on whether the patient is a new or retreatment case (Department of Health 2000c).

The costing of TB treatment has been broken into three categories: clinic visits for treatment initiation and monitoring; DOTS; and patient-specific costs (TB medicines, laboratory investigations, imaging and procedures). Costs in these categories have been calculated separately for new and retreatment cases (see Table 11). The costs of clinic visits for treatment initiation and monitoring have been based on HIV-positive patients at Nyanga clinic (see second column in Table 11). The costs of clinic visits for initiation and monitoring have also been sourced from secondary data for patients who were not necessarily HIV-positive (Sinanovic, Floyd et al. 2000; Sinanovic, Floyd et al. 2003). The overall magnitude of clinic visit costs from secondary data was similar to the costs at Nyanga, especially if the additional medicines (mainly multivitamins and cotrimoxazole) offered to HIV-positive patients were excluded from the Nyanga visit cost. For this reason, despite the sample size at Nyanga being small, the average cost per clinic visit for initiation and monitoring was based on these costs and secondary data were not included. While DOTS costs were also calculated for Nyanga, a number of alternative TB supervision models are employed in the health system, with different associated costs. To capture these differences, weighted average costs for TB DOTS included primary and secondary data, and costs were weighted according to the total patient sample in each setting. Patient-specific costs, including TB medicine, laboratory investigation and imaging, were based on the National Protocol. Secondary data were not included in the calculation, but are presented in Table 11 for comparative purposes. The weighted average cost per patient treated was US$545 for new patients and US$777 for retreatment patients.

Table 11: Unit costs of TB treatment (US$; 2004 prices)

Venue	Nyanga; Clinic DOTS	Nyanga; Clinic DOTS *	Chapel St; Clinic DOTS *	Guguletu; Clinic DOTS *	Guguletu; Work-place DOTS *	Guguletu; Com-munity DOTS *	Average
Total patient samples	26	207	135	614	11	177	1170
Costs of TB treatment for new patients							
New patient samples	17	136	93	383	9	145	783
Costs of 4 clinic visits for initiation and monitoring							
Medicines	5.67						5.67
Clinical staff	28.70						28.70
Overheads	8.26						8.26
Buildings	1.23						1.23
Equipment	0.23						0.23
Total	**44.08**	**29.73**	**32.86**	**27.34**	**28.04**	**28.02**	**44.08**
Costs of ongoing DOTS visits							
Overheads	268.43						
Buildings	39.83						
Equipment	7.32						
Total	**315.57**	**508.50**	**591.47**	**455.45**	**84.94**	**148.14**	**427.45**
Patient-specific costs for new TB case							
TB Medicines	43.39	49.49	49.49	49.49	49.49	49.49	43.39
Laboratory investigations	11.13	7.91	7.91	7.91	7.91	7.91	11.13
Imaging & procedures	18.78	22.65	22.65	22.65	22.65	22.65	18.78
Total	**73.30**	**80.04**	**80.04**	**80.04**	**80.04**	**80.04**	**73.30**
Total cost per new patient treated	**432.94**	**618.27**	**704.37**	**562.83**	**193.02**	**256.21**	**544.82**
Costs of TB treatment for retreatment patients							
Retreatment patient samples	9	71	42	231	2	32	387
Costs of 5 clinic visits for initiation and monitoring							
Medicines	7.08						7.08
Clinical staff	35.87						35.87
Overheads	10.32						10.32
Buildings	1.53						1.53
Equipment	0.28						0.28
Total	**55.10**	**37.16**	**41.07**	**34.17**	**34.65**	**35.12**	**55.10**
Costs of ongoing DOTS visits							
Overheads	359.28						
Buildings	53.31						
Equipment	9.79						
Total	**422.38**	**680.61**	**791.65**	**609.44**	**203.94**	**270.31**	**583.78**
Patient-specific costs for retreatment TB case							
TB Medicines	84.39	104.69	104.69	104.69	104.69	104.69	84.39
Laboratory investigations	34.71	23.82	23.82	23.82	23.82	23.82	34.71
Imaging & procedures	18.78	22.65	22.65	22.65	22.65	22.65	18.78
Total	**137.89**	**151.15**	**151.15**	**151.15**	**151.15**	**151.15**	**137.89**
Total cost per retreatment patient treated	**615.36**	**868.92**	**983.88**	**794.76**	**389.73**	**456.58**	**776.77**

Table 12 shows the final average weighted cost (US$622) to be used for TB services. This is a weighted average of new and retreatment costs, where the weighting is based on the percentage of patients in each category. Overall, 33 per cent of patients were retreatment cases. This was similar to the percentage of HIV-positive retreatment TB patients in Nyanga (35 per cent). The costs of DOTS monitoring comprised 78.1 per cent of the cost of TB treatment.

Table 12: Comparison of costs of managing new and retreatment TB patients (US$; 2004 prices)

Type of TB treatment	New patients	Retreatment patients	Average	%
Patient samples	783	387	1170	
Costs - clinic visits				
Medicines	5.67	7.08	6.14	1.0%
Clinical staff	28.70	35.87	31.07	5.0%
Overheads	8.26	10.32	8.94	1.4%
Buildings	1.23	1.53	1.33	0.2%
Equipment	0.23	0.28	0.24	0.0%
Total	**44.08**	**55.10**	**47.72**	**7.7%**
Costs - DOTS visits				
Total	**427.45**	**583.78**	**479.16**	**77.1%**
Patient-specific costs for TB				
TB Medicines	43.39	84.39	56.95	9.2%
Laboratory investigations	11.13	34.71	18.93	3.0%
Imaging & procedures	18.78	18.78	18.78	3.0%
Total	**73.30**	**137.89**	**94.66**	**15.2%**
Total costs of TB treatment	**544.82**	**776.77**	**621.54**	**100%**

4 Unit cost per inpatient day

This section describes and summarizes a number of hospital cost analyses that will be used to inform the costs of inpatient care for ART and No-ART patients. The section includes a discussion of patient-specific costs, clinical personnel costs, overhead and capital costs and concludes with the calculation of average weighted unit costs per inpatient day.

Table 13 summarizes the patient-specific cost per inpatient day. The "sample of inpatient days" refers to the number of inpatient days that was used to calculate patient-specific costs. The "sample of admissions" refers to the number of patients. At the secondary level, one of the cost

analyses calculated inpatient costs separately for ART and No-ART patients. Because ART-specific costs are not available in other facilities, and because the cost was similar, these data are merged. In other words, it is assumed that the cost per inpatient day for HIV-positive patients is the same whether the patient is on or off ART. Costs at Jooste derived from Smith De Scherif, Schoeman et al. (2005) were far lower than from Haile (2000). Jooste had been open for one year at the time of Haile's study and had patient-specific costs that were similar to those at tertiary facilities. Since that time, the facility has experienced an increasing HIV-related burden in its medicine wards and has introduced step-down care to cope with this increased load. This is likely to be the explanation for the higher patient-specific costs from the Haile study.

A weighted average patient-specific cost for tertiary and secondary hospitals has been calculated based on the proportion of the sample of inpatients in facility x out of the total inpatient sample at each level of care. As before, this calculation method ensures that larger costing samples are given relatively higher weight in average cost calculations. The weighted average patient specific cost at tertiary and secondary facilities was US$23.71 and US$15.21 respectively.

Table 13: Patient-specific cost per inpatient day (US$; 2004 prices)

No-ART	Tygerberg	Groote Schuur §	Average Tertiary	Jooste ±	Jooste & Carnation ART †	Jooste & Carnation No-ART †	Average Secondary
Sample of inpatient days	243	777	1,020	367	1,003	654	2,024
Sample of admissions	61	100	161		70	60	130
Medicines	4.48	5.67	5.39	9.90	1.98	1.90	3.39
Laboratory investigations	6.40	8.93	8.32	12.97	6.15	6.27	7.42
Imaging	18.10	3.22	6.76	4.16	3.74	4.43	4.04
Consumables	-	4.24	3.23	-	-	-	-
Diagnostic and treatment procedures	-	-	-	0.63	0.40	0.13	0.35
IV fluids and blood	-	-	-	-	3.04	1.75	2.07
Total patient-specific costs	**28.97**	**22.06**	**23.71**	**27.66**	**12.26**	**12.74**	**15.21**

Sources:
§ Govender, McIntyre et al. (2000)
± Haile (2000)
† Smith De Scherif, Schoeman et al. (2005)

Table 14 contains details of overhead, building and equipment costs per inpatient day. In all facilities, overhead costs have been calculated from primary data[6], and capital costs are based on the NDoH model. The table also contains information on the overall inpatient PDE in each facility, which gives a sense of the size of each facility. Overheads ranged between US$125.46 at Tygerberg and US$45.95 at Groote Schuur. The latter is smaller because clinical staff costs for nurses and ward support staff have been calculated separately (see Table 15).

Weighted average overhead and capital costs for secondary and tertiary facilities have been based on the proportion of the facility's inpatient PDE out of the total inpatient PDE, implying that the larger facility received greater emphasis in average cost calculations. The average weighted overhead cost for tertiary facilities was US$85.57 and US$54.05 for secondary facilities. Equipment and building costs are also higher in tertiary than secondary hospitals.

[6] While patient-specific costs were derived from secondary data in certain facilities, all overheads have been calculated using more recent routine facility expenditure data.

Table 14: Overhead and capital costs per inpatient day (US$; 2004 prices)

	Tygerberg	Groote Schuur	Average tertiary	Jooste/Carnation
Overhead and capital costing sample (PDE inpatients)	516,473	424,942		93,720
Overhead cost per inpatient day	125.46	45.95	89.57	54.05
Building replacement cost	248,299,794.18	198,774,597.78		17,971,559.39
Equipment replacement cost	124,149,897.09	99,387,298.89		5,391,467.82
Building annuitized cost	22,055,800.79	17,656,611.22		1,596,365.13
Equipment annuitized cost	31,094,221.23	24,892,253.08		1,350,331.31
Building cost per inpatient day	42.70	41.55	42.18	17.03
Equipment cost per inpatient day	60.20	58.58	59.47	14.41

Staff costs were calculated from primary data at Tygerberg and at Jooste and Carnation, and were taken from secondary data for Groote Schuur. Separate staff costs were calculated for the Carnation ward because it has substantially lower clinical staff intensity than the Jooste medicine wards. Jooste and Carnation costs were combined to calculate an average weighted clinical staff cost using the proportion of inpatient days spent in each facility by patients included in the study by Smith De Cherif, Schoeman et al. (2005) as the weighting factor. As mentioned, Groote Schuur staff costs (Govender, McIntyre et al. 2000) covered a wider range of staff than in other facilities, and are therefore higher. At Tygerberg, it was not possible to calculate separate nursing costs for wards because nurses work on a flexible basis throughout the facility. A weighted average staff cost for tertiary facilities has been based on the inpatient PDE, once again giving greater weight to staff costs from the larger facility.

Table 15: Staff cost per inpatient day (US$; 2004 prices)

	Tygerberg	Groote Schuur §	Average tertiary	Jooste	Carnation	Average Jooste/ Carnation
Medical officer	35.93	28.41	32.53	14.28	1.45	11.71
Professional nurse		42.04	18.97	6.33	5.45	6.15
Ward support staff		1.71	0.77			
Total clinical staff costs	**35.93**	**72.16**	**52.28**	**20.60**	**6.90**	**17.86**

Source: §Govender, McIntyre et al. (2000)

Table 16 summarizes the final tertiary and secondary level inpatient unit costs that will be used in subsequent chapters. The average weighted cost per inpatient day in tertiary care was US$267.21 in comparison with US$118.56 for secondary level care. In both instances, the biggest cost drivers were overheads and capital. Patient-specific costs were 8.9 per cent and 12.8 per cent of the total in tertiary and secondary hospitals respectively.

Table 16: Average weighted unit cost per inpatient day (US$; 2004 prices)

	Average Tertiary	%	Average Secondary	%
Sample sizes				
Patient-specific costing sample	1,020		2,024	
Overhead and capital costing sample (PDE inpatients)	941,415		93,720	
Costs				
Patient-specific	23.71	8.9%	15.21	12.8%
Clinical staff	52.28	19.6%	17.86	15.1%
Overhead	89.57	33.5%	54.05	45.6%
Buildings	42.18	15.8%	17.03	14.4%
Equipment	59.47	22.3%	14.41	12.2%
Total cost per IPD	267.21	100.0%	118.56	100.0%

Sources:
§ Patient specific costs from Govender, McIntyre et al. (2000)
*Patient specific costs from Haile (2000) and Smith De Scherif, Schoeman et al. (2005)

5 ARV and laboratory investigation costs

South Africa recently awarded tenders for the supply of ARV drugs to the public sector, for the period 2005 to 2008. Although all other costs in this dissertation have been expressed in 2003/04 prices, the prices of ARVs reflect those in the tender as of October 2005. This was necessary to capture the substantial reductions in ARV prices that have occurred in recent years (Luchini, Cisse et al. 2003). The majority of the tender has been awarded to South African generic firm Aspen, with smaller quantities to Indian generic manufacturer Cipla Medpro. With the exception of 20 per cent of the lamivudine market share to GlaxoSmithKline, the originator manufacturers have only been successful when no generic alternative has been available (as is the case with efavirenz - manufactured by MSD - and lopinavir/ritonavir - manufactured by Abbott). In all cases, the tendering firm has been in possession of Medicines Control Council registration and appropriate Intellectual Property rights (patents in the case of originator manufacturers and voluntary licenses in the case of generics – there are currently no compulsory licenses for ARVs in South Africa). Table 17 contains a summary of ARV medicines and formulations, the manufacturer (and share of market where appropriate) and the cost per unit.

Table 17: ARV medicines, formulations, manufacturers and cost (US$; October 2005 prices)

Medicine name, formulation and quantity	Manufacturer	Share of tender	Cost per unit
First-line regimen			
Stavudine Capsules 30mg; 60'S	Aspen	70%	2.67
Stavudine Capsules 30mg; 60'S	Cipla Medpro	30%	2.94
Stavudine Capsules 30mg; 60s - average weighted cost			2.75
Stavudine Capsules 40mg; 60'S	Aspen	60%	2.97
Stavudine Capsules 40mg; 60'S	Cipla Medpro	40%	3.20
Stavudine Capsules 40mg; 60s - average weighted cost			3.06
Lamivudine Tablets 150mg; 60'S	Glaxo	20%	5.58
Lamivudine Tablets 150mg; 60'S	Aspen	80%	4.68
Lamivudine Tablets 150mg; 60'S - average weighted cost			4.86
Efavirenz Tablets 600mg; 30'S	MSD	100%	28.71
Nevirapine Tablets 200mg; 60'S	Aspen	100%	5.71
Second-line regimen			
Didanosine Tablets 25mg; 60'S	Aspen	100%	9.38
Didanosine Tablets 100mg; 60'S	Aspen	100%	10.89
Zidovudine Tablets 300mg; 60'S	Aspen	100%	10.73
Lopinavir/Ritonavir Capsules 133.3mg;33.3mg;180'S	Abbott	100%	47.60

ART is composed of a combination of three different ARV medicines, some of which need to be taken in different formulations by different patients. In the first-line regimen, patients can take either efavirenz or nevirapine in combination with stavudine and lamivudine. The share of patients on nevirapine/efavirenz has been based on primary data from the Khayelitsha programme (see Chapter 6 for details). Dosages for stavudine and didanosine differ depending on whether the patient weighs more or less than 60kg. If the former, the patient receives stavudine 30mg bd (twice daily) / didanosine 250mg od (once daily); if the latter, the patient receives stavudine 40mg bd / didanosine 400mg od. It was assumed that 50 per cent of patients would fall into each group. Table 18 presents the cost per three-month period for each ARV. This ARV unit cost information will be combined with data on the proportion of patients on different ARVs in the next chapter to calculate costs per Markov state.

Table 18: Cost per three-month period for each ARV (US$; October 2005 prices)

Medicine name and formulation	Daily dose	Unweighted cost per quarter
First-line		
Stavudine (30mg)	30mg bd	8.25
Stavudine (40mg)	40mg bd	9.18
Lamivudine (150mg)	150mg bd	14.58
Efavirenz (600mg)	600mg at night	86.14
Nevirapine (200mg)	200mg bd	17.12
Second-line		
Zidovudine (300mg)	300mg bd	32.19
Didanosine (100mg+25mg)	2*100mg+2*25mg od	60.80
Didanosine (100mg)	4*100mg od	65.33
Lopinavir/ritonavir (133.3/33.3mg)	3*133.3/33.3 mg od	142.81

bd = twice daily
od = once daily

Table 19 shows the schedule of laboratory investigations for patients on ART, as derived from the national guidelines (Department of Health 2004). Viral load testing occurs at base case (i.e. treatment initiation) and then 6-monthly thereafter. CD4 count testing occurs when a patient is staged for ART and then 6-monthly thereafter. Patients on nevirapine require ALT testing at weeks 2, 4, 8 and 6-monthly thereafter. Second-line patients require a full blood count and white cell diff at base case, weeks 4, 8, 12 and 6-monthly thereafter as well as less frequent fasting cholesterol and glucose tests. Testing is intensive during the first 6 months of treatment for both first and second-line regimens, and drops off to six-monthly for most tests for the remaining time on each regimen.

Table 19: Laboratory investigation schedule

	Staging	Base-line	Week 2	Week 4	Week 8	Week 12	Week 24	6 Monthly	12 Monthly
Regimen 1									
Viral load		X						X	
CD4 count	X							X	
Patients on nevirapine									
ALT			X	X	X			X	
Regimen 2									
Viral load								X	
CD4 count	X							X	
Full blood count and white cell diff		X		X	X	X		X	
Fasting cholesterol and triglyceride		X					X		X
Fasting glucose									X

Staging = initial testing for all patients when being referred for ART
Base case = testing for ARV eligible patients, at initiation of ART
Source: Department of Health (2004)

Table 20 shows the unit cost per laboratory test. Costs are highest for viral loads.

Table 20: Unit cost per laboratory test (US$; 2004 prices)

Test	Cost per test
Viral load	39.68
CD4 count	7.94
ALT	3.36
Full blood count and white cell diff	6.35
Fasting cholesterol and triglyceride	2.65
Fasting glucose	2.18

Table 21 combines data on quantities of laboratory tests and unit costs to estimate laboratory costs per patient-quarter on ART. For the first-line regimen, Markov tunnel states allow separate costs to be calculated in months 0-3, 3-6 and 6-12 therefore separate laboratory costs have been calculated in each period. An average quarterly cost is calculated for periods beyond the first year. Because of the Markovian assumption (no memory of previous cycles) it is not possible to keep track of the quarters for the second-line regimen, and a weighted average laboratory test cost per quarter has been calculated instead.

Costs are highest during the early period on each regimen, but fall to US$24.46 and US$28.19 per quarter on first and second-lines respectively after the first year.

Table 21: Laboratory costs per patient-quarter

	0-3 months	3-6 months	6-12 months	Quarterly
First-line				
Viral load	39.68		19.84	19.84
CD4 count	7.94		3.97	3.97
ALT	3.93		-	0.65
Total per patient-quarter	**51.55**		**23.81**	**24.46**
Second-line				
Viral load				19.84
CD4 count				3.97
Full blood count and white cell diff				3.17
Fasting cholesterol and triglyceride				0.66
Fasting glucose				0.54
Total per patient-quarter				**28.19**

6 Summary

These unit costs, which have been compiled from primary and secondary data to aid in enhancing generalizability, will be a component of the models used to calculate patient-level and population-level costs and consequences in alternative treatment strategies in the next 3 chapters

Chapter 6: Patient-level results

1 Introduction

This chapter presents patient-level lifetime costs and individual level health gains associated with alternative mutually exclusive HIV-treatment interventions. The chapter begins by describing the process of calculating and attaching costs and outcomes to Markov states and presents the transition probabilities that determine movements between these states. The next section describes the results of an extensive validation of the models which includes a detailed review of similar analyses in the literature. The chapter concludes by presenting the patient-level results associated with a number of interventions. The implications of defining need as illness or as capacity to benefit are also assessed.

2 Attaching rewards to Markov states

Each Markov state is associated with costs and outcomes, which together are known as rewards. Because the chosen cycle length of the model is three months, these rewards need to reflect either three months of costs or three months of outcomes.

2.1. Cost per Markov state

The cost per Markov state is a product of the utilisation of health services and the unit costs of that utilisation. Given that unit costs have been presented in the previous chapter, this section focuses on the calculation of utilisation and the calculation of the cost per Markov state. HIV clinic utilisation was calculated from 1,729 patients with 1,146 No-ART patient years and 2,229 ART patient years, over a median No-ART and ART follow-up of 0.63 years (IQR 0.33-1.32, max 4.35) and 1.03 years (IQR 0.68 – 1.70, max 4.08) respectively. Utilisation of tuberculosis and inpatient care was based on a sub-sample of 670 patients, with 501 patient-years for No-ART and 693 patient-years for ART patients. An additional sample of 83 patients was used to calculate inpatient utilisation prior to death for No-ART patients. The same procedure was followed for the 81 patients on ART who died of HIV-related causes. During the follow-up

period, No-ART patients had 10,892 visits, 1,342 inpatient days and 159 treated episodes of tuberculosis. ART patients had 39,450 visits, 840 days in hospital and 86 TB episodes.

When specified by Markov state, results indicated that patients commencing ART with CD4<50 cells/μl had 10.5 (95% CI 10.2-10.7) clinic visits, 1.1 (1.0-1.3) days in hospital and 0.08 (0.05-0.12) tuberculosis cases per patient in the first 3 months. Note that these clinic visits include patient work-up prior to ART commencement. During the same period, patients commencing ART with CD4 50-199 cells/μl had 9.7 (9.5-9.9) clinic visits, 0.6 (0.5-0.7) days in hospital and 0.06 (0.04-0.1) cases of tuberculosis. All health care utilisation dropped markedly over time. During the third year on treatment 2.6 (2.5-2.7) clinic visits, 0.1 (0.08-0.13) inpatient days and 0.02 (0.01-0.03) tuberculosis cases occurred per patient period. Patients not accessing ART with CD4<50 cells/μl utilised 3.4 (3.2-3.5) clinic visits, 0.7 (0.6-0.8) inpatient days, and had 0.1 (0.07-0.13) tuberculosis cases per quarter. All patients spent between 4 and 6 days in hospital prior to death. So as not to underestimate costs for patients lost to follow-up, it was assumed that these patients also incurred the inpatient utilisation associated with dying. Table 22 shows utilisation of health services in different Markov states per three-month period.

Table 22: Health service utilisation in Markov states per three-month period

Markov states:	Mean visits (95%CI)	Mean inpatient days (95%CI)	Mean TB cases (95%CI)
ART			
CD4<50 cells/μl months 0-3	10.47 (10.21-10.74)	1.09 (0.96-1.25)	0.08 (0.05-0.12)
CD4<50 cells/μl months 3-6	3.67 (3.52-3.84)	0.77 (0.66-0.90)	0.04 (0.02-0.08)
CD4 50-199 cells/μl months 0-3	9.69 (9.51-9.88)	0.60 (0.52-0.7)	0.06 (0.04-0.1)
CD4 50-199 cells/μl months 3-6	3.46 (3.34-3.57)	0.12 (0.09-0.17)	0.02 (0.01-0.04)
All patients months 6-12	3.62 (3.54-3.70)	0.21 (0.18-0.24)	0.02 (0.01-0.03)
All patients months 12-24	2.72 (2.65-2.79)	0.10 (0.08-0.13)	0.02 (0.01-0.03)
All patients months 24-36	2.58 (2.48-2.69)	0.10 (0.08-0.13)	0.02 (0.01-0.03)
All patients month 36 onwards	2.76 (2.56-2.97)	0.10 (0.08-0.13)	0.02 (0.01-0.03)
Utilisation for patients who die / default		3.95 (1.72-6.19)	
No-ART			
CD4 <50 cells/μl	3.35 (3.24-3.49)	0.7 (0.64-0.75)	0.1 (0.07-0.13)
CD4 50-199 cells/μl	2.58 (2.53-2.64)	0.3 (0.28-0.32)	0.07 (0.06-0.09)
Utilisation for patients who die / default		5.75 (3.51-7.98)	

The generalizability of these utilisation estimates has been assessed by comparing these results against those of a South African cost-effectiveness analysis based on a cohort of ART patients enrolled in a clinical trial and a natural history No-ART group (Badri, Cleary et al. 2006).

Although data are not presented in a directly comparable manner, some general conclusions can be made. For ART patients, quarterly utilisation of inpatient care varied between a high of 0.45 days for patients with AIDS and CD4<200 cells/μl and a low of 0.04 for patients with CD4>350 cells/μl and No-AIDS. Inpatient utilisation in Khayelitsha was higher, but this could relate to the lower CD4 levels of patients. Quarterly outpatient utilisation in Badri, Cleary et al. (2006) varied between 1.4 and 1.9 visits. Although utilisation of clinic visits in Khayelitsha may appear high compared to these data, this may reflect the model of care, where clinical visits were to a combined doctor-nurse team, with the implication that patients would not see a doctor at every visit. This differs from settings where visits are counted only as those in which a doctor is consulted.

In the No-ART natural history cohort (Badri, Cleary et al. 2006), quarterly inpatient utilisation was 4.43 days and outpatient utilisation was 1.93 visits for patients with AIDS and CD4<200 cells/μl. Inpatient utilisation is similar in Khayelitsha, especially if one takes utilisation at death into account (which would be included in the 4.43 measure from the natural history cohort) while outpatient visits were again considerably higher in Khayelitsha.

To enhance the generalizability of visit utilisation, scenario analysis will consider the following adjustments:

- Basing ART visits on national ART guidelines
- Basing No-ART visits on data from Badri, Cleary et al. (2006) under the assumption that patients with CD4<50 cells/μl use the same visits as those with CD4<200 cells/μl and AIDS while patients with CD4 50-199 cells/μl use the same visits as those with CD4<200 cells/μl who are in other WHO stages. While this is a crude assumption, there are no other data to which this comparison can currently be made.

These utilisation estimates are presented in Table 23.

Table 23: Alternative estimates of visit utilisation per three-month period

Markov states:	Mean visits (95%CI)
ART	
CD4<50 cells/µl months 0-3	6.00
CD4<50 cells/µl months 3-6	3.00
CD4 50-199 cells/µl months 0-3	6.00
CD4 50-199 cells/µl months 3-6	3.00
All patients months 6-12	3.00
All patients months 12-24	3.00
All patients months 24-36	3.00
All patients month 36 onwards	3.00
No-ART	
CD4 <50 cells/µl	1.93
CD4 50-199 cells/µl	1.40

Table 24 illustrates the proportion of patients on different antiretroviral drugs. In the first-line regimen, all patients were assumed to receive stavudine and lamivudine in line with "National Antiretroviral Treatment Guidelines" (National Antiretroviral Treatment Guidelines 2004). Data indicated that 47 per cent (44 per cent-49 per cent) of patients received efavirenz and the remainder received nevirapine. All patients are assumed to receive zidovudine, didanosine, and lopinavir/ritonavir as the second-line.

Table 24: Proportion of patients on each ARV

Medicine name / formulation	proportion patients (95% CI)
First-line	
Stavudine (30mg)	0.50
Stavudine (40mg)	0.50
Lamivudine (150mg)	1.00
Efavirenz (600mg)	0.47 (0.44-0.49)
Nevirapine (200mg)	0.53 (0.51-0.56)
Second-line	
Zidovudine (300mg)	1.00
Didanosine (250mg)	0.50
Didanosine (400mg)	0.50
Lopinavir/ritonavir (133.3/33.3mg)	1.00

Utilisation of health care services and ARVs is multiplied by the unit cost of each to calculate the cost per Markov state. The following is a generic formula for calculating the cost per three-month period in Markov state n:

$$Stage\_cost_n = [ARV^{1...7}{}_n * c\_ARV^{1...7}{}_n + ARV^{1...7}{}_n * c\_ARV^{1...7}{}_n + ARV^{1...7}{}_n * c\_ARV^{1...7}{}_n + c\_labs_n + visit_n * c\_visit_n + ip_n * c\_ip + tb_n * c\_tb] / (1 + disc\_rate)$$

Where:

$Stage\_cost_n$ is the total cost per three-month period in Markov state n

$ARV^{1...7}{}_n$ is the proportion of patients on each individual ARV[7] in Markov state n

$c\_ARV^{1...7}{}_n$ is the three-month cost of $ARV^{1...7}$ in Markov state n

c_labs_n is the three-month cost of laboratory investigations in Markov state n

$visit_n$ is the distribution (mean and 95 per cent confidence interval) of visits per three-month period in Markov state n

c_visit_n is the unit cost per visit in Markov state n

ip_n is the distribution (mean and 95 per cent confidence interval) of inpatient days in Markov state n

c_ip is the unit cost per inpatient day

tb_n is the distribution (mean and 95 per cent confidence interval) of TB incidence in Markov state n

c_tb is the unit cost per TB treatment completed

$disc\_rate$ is the discount rate per three-month period

Table 25 summarizes utilization information and presents the cost per Markov state. Note that the cost per Markov state is calculated by multiplying unit costs (as presented in the previous chapter) by the utilization of different service categories. ART costs were highest for patients with CD4<50 cells/μl during the first three months on treatment, at US$548 excluding costs for patients who died. Costs remained steady at around US$180 per quarter from 12 months onwards while patients remained on first-line, but increased to over US$340 per quarter when second-line treatment was initiated. No-ART costs in the CD4<50 cells/μl category were US$239 per patient-quarter excluding patients who died. Obviously, No-ART patients do not incur ARV or ARV

---

[7] ARVs include combinations of stavudine, lamivudine, nevirapine, efavirenz, didanosine, zidovudine or lopinavir/ritonavir as appropriate to each Markov state. For No-ART states, $ARV^{1...7}{}_n = 0$.

related laboratory costs – these categories are therefore not applicable. A mean cost of over US$1000 was incurred at hospitals during the period preceding death.

Table 25: Summary of service utilisation and the cost per Markov state (US$)

Health state	Clinic visits		Inpatient days		Tuberculosis treatment		ARV Cost (95% CI)	Safety and monitoring laboratory costs	Cost per Markov state (95%CI)	Additional cost for dying patients	
	Mean visits (95%CI)	Cost (95%CI)	Mean IP days (95%CI)	Cost (95%CI)	Mean cases (95%CI)	Cost (95%CI)				Mean IP days (95%CI)	Cost (95%CI)
ART CD4<50 cells/µl months 0-3		199 (194-204)		177 (155-201)		48 (30-76)	72.7 (70.2-75.1)	52	**548 (502-609)**		704 (306-1102)
ART CD4<50 cells/µl months 3-6		70 (67-73)		125 (107-146)		25 (13-49)	72.7 (70.2-75.1)	0	**293 (257-343)**		704 (306-1102)
ART CD4 50-199cells/µl months 0-3		184 (181-188)		98 (84-114)		38 (24-60)	72.7 (70.2-75.1)	52	**444 (410-488)**		704 (306-1102)
ART CD4 50-199 cells/µl months 3-6		66 (64-68)		20 (14-27)		11 (4-26)	72.7 (70.2-75.1)	0	**169 (152-196)**		704 (306-1102)
First-line ART months 6-12		69 (67-70)		33 (29-39)		13 (8-21)	72.7 (70.2-75.1)	24	**212 (198-229)**		704 (306-1102)
Second-line ART months 6-12							238	28	**381 (370-397)**		
First-line ART months 12-24		52 (50-53)		17 (14-20)		12 (8-20)	72.7 (70.2-75.1)	24	**178 (167-193)**		704 (306-1102)
Second-line ART months 12-24							238	28	**347 (338-360)**		
First-line ART months 24-36		49 (47-51)		17 (14-20)		12 (8-20)	72.7 (70.2-75.1)	24	**175 (164-191)**		704 (306-1102)
Second-line ART months 24-36							238	28	**344 (335-358)**		
First-line ART beyond 36 months		52 (49-56)		17 (14-20)		12 (8-20)	72.7 (70.2-75.1)	24	**179 (165-196)**		704 (306-1102)
Second-line ART beyond 36 months							238	28	**348 (337-363)**		
No-ART CD4 <50 cells/µl		66 (63-68)		113 (104-122)		60 (46-79)	N/A	N/A	**239 (213-269)**		1023 (626-1422)
No-ART CD4 50-199 cells/µl		51 (49-52)		49 (45-52)		45 (37-55)	N/A	N/A	**145 (132-159)**		1023 (626-1422)

IP: Inpatient care

2.2. Outcomes in each Markov state

In order for the model to calculate life-years or QALYs, outcome rewards are attached to each Markov state (other than the absorbing state - death - where no outcomes are accrued). If outcomes are life years, the reward is the same as the Markov cycle length, which in this case is three months. In the case of QALYs, each three-month period is multiplied by the HRQoL value that is appropriate to each Markov state. HRQoL has been measured in Khayelitsha on a subset of patients over a 12 months period using the EQ-5D instrument (Jelsma, MacLean et al. 2005) and has been valued using a United Kingdom general population TTO value set (Dolan, Gudex et al. 1995). Results are presented in Table 26.

Table 26: HRQoL values

Health state	Value
ART 0-3 months	0.71
ART 3-6 months	0.81
ART 6-12 months	0.82
ART >12 months	0.85
No-ART	0.71

3 Transition probabilities

Table 27 contains transition probabilities and data sources for the ART and No-ART models. As explained in the methods, ART transition probabilities were calculated from 1,729 patients followed over a maximum duration of 48 months.

Sixty-three per cent of patients initiated ART with CD4 50-199 cells/μl and the remainder with CD4 counts <50 cells/μl. Deaths were concentrated in the first year. At six months on ART, 83.9 per cent of the 643 patients initiating with a CD4 count <50 cells/μl were alive in comparison with 93.9 per cent of those initiating with CD4 50-199 cells/μl. No patients were changed to second-line during the first 6 months, but by 48 months, 16 per cent of those surviving had switched. The product limit estimate of survival was 86.9 per cent, 83.4 per cent and 76.2 per cent at 12, 24 and 48 months respectively. Extrapolated results indicate 71 per cent surviving at 60

months. These outcomes are slightly worse than results from a 7-year Senegalese cohort study (Etard, Ndiaye et al. 2006) which indicated 88.3 per cent, 82.6 per cent and 75.4 per cent surviving at 12, 24 and 60 months respectively but, given the higher base case CD4 counts of the Senegalese patients, results are comparable.

Local natural history data indicated that fifty per cent of No-ART patients with CD4<200 cells/μl were alive at 24 months (Badri, Bekker et al. 2004).

Table 27: Transition probabilities (per three-month period) and data sources

Input parameters		Data sources	Transition probability (range)
CD4 category at baseline for ART and No-ART			
	CD4<50 cells/ml		0.372 (0.349-0.395)
	CD4 50-199 cells/ml		0.628 (0.605-0.651)
Probabilities of transitioning between alive Markov states			
First-line regimen to second-line regimen			
	0-6 months	N/A - no patients switched	0.000
	6-12 months	0.48% switched by 12 months	0.002 (0.001-0.006)
	12-24 months	4.66% switched by 24 months	0.011 (0.007-0.015)
	24-36 months	11.73% switched by 36 months	0.019 (0.014-0.026)
	36-48 months	15.9% switched by 48 months	0.012 (0.005-0.024)
	>48 months	Average over 0-48 months	0.011 (0.007-0.017)
No-ART, CD4 50-199 cells/ml to CD4<50 cells/ml			
	all quarters	Calculated to ensure 50% surviving at 24 months	0.040 (0.026-0.043)
Probabilities of transitioning to dead Markov states			
ART CD4<50 cells/ml			
	0-3 months	86.9% surviving at 3 months	0.131 (0.107-0.159)
	3-6 months	83.9% surviving at 6 months	0.034 (0.031-0.038)
ART CD4 50-199 cells/ml			
	0-3 months	95.9% surviving at 3 months	0.041 (0.030-0.054)
	3-6 months	93.9% suriviving at 6 months	0.021 (0.019-0.024)
All patients on ART, irrespective of regimen			
	6-12 months	86.9% suriviving at 12 months	0.018 (0.017-0.020)
	12-24 months	83.4% suriviving at 24 months	0.010 (0.008-0.012)
	24-36 months	79.5% surviving at 36 months	0.012 (0.009-0.016)
	36-48 months	76.2% suriviving at 48 months	0.010 (0.005-0.017)
	>48 months	Average over 0-48 months	0.017 (0.013-0.021)
No ART CD4 count <50 cells/ml			
	all quarters	20% surviving at 24 months[1]	0.182 (0.147-0.227)
No ART CD4 50-199 cells/ml			
	all quarters	50% surviving at 24 months with CD4<200 cells/ml[2] divided by hazard ratio[3]	0.039 (0.034-0.043)

Sources:
[1] No-ART survival with CD4<50 cells/μl from Post, Wood et al. (2002)
[2] No-ART survival with CD4<200 cells/μl from Badri, Bekker et al. (1996)
[3] Hazard ratio for death in CD4<50 versus 50-199 cells/μl from Coetzee, Hildebrand et al. (2004)

While ART data were similar to those reported in the Senegalese cohort, generalizability has also been assessed by comparing 12-month clinical outcomes against ART-LINC data. These data are from a collaboration of ART cohorts in low income countries. A recent publication includes data from 4,810 patients receiving ART in Africa (including 278 Khayelitsha patients), South America and Asia and provides information about base case CD4 stratum, loss to follow up and deaths after one year of treatment (ART-LINC and ART-CC 2006). Because a number of these patients were enrolled in cohorts without active follow-up, this comparison focuses on 2,725 ART-LINC patients who were actively followed. Results are presented in Table 28.

Table 28: Comparison of outcomes at one-year between ART-LINC and Khayelitsha

	ARTLinc	Khayelitsha
Number of patients	2725	1729
Proportion loss to follow-up	12%	coded as deaths
Proportion deaths	6%	13%
Combined deaths and loss to follow-up	18%	13%

Thirty-one per cent of patients in ART-LINC initiated ART with CD4<50 cells/μl in comparison with 37 per cent in Khayelitsha. Loss to follow-up was higher in ART-LINC than in Khayelitsha, but the rate of death was far higher in Khayelitsha than in ART-LINC which suggests that many of the patients that were lost to follow-up in ART-LINC had died. The final row in the table therefore calculates combined deaths and loss to follow-up. When these data are compared, one could conclude that outcomes at one-year in Khayelitsha might be more favourable than in routine settings despite the higher proportion of patients initiating ART in lower CD4 strata. To increase the generalizability of results, death transition probabilities will be increased to match ART-LINC results in scenario analysis (see section 5.2). This would represent a conservative overestimate of mortality because patients who were lost to follow-up would not necessarily have died within this one-year period.

The generalizability of No-ART outcomes has been assessed by comparing results reported in Table 27 against a literature review of natural history data from developed and developing countries (Schneider, Zwahlen et al. 2004). This indicated that patients with CD4<200 cells/μl had a median survival of 11 months (ranging between 7-38 months), whereas calculations from secondary data for this thesis (Badri, Bekker et al. 2004) indicate a median survival of 15 months. Similar to ART, No-ART outcomes are therefore slightly higher than in other resource-poor settings. This will also be addressed in scenario analysis.

While Table 27 presents the probabilities of dying from HIV-related causes, one also has to include the probability of dying from other causes, adjusted for socioeconomic status, age and gender of HIV positive people. These data were taken from a sample of 749 patients accessing ART in a number of clinics in the Western Cape (Pienaar, Myer et al. 2006). Seventy-seven per cent of these patients were women. Men had a median age of 37 as compared to 32 for women and an average of 33 for both genders. Socio-economic status was estimated by constructing an asset index using the 1998 Demographic and Health Survey adult dataset. Six socio-economic groups were calculated based on this asset index, and by mapping the same asset index onto patients in the ART sample, it was possible to estimate the socio-economic strata of these patients. Group 1 is the lowest socio-economic stratum. Results indicate that 33 per cent of ART users fell into the poorest three groups and 67 per cent were in groups 4, 5 or 6. The majority of the sample (32.6 per cent) fell into group 4. See Table 29. There were no major differences in strata between men and women.

Table 29: Socioeconomic groupings of men and women in the ART sample

	Women	%	Men	%	All	%
Group 1	6	1.0%		0.0%	6	0.8%
Group 2	61	10.6%	26	14.9%	87	11.6%
Group 3	117	20.3%	37	21.3%	154	20.6%
Group 4	200	34.8%	44	25.3%	244	32.6%
Group 5	157	27.3%	54	31.0%	211	28.2%
Group 6	34	5.9%	13	7.5%	47	6.3%
Total	575	100.0%	174	100.0%	749	100.0%

Armed with age, gender and socio-economic status, it is finally possible to estimate the probability of dying from causes other than HIV per three-month period as a function of the age of the cohort. This is a time dependent probability which increases from the age of entry into care (average of 33 years) until the maximum age of 100 is reached when it is assumed that the probability of dying is 1.

4 Model validation

This section covers validation of the model under the categories of technical, predictive, face and modelling process validity. In terms of technical and predictive validity, the overall objective is to ascertain whether the model is reproducing primary data accurately whereas face validity assesses whether the model produces the output that one would expect. Non-HIV-related mortality probabilities have therefore been excluded from these sections. Modelling process validity ascertains whether the results in this thesis are comparable to other published cost-effectiveness studies. This section therefore includes non HIV-related mortality because this should be included in the results in the literature (Sonnenberg and Beck 1993).

4.1. Technical validity

Technical validity involves identifying and correcting for modelling bugs such as unexpected model behaviour, redundant variables, programming errors and typing errors (Sendi, Craig et al. 1999). A useful method of debugging a model is to vary one or more parameter over its entire range and to examine any anomalies in model outputs. This process was done using cohort simulation on the following transition probabilities:

- Probabilities of dying (ART[8] and No-ART model)
- Probabilities of failing first-line (ART model)

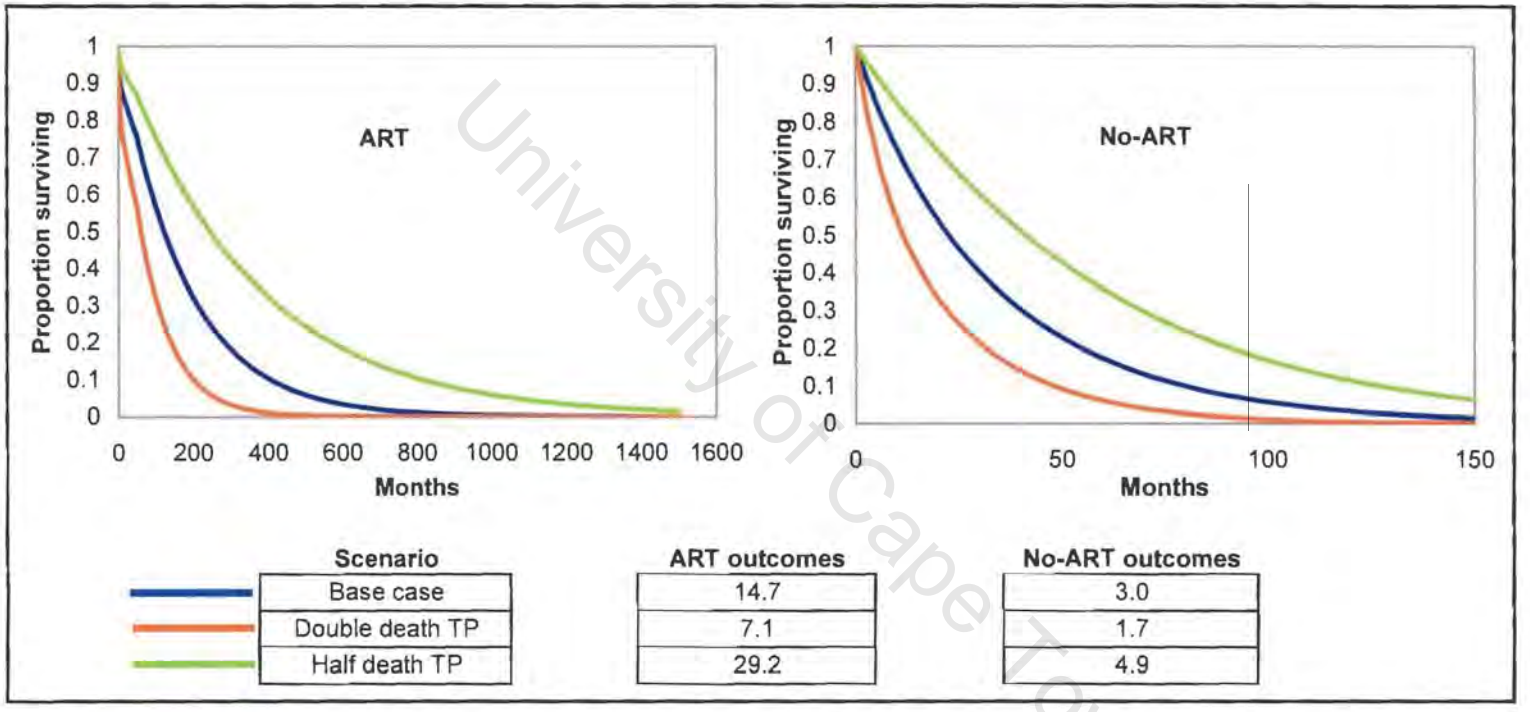
In order to validate death transition probabilities (TPs), the TPs for death were doubled and halved, and comparison was made to base case. Figure 15 shows total outcomes in the ART and No-ART models when these adjustments are made. Under base case assumptions, total ART outcomes are 14.7; when the death TPs are doubled, total outcomes are 7.1; and when death TPs are halved, total outcomes are 29.2. Adjustments to the death TP will never lead to exactly the same adjustment in outcomes because outcomes are also a function of the number of patients remaining in any given state facing a particular probability of dying. However, the validation shows that the model is performing as expected. Similarly, No-ART outcomes are 3.0 under base case TPs; 1.7 when the TPs are doubled; and 4.9 when the TPs are halved.

[8] Given that the first and second-line and first-line models are structured similarly, this validation has only been performed on the first and second-line model.

Technical validity has also been established for second-line TPs. While transitions to second-line will have no impact on total outcomes[9], there will be an impact on total costs because second-line drugs are more expensive. Therefore, the key purpose of the second-line TP is to ensure an appropriate split between first and second-line ARV regimens in order to capture cost differences. As before, the second-line TP was halved and doubled and results were compared to base case. For the purposes of technical validity, one would expect that doubling the TP would increase the percentage of total outcomes that are gained on second-line whereas halving it would do the opposite. While the second-line TPs function is to make cost adjustments, one can judge whether these are producing reasonable results by assessing the percentage of outcomes (e.g. life years) that are gained in first-line and second-line Markov states. Results are shown in Figure 16. Under the base case TP, 38 per cent of total outcomes are gained on second-line; when the second-line TP is doubled, 55 per cent of total outcomes are gained on second-line; and when the TP is halved, 23 per cent of outcomes are gained on second-line.

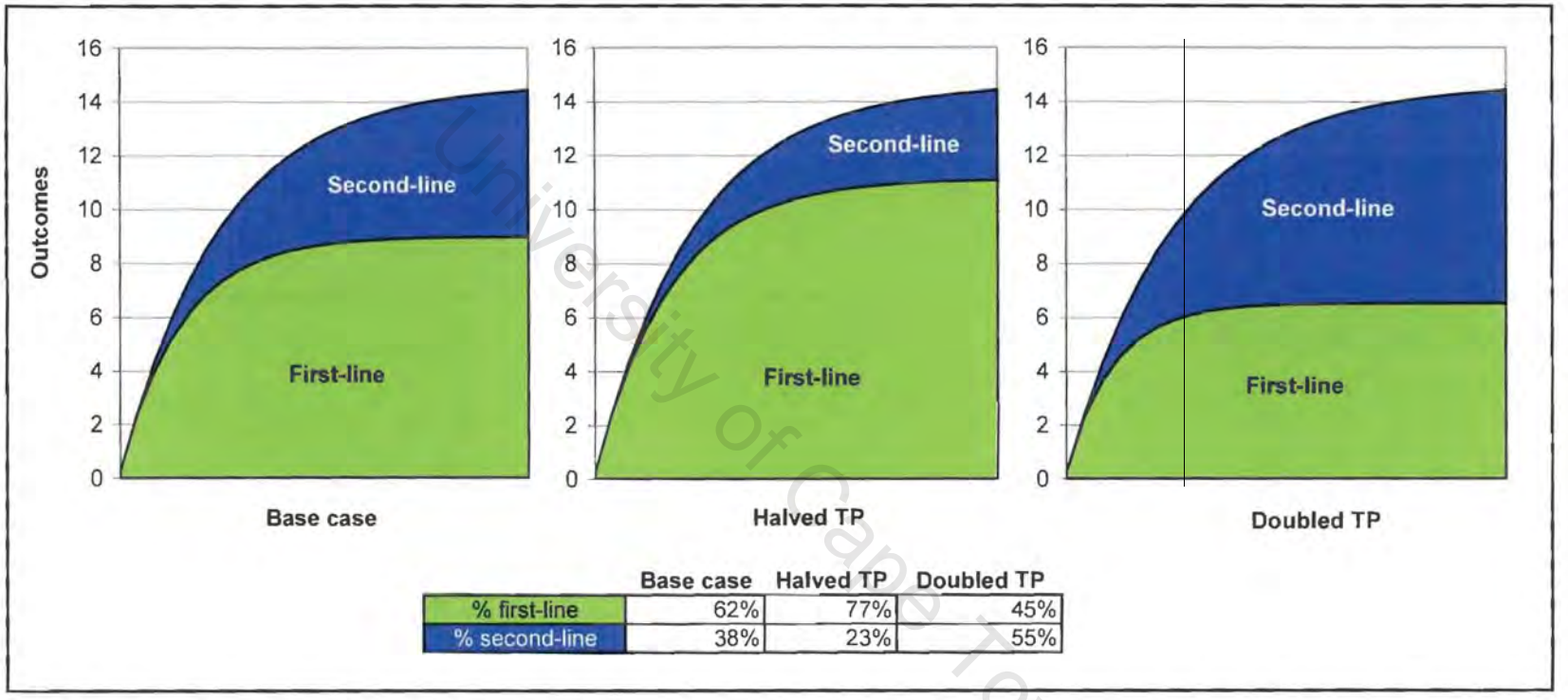
[9] A simplifying assumption of the model is that the probability of dying is a function of duration on treatment and does not vary in first or second-line specifically, although the overall chance of being on second-line does increase as time on treatment increases.

Figure 15: Modelled survival curves and outcomes for ART and No-ART under base case, halved and doubled death transition probabilities



Scenario	ART outcomes	No-ART outcomes
Base case	14.7	3.0
Double death TP	7.1	1.7
Half death TP	29.2	4.9

Figure 16: Cumulative time on first and second-line under base case, halved and doubled second-line transition probabilities



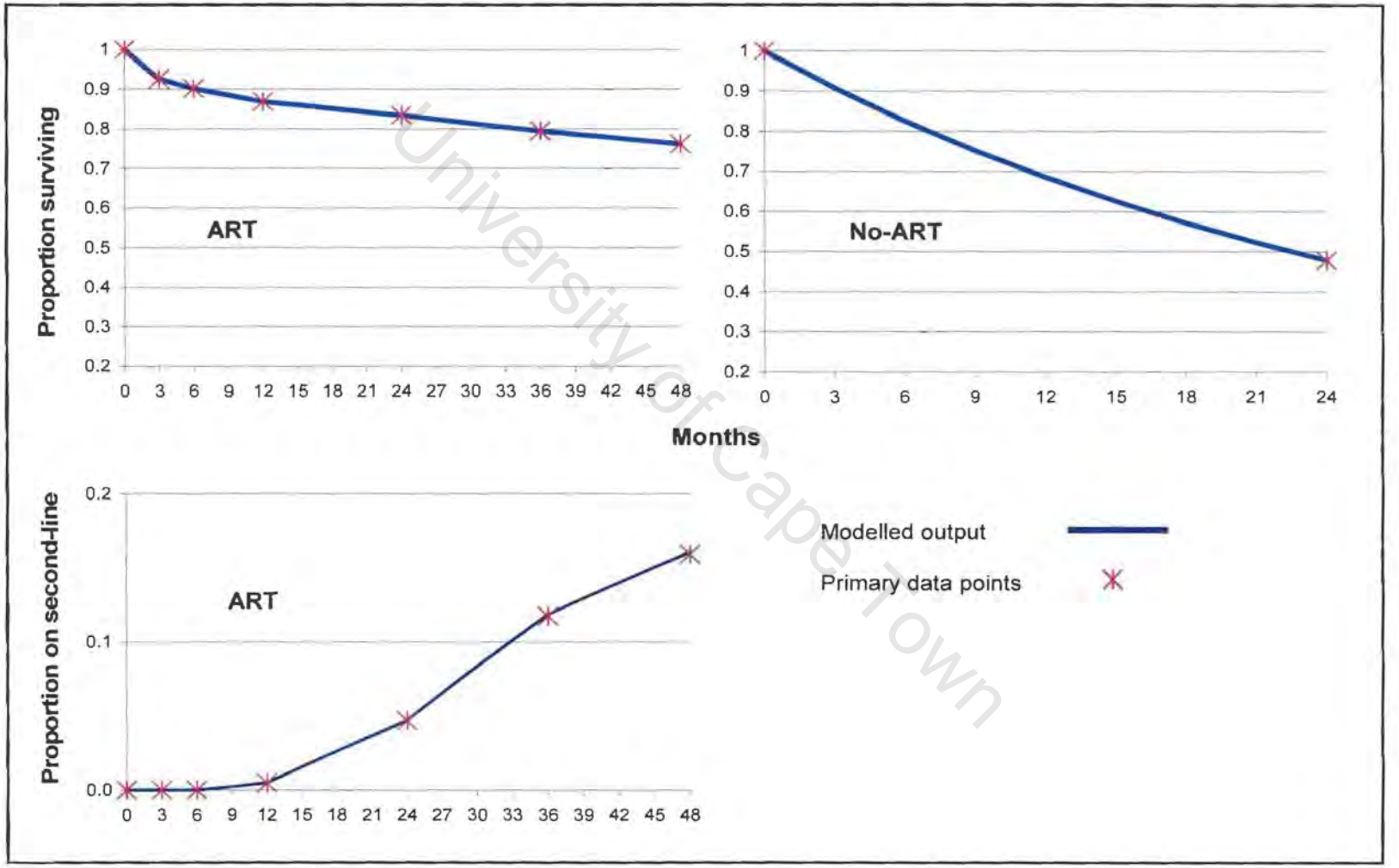
	Base case	Halved TP	Doubled TP
% first-line	62%	77%	45%
% second-line	38%	23%	55%

4.2. Predictive validity

Predictive validity of a model can be tested by comparing intermediate and final outcomes from the model against outcomes from primary data (Sendi, Craig et al. 1999) in order to assess whether the model is adequately reflecting the data from which it is derived. Outcomes have also been compared against secondary data – see section 3. Figure 17 compares the modelled survival curves for ART and No-ART to the primary/secondary data points from which these are derived[10]. The figure shows that the model perfectly replicates death data thereby indicating very strong predictive validity. Similarly, when modelled output is compared to primary data detailing the proportion of patients (out of those surviving) switching on to the second-line regimen, the modelled output shows a highly accurate replication of the data from which it is derived.

[10] The No-ART modelled output is validated against an exponential curve fitted to a point estimate of just under 50% surviving with CD4<200 cells/μl at 24 months from secondary data (Badri, Bekker et al. 2004).

Figure 17: Comparison of proportion surviving and proportion on second-line between primary data and modelled output



4.3. Face validity

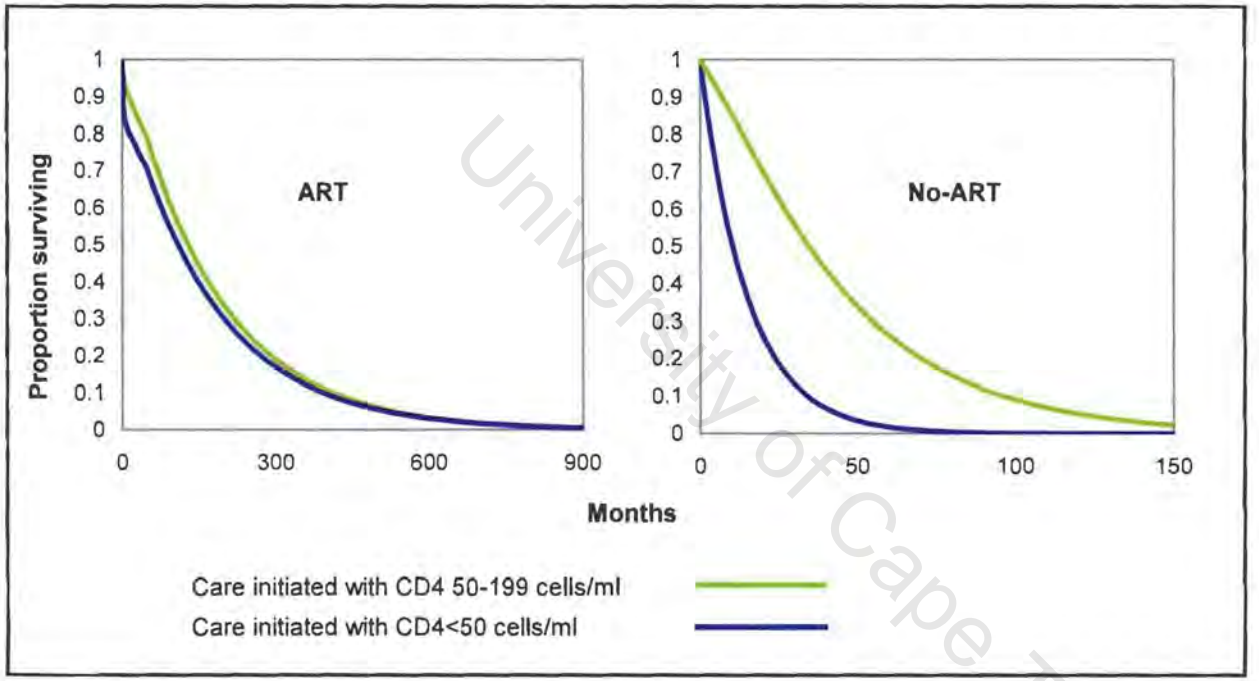
Face validity assesses whether the model produces the output that one would expect (Sendi, Craig et al. 1999). Expected model output could include:

- Higher outcomes for ART than for No-ART
- Higher outcomes for patients starting with a CD4 level between 50 and 199 cellsμl versus CD4<50 cellsμl
- Higher lifetime costs for ART versus No-ART

It has already been shown that patients on ART survive for longer than those receiving No-ART (see Figure 15) and lifetime cost results will be presented in section 5. It remains to be established whether outcomes are lower for patients entering care with lower CD4 levels. This is validated by altering the base case "seeding" of the models – in other words, the entire cohort is assumed to enter care with CD4 50-199 cellsμl or is assumed to enter care with CD4<50 cellsμl and survival is established. Results are presented in Figure 18.

The figure indicates that survival is worse for patients entering care later, and this is especially marked for the No-ART group. In the ART group, although initial survival is impacted by a lower CD4 level at base case, patients who survive beyond the first year tend to recover and perform similarly to patients entering care with higher CD4 levels.

Figure 18: Survival curves for patients initiating care with CD4 50-199 cells/μl or CD4<50 cells/μl



4.4. Modelling process validity

The final step in model validation is establishing the modelling process validity. Here, the lifetime costs, outcomes and cost-effectiveness conclusions of published studies have been compared to the results obtained through the models in this thesis. Results include non HIV-related mortality because one is comparing against published results that ought to include deaths from all causes. A published review of HIV-treatment and prevention cost-effectiveness studies was the primary source of published studies (Harling, Wood et al. 2005). A small number of additional studies were also identified.

Studies were included if:

- Total direct health care costs were calculated (studies were excluded if they only calculated the incremental costs of the programme under evaluation[11])
- Final outcomes were calculated (final outcomes could be life years or QALYs)
- Interventions included treatment and prophylaxis[12] of opportunistic and HIV-related infections and/or ART for adult patients

A total of 57 English language studies were identified based on Harling, Wood et al. (2005) and an additional four studies were identified that have been published more recently, leading to a total of 61 studies.

Of these studies:

- Three were excluded because they focused on a subset of HIV disease (e.g. only patients with pneumocystis carnii pneumonia)
- Nine were excluded because they did not calculate final outcomes
- Four were excluded that only calculated incremental costs (e.g. only isoniazid prophylaxis for tuberculosis was included in costs and all other HIV-related health care costs were excluded)

[11] These studies were excluded because the overall conclusions of the study in terms of cost-effectiveness were not comparable.
[12] Prophylactic interventions were included in the review if the type of prophylaxis is currently recommended in South Africa.

- Seventeen were excluded because it was found that they did not include either ART or No-ART (for example, a number of studies evaluated ARV monotherapy or dual therapy)
- Two were excluded because they focused on a different patient population (e.g. children)
- Three were excluded because no journal access was available in South Africa either from the University of Cape Town or interlibrary loans

After these 38 exclusions, 23 studies remained.

Of these, only three studies explicitly compared ART to No-ART (Sendi, Bucher et al. 1999; Freedberg, Losina et al. 2001; Bachman 2006) and another 3 compared the cost-effectiveness of starting ART within different CD4 bands and included a No-ART comparator (Schackman, Goldie et al. 2001; Schackman, Freedberg et al. 2002; Badri, Cleary et al. 2006). One study compared monotherapy to No-ART (Moore, Hidalgo et al. 1994). Five studies compared triple ART to dual therapy (Moore and Bartlett 1996; Cook, Dasbach et al. 1999; Anis, Guh et al. 2000; Miners, Sabin et al. 2001). Eleven studies evaluated various prophylactic agents for opportunistic infections versus No-ART without prophylaxis or ART in the absence of prophylaxis (Moore and Chaisson 1997; Bayoumi and Redelmeier 1998; Freedberg, Scharfstein et al. 1998; Paltiel and Freedberg 1998; Goldie, Weinstein et al. 1999; Sendi, Craig et al. 1999; Paltiel, Goldie et al. 2001; Goldie, Kaplan et al. 2002; Yazdanpanah, Goldie et al. 2003; Hoffmann and Brunner 2004; Yazdanpanah, Losina et al. 2005).

Of all studies included, only 3 were based in developing countries (Yazdanpanah, Losina et al. 2005; Bachman 2006; Badri, Cleary et al. 2006).

From these studies, it is only possible to draw broad conclusions about the range of plausible outcomes for ART and No-ART. This is partly because in order to evaluate different HIV-treatment interventions, the studies had different assumptions about when patients would enter care. In addition, a number of studies only presented discounted results. Results are therefore discussed separately for zero and other discount rates. The majority of studies presented outcomes as life years, and some presented outcomes as life years and QALYs. The discussion therefore focuses on life years because this allows comparison across all studies with the exception of one that only calculated QALYs. However, full details of outcomes are presented in Appendix B.

The undiscounted outcomes for No-ART patients receiving no prophylaxis for opportunistic infections ranged between 0.6 life years for patients with AIDS (Moore, Hidalgo et al. 1994) to 8.2 life years for patients entering care with CD4>350 cells/μl (Bachman 2006). The latter was particularly high – most studies with patients entering care with CD4 350-500 calculated outcomes between 3.2 and 6.6 life years. This is comparable to the 2.9 life years calculated in this thesis given that patients had much lower CD4 levels.

In studies that presented discounted results (all at 3 per cent per annum), the majority calculated outcomes between 2 and 4 life years for patients entering care in a variety of CD4 strata (Sendi, Bucher et al. 1999). In this thesis, outcomes are 2.7 life years for No-ART at a 3 per cent discount rate.

For ART undiscounted outcomes ranged between 5.8 (Hoffmann and Brunner 2004) and 18.8 life years (Badri, Cleary et al. 2006) while the majority of studies calculated outcomes between 14 and 15 life years. For studies that discounted (majority at 3 per cent but one at 4 per cent), outcomes ranged between 5.3 (Freedberg, Losina et al. 2001) and 11.4 (Yazdanpanah, Goldie et al. 2003) life years, but most studies calculated lifetime outcomes that were closer to 10 years than to 5 years.

Because ART is still a new intervention, a number of studies have tended to be conservative about the durability of treatment (Freedberg, Losina et al. 2001; Schackman, Goldie et al. 2001; Schackman, Freedberg et al. 2002). Other studies that have relaxed these durability assumptions have calculated much higher benefits from ART (Sendi, Bucher et al. 1999; Bachman 2006; Badri, Cleary et al. 2006) In this thesis, life years are 13 at a zero discount rate and 9.5 at a 3 per cent annual discount rate[13]. These results are plausible when compared with published studies.

The final step in modelling process validity is to compare the ICERs of published studies to the results in this thesis. Because costs are highly context specific, comparison will only be made between this thesis and studies conducted in other developing countries. Based on the candidate's

[13] Given the low discount rate, one might be surprised by the large drop in life years from 13 to 9.5. In fact, the model runs for considerably longer than these years to calculate these results owing to the shape of the survival curve that is assumed in Markov modelling.

own calculations[14], the two studies from developing countries had ICERs of ART versus No-ART of US$2,486 (Bachman 2006) and US$504 (Badri, Cleary et al. 2006) per life year gained. However, the former derives ARV costs from Cleary, Boulle et al. (2004) at about U$1,475 including laboratory investigations whereas the latter used the South African government tender prices of US$558 per annum for first-line. This thesis calculates an ICER of US$1,023 per life year gained which is higher than Badri, Cleary et al. (2006). However, there are a number of key differences that would influence ICER results. In the Badri, Cleary et al. (2006) study, the ICER was heavily dependent on high lifetime costs for No-ART patients (of US$7,877 versus US$2,966 in this thesis). This is because patients in the Badri, Cleary et al. (2006) study enter the model with CD4>350 cells/μl whereas in this thesis, they enter the model with CD4<200 cells/μl. The lifetime costs of ART in the two studies are more comparable, at approx US$13,191 in this thesis and US$14,230 in Badri, Cleary et al. (2006). ICERs from developed countries are much higher and are not comparable given the much higher lifetime costs of treatment in these settings (although the outcomes have been found to be similar in the above sections).

Across all studies, none that took a provider's perspective found ART to be cost-saving over No-ART, but instead based their cost-effectiveness recommendations on whether the ICER was a good buy in the particular context. One study that took a societal perspective (including productivity gains) in Switzerland found ART to be cost-saving, but not from a provider's perspective (Sendi, Bucher et al. 1999). A recent study of the costs and benefits of HIV treatment to a Ugandan firm found ART in combination with cotrimoxazole to be cost saving over a 5-year time frame, but if costs and outcomes were calculated over a ten-year period, this result no longer held (Marseille, Saba et al. 2006). The latter study illustrates the difficulty of decision-making for HIV-treatment where uncertainty relating to future effectiveness requires assumptions to be made that can radically alter the study conclusions.

Results of all studies included in the review are presented in Appendix B.

[14] These studies included interventions that are not comparable with this thesis. ICERs were therefore re-calculated in order to compare the incremental cost per life year gained between ART and No-ART.

5 Results

Having described the process of attaching costs and outcomes to Markov states and the extensive validation of Markov models, this section finally presents patient-level lifetime costs, individual health gains and ICERs under a base case scenario, where data from the Khayelitsha pilot are unadjusted; a generalized scenario, where a number of adjustments are made based on secondary literature; and a low cost generalized scenario which explores the possibilities for additional cost savings. The final section considers the implications of initiating ART within different CD4 categories. This section draws on earlier work (Badri, Cleary et al. 2006) as well as insights from primary data.

5.1. Base case scenario

For the base case results, uncertainty relating to data requirements has been assessed using probabilistic sensitivity analysis (PSA), which propagates parameter and underlying modelling uncertainty through the model by means of first and second-order Monte Carlo simulation. Simulations have been run using 1,000 distribution values, and each distribution value has been subjected to 1,000 first-order simulations. This allows overall parameter uncertainty to be captured as confidence intervals around lifetime costs, outcomes and ICERs (Briggs 1995; O'Hagan, McCabe et al. 2005).

Figure 19 shows scatterplots of discounted ICERs that have been generated through these simulations. Each of the 1,000 points (derived from second-order Monte Carlo sampling) shows the mean incremental costs and incremental outcomes of first-line ART over No-ART and first and second-line ART over first-line ART. Each of these points has been generated from 1,000 first order simulations.

Figure 19: ICER scatterplots (QALYs, discounted at 3 per cent per annum)

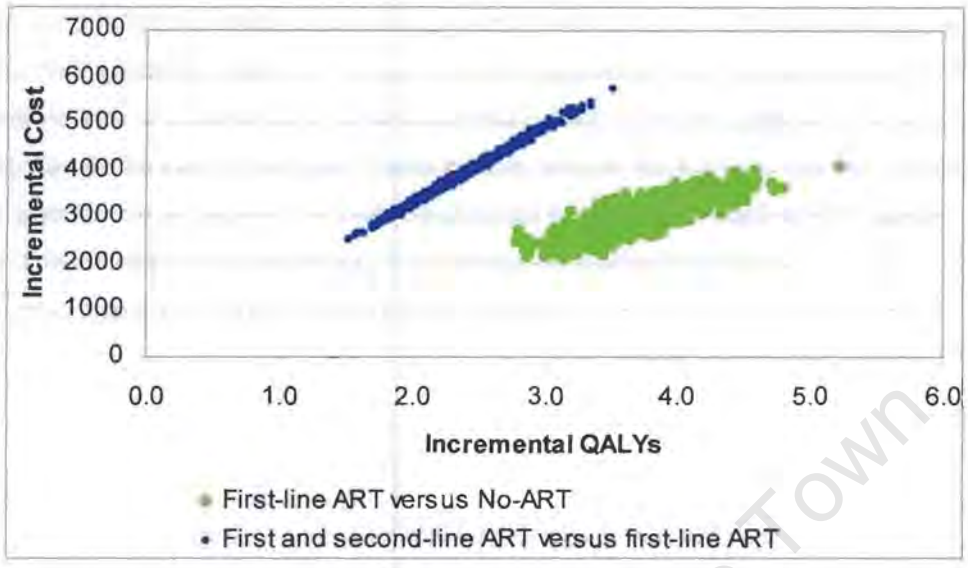
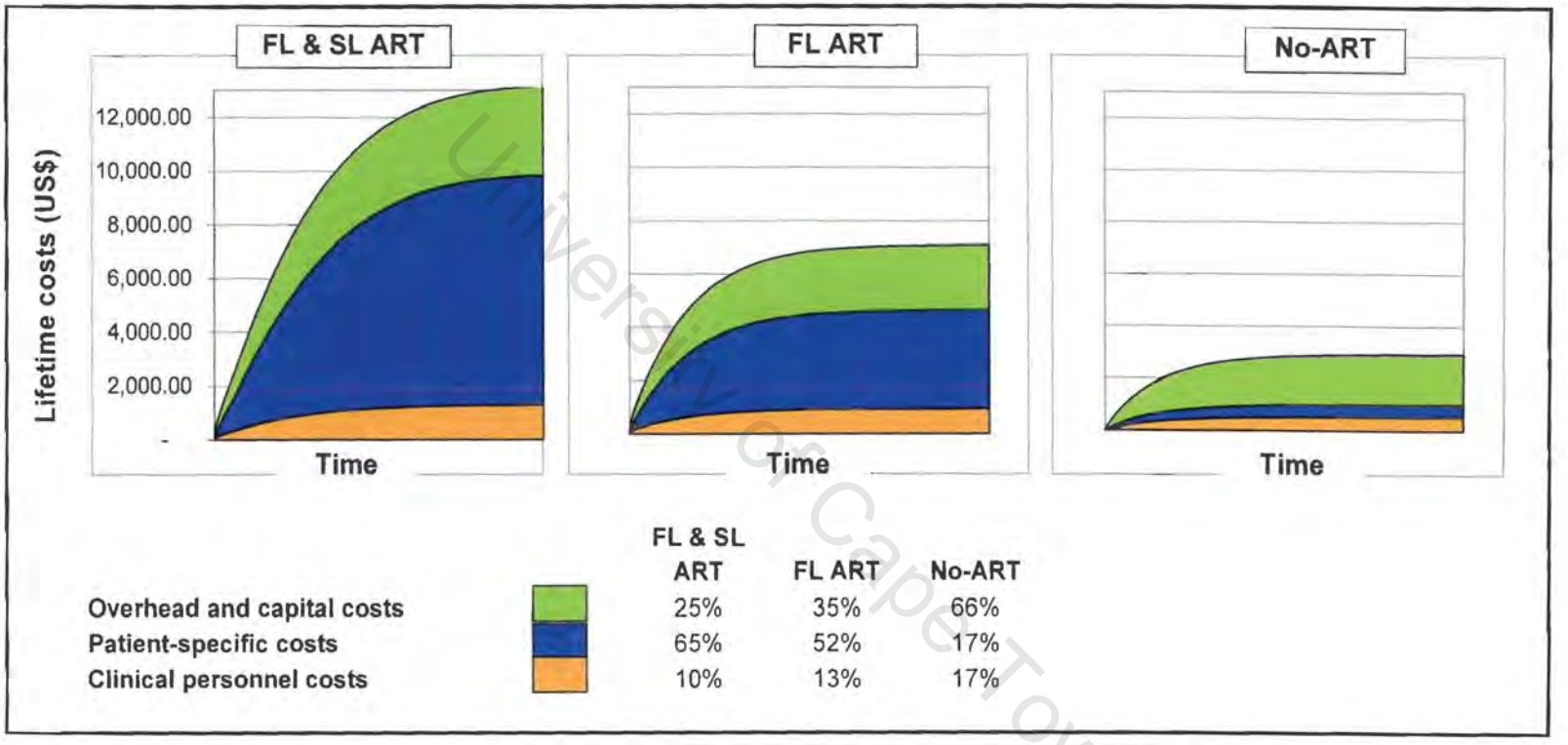


Table 30 shows the results of these simulations as 95 per cent confidence intervals around lifetime costs, outcomes and ICERs – these indicate, with 95 per cent probability, the range in which results can be expected to fall. The mean results have been calculated from cohort simulation as this gives expected values whereas Monte Carlo simulation gives a different result with every simulation. At a zero annual discount rate, the model estimated a mean survival of 2.9 years for No-ART, 8.5 life years for first-line ART and 12.9 life years for first and second-line ART which translates into 2.1, 7.1 and 10.8 QALYs respectively. Discounted per patient lifetime costs were US$2,966 for No-ART, US$5,779 for first-line ART and US$9,435 for first and second-line ART. The full breakdown of these costs by service category and by human resource components (doctors and nurses) is presented in Figure 20. Sixty-five per cent of lifetime costs on first and second-line ART were for ARVs, laboratory investigations and other medicines while overhead and capital costs were the largest cost driver for No-ART at 66 per cent of lifetime costs. Lifetime clinical staff costs were US$516 for No-ART, US$931 for first-line ART and US$1,305 for first and second-line ART. First-line ART had a discounted ICER of US$795 per QALY gained versus No-ART while first and second-line ART had an ICER of US$1,625 versus first-line ART. ICERs were slightly lower per LY gained. Although not presented in the table, the discounted ICER for first and second-line ART versus No-ART would be approximately US$1,102 per QALY gained.

Table 30: Base case results, including PSA

Treatment option	Lifetime costs (95%CI)	Outcomes		ICER	
		Life Years (95%CI)	QALYs (95%CI)	Life Years (95%CI)	QALYs (95%CI)
Undiscounted					
No-ART	2,966 (2,611-3,343)	2.9 (2.6-3.3)	2.1 (1.8-2.3)		
First-line ART	7,085 (6,099-7,988)	8.5 (7.1-9.6)	7.1 (5.9-8.1)	736 (663-831)	816 (726-917)
First and second-line ART	13,191 (11,167-16,056)	12.9 (11.1-15.2)	10.8 (9.1-12.5)	1,388 (1,362-1,414)	1,641 (1,621-1,683)
Discounted					
No-ART	2,743 (2,414-3,057)	2.7 (2.4-3.0)	1.9 (1.7-2.1)		
First-line ART	5,779 (5,149-6,351)	6.9 (6.0-7.6)	5.7 (5.0-6.3)	723 (652-846)	795 (706-911)
First and second-line ART	9,435 (8,414-10,891)	9.5 (8.5-10.7)	8.0 (7.3-8.6)	1,365 (1,344-1,398)	1,625 (1,601-1,665)

Figure 20: Lifetime cost curves for first and second-line ART, first-line ART and No-ART in the base case scenario



5.2. Generalized scenario

This section makes a number of adjustments to the base case scenario in order to enhance the generalizability of estimates. The following adjustments have been made to utilisation data:

- ART visits are based on national ART guidelines
- No-ART visits are based on a local natural history cohort (Badri, Cleary et al. 2006) under the assumption that patients with CD4<50 cells/μl use the same visits as those with CD4<200 cells/μl and AIDS while patients with CD4 50-199 cells/μl use the same visits as those with CD4<200 cells/μl who are in other WHO stages

The following adjustments have been made to unit costs:

- Inpatient care is at secondary level hospitals
- No-ART visit costs are based on the costs from clinics and community health centres (costs from hospital outpatient departments have been excluded)

Outcomes have been adjusted by increasing the death transition probabilities on ART and No-ART such that 18 per cent of the ART cohort is dead by one year based on ART-LINC data (ART-LINC and ART-CC 2006) and the median No-ART survival is 11 months based on a review of natural history data from resource poor settings (Schneider, Zwahlen et al. 2004). These adjustments might make results more applicable to routine settings.

Table31: Generalized results

Treatment option	Lifetime costs	Outcomes		ICER	
		Life Years	QALYs	Life Years	QALYs
Undiscounted					
Generalized No-ART	1,813	2.3	1.6		
Generalized first-line ART	5,615	6.9	5.7	824	919
Generalized first and second-line ART	9,474	9.7	8.1	1,387	1,651
Discounted					
Generalized No-ART	1,706	2.1	1.5		
Generalized first-line ART	4,716	5.7	4.8	835	921
Generalized first and second-line ART	7,215	7.6	6.3	1,374	1,635

When comparisons are made between the base case and generalized scenario, lifetime costs were reduced by around US$1,150 for No-ART, US$1,500 for first-line ART and US$4,000 for first and second-line ART. No-ART life expectancy was decreased by 0.6, first-line ART life-expectancy was decreased by 1.6 and first and second-line ART life-expectancy was decreased by 3.2. Although costs and outcomes were reduced, note the similarity between the generalized and base case ICERs.

5.3. Low cost generalized scenario

This scenario extends the generalized framework to consider the impact of additional cost saving assumptions:

- No viral load laboratory investigations
- All patients receive nevirapine (US$17.12 per quarter) instead of efavirenz (US$ 86.14 per quarter) in the first-line regimen

While data are insufficient to make adjustments to clinical outcomes when efavirenz is replaced by nevirapine, it should be noted that outcomes might be lower because nevirapine has been shown to be associated with more side-effects than efavirenz.

Table 32: Low cost generalized results

Treatment option	Lifetime costs	Outcomes		ICER	
		Life Years	QALYs	Life Years	QALYs
Undiscounted					
Low cost No-ART	1,813	2.3	1.6		
Low cost first-line ART	4,204	6.9	5.7	518	578
Low cost first and second-line ART	7,841	9.7	8.1	1,307	1,556
Discounted					
Low cost No-ART	1,706	2.1	1.5		
Low cost first-line ART	3,541	5.7	4.8	509	561
Low cost first and second-line ART	5,895	7.6	6.3	1,294	1,541

Table 32 indicates that additional cost savings are possible through the use of cheaper ARVs and by eliminating viral load tests. Cost savings are around US$3,000 and US$5,000 when the low cost first-line and first and second-line results are compared to the base case. The ICERs of low cost options are also more favourable than the base case ICERs presented in Table 30 indicating that the cost-effectiveness of either ART option might be higher in routine settings if low cost options are pursued.

5.4. Debating when to start ART

In South Africa, a patient is medically eligible for ART if she or he has an AIDS diagnosis at any CD4 level or a CD4 count of less than 200 cells/μl at any WHO stage. While these medical eligibility criteria are usually thought of as merely technical, in reality they reflect a trade-off between maximising individual patient benefits and increasing coverage. Because of this trade-off, patients are eligible for ART at a point when their capacity to benefit is reduced owing to their advanced stage of illness. This section considers the impact of prioritizing patients according to their capacity to benefit. The discussion in this section will draw on two separate analyses. Firstly, data in this thesis allows an assessment of the cost-effectiveness of initiating ART with CD4 50-199 cells/μl and CD4<50 cells/μl. Secondly, earlier work reported in Badri, Cleary et al. (2006) allows an assessment of the cost-effectiveness of initiating ART with CD4>350 cells/μl, CD4 200-350 cells/μl and CD4<200 cells/μl.

The data in this thesis, as presented in Figure 18, have shown that survival is slightly lower for patients who start ART with CD4<50 cells/μl in comparison to those who start with CD4 50-199

cells/μl. However, the relatively ill patients that survive the first year on ART can still attain substantial benefits. What this analysis did not show is that many patients would die prior to starting ART if therapy initiation were delayed.

In order to provide further understanding of the implications of starting ART once severely immune compromised, it is necessary to assume that all enter the models with CD4 50-199 cells/μl, from where they are "randomised" to start ART immediately, start ART once their CD4 has dropped below 50 cells/μl, or to continue with the No-ART option[15]. For simplicity, only the first and second-line ART option is modelled. Results are presented in Table 33. Discounted (at 3 per cent annual rate) QALYs are 2.5, 3.3 and 8.3 and lifetime costs (US$) are 3,090, 4,364 and 9,695 for No-ART, starting ART with CD4<50cells/μl and starting with CD4 50-199cells/μl respectively. Following Siegel, Weinstein et al. (1996), ICERs have been calculated by first ordering interventions from the least effective to the most effective. Then the ICER of starting ART at CD4 50-199 cells/μl is calculated against starting ART at CD4<50 cells/μl which in turn is calculated against No-ART. However, if an option is dominated, then the ICER is calculated against the next un-dominated intervention. Absolute dominance is defined as the case where a more effective option is less costly. While there is no absolute dominance in this scenario, starting ART with CD4<50 cells/μl is dominated through extended dominance. Essentially this means that one could maximise benefits through following a mixed strategy of starting ART with CD4 50-199 cells/μl and No-ART. Because starting ART with CD4<50 cells/μl is dominated, the ICER for starting ART earlier is calculated against No-ART.

The following conclusions can be drawn about starting ART in the lower CD4 stratum:

- Outcomes are considerably lower if ART is started during late immune decline, although there are still benefits in comparison to No-ART
- The majority of deaths occur prior to ART initiation (this is deduced by comparing the considerably higher outcomes that are calculated in Figure 18 which showed outcomes for patients with CD4<50 cells/μl from the time of initiating ART)

[15] Note that the results in Table 12 differ from those in Table 9 because all patients enter care in the higher CD4 stratum.

This result has implications for prioritizing patients if one is unable to offer treatment for all patients presenting with CD4<200 cells/μl. If patients with CD4<50 cells/μl are prioritized, the outcomes of healthier patients are jeopardized. If one is concerned with maximising outcomes and maximising the coverage of patients in need in the short-run, then patients with higher CD4 levels should be prioritised for treatment over sicker patients. In the long-run however starting ART in patients with higher capacity to benefit has a higher opportunity cost in terms of health care resources (i.e. higher lifetime cost) than starting in the lower stratum.

Table 33: Implications of starting ART with CD4 50-199 cells/μl versus CD4<50 cells/μl

Treatment option	Lifetime costs	Outcomes		ICER	
		Life Years	QALYs	Life Years	QALYs
Undiscounted					
No-ART	3,401	3.8	2.7		
Start ART at CD4<50 cells/ml	6,366	5.7	4.8	dominated	dominated
Start ART at CD4 50-199 cells/ml	13,605	13.5	11.3	1,061	1,192
Discounted					
No-ART	3,090	3.5	2.5		
Start ART at CD4<50 cells/ml	4,364	3.9	3.3	dominated	dominated
Start ART at CD4 50-199 cells/ml	9,695	9.9	8.3	1,031	1,135

While the preceding analysis provides insight into the alternative costs and outcomes associated with ART in patients with CD4<200 cells/μl, it should be borne in mind that the medical eligibility criteria for ART in South Africa are relatively conservative. Badri, Cleary et al. (2006) have used a similar approach to that described in this thesis to assess the cost-effectiveness of starting ART within higher CD4 strata. In this analysis, all patients enter care with CD4>350 cells/μl from where they start ART immediately, start ART with CD4 200-350 cells/μl, start with CD4<200 cells/μl or to continue with the No-ART status quo. Results indicate that undiscounted life expectancy could be 23, 21, 19 and 6 years from each of these strategies. However, earlier initiation was inevitably associated with higher lifetime costs given that at least 50 per cent of ART costs are associated with recurrent medication and laboratory investigations. The study found that the ICER was lowest for starting ART at CD4<200 cells/μl (US$611 per QALY gained at a zero discount rate), followed by 200-350 cells/μl (US$915 per QALY gained) and finally >350 cells/μl (US$1,236 per QALY gained).

6 Summary

This chapter has presented extensively justified patient-level lifetime costs and individual health gains associated with alternative HIV-treatment interventions. The building blocks of these estimates have included widely generalizable unit costs, estimates of the utilisation of clinic visits from 1,729 patients, estimates of the utilisation of TB and inpatient services from 670 patients, secondary HRQoL data from a sample of patients in the same clinics (Jelsma, MacLean et al. 2005), ART transition probabilities estimated from 1,729 patients over a maximum follow-up of four years, and No-ART transition probabilities from a local natural history cohort (Post, Wood et al. 1996; Badri, Bekker et al. 2004). These building blocks have been fed into the Markov models, which have been shown to perform well in terms of technical, predictive, face and modelling process validity. Uncertainty relating to data requirements has been assessed in probabilistic sensitivity analysis; the very narrow confidence intervals generated through this analysis suggest that one can place confidence in the results that have been produced.

Patient-level results have been produced in four scenarios. In the base case scenario, results were estimated from Khayelitsha data. The generalized scenario made a number of adjustments to cost, utilisation and clinical outcomes based on secondary literature. While ICERs in this scenario were similar to base case, lifetime costs and outcomes were reduced. The low cost generalized scenario explored the implications of using cheaper first-line ARVs and removing viral load testing. This scenario indicated that cost-effectiveness could be considerably improved through these strategies. In each of these scenarios, it was indicated that the highest individual benefits could be achieved through first and second-line ART. However, if access to second-line were restricted, it is likely that more patients could access first-line ART.

The final section considered the implications of starting ART in alternative CD4 strata. This concluded that capacity to benefit would be improved if ART were initiated at CD4 50-199 cells/μl. However, defining need as capacity to benefit does not necessarily mean that the most effective intervention should be implemented. The opportunity cost of the additional resources associated with starting ART earlier would need to be considered given resource scarcity. On the other hand, initiating care in a state of advanced immune suppression (CD4<50 cells/μl) is dominated through extended dominance. What this means is that a health maximising strategy would suggest that patients who present for care with CD4 50-199 cells/μl should be prioritised.

Chapter 7: Population-level results

1 Introduction

This chapter extends patient-level results to the population-level. While patient-level analyses were concerned with calculating lifetime costs and individual health gains, at the population-level one is concerned with assessing the total costs and total QALYs that can be gained through delivering each intervention to a defined population in need over a given planning time frame. These results are fed into mathematical programming algorithms in order to estimate the proportion of patients receiving alternative HIV-treatment interventions, the QALYs that can be gained and the percentage of need that can be met in each of three different social choice rules. In health maximisation, technical efficiency is assessed through choosing the mutually exclusive HIV-treatment intervention(s) that maximises health within the budget allocated to one independent health care programme (HIV-treatment). In the decent minimum, health is maximised within the budget constraint with the proviso that all patients in need receive one of the mutually exclusive HIV-treatment interventions. In equal health, health is maximised within the budget constraint with the constraint that all patients receive the same health gain (in other words, they receive the same HIV-treatment intervention). By comparing the percentage of need that can be met and the QALYs that can be gained at the same budget in each social choice rule, the equity/efficiency trade-off is apparent. These analyses are described in detail in this chapter, and form a key part of the discussion in Chapter 7.

The chapter has three parts. The first part sets the scene for subsequent sections by discussing the HIV-treatment scenarios that will be considered; the choice of the decision-making time frame; the need for HIV-treatment; the patient targets contained in the "Operational Plan for HIV and AIDS Care" (Department of Health 2003); and the health care budget and clinical personnel available for HIV-treatment. The second part of the chapter presents total costs and QALYs if each of the different HIV-treatment programmes were to reach 100 per cent of need over the decision-making period. These totals are then used in the mathematical programming algorithms to assess the implications of alternative social choice rules. The third part models the resources required to implement the targets contained in the Operational Plan and considers how the current programme could be offered in a more efficient or equitable manner.

2 Background

2.1. Scenarios

Total costs and QALYs will be calculated for three mutually exclusive HIV-treatment interventions grouped into two scenarios:

Base case scenario (derived from patient-level costs and outcomes as per Khayelitsha data)

- No-ART
- First-line ART
- First and second-line ART

Generalized scenario (includes adjustments to unit costs, utilisation and outcomes to enhance generalizability of results)

- No-ART
- First-line ART
- First and second-line ART

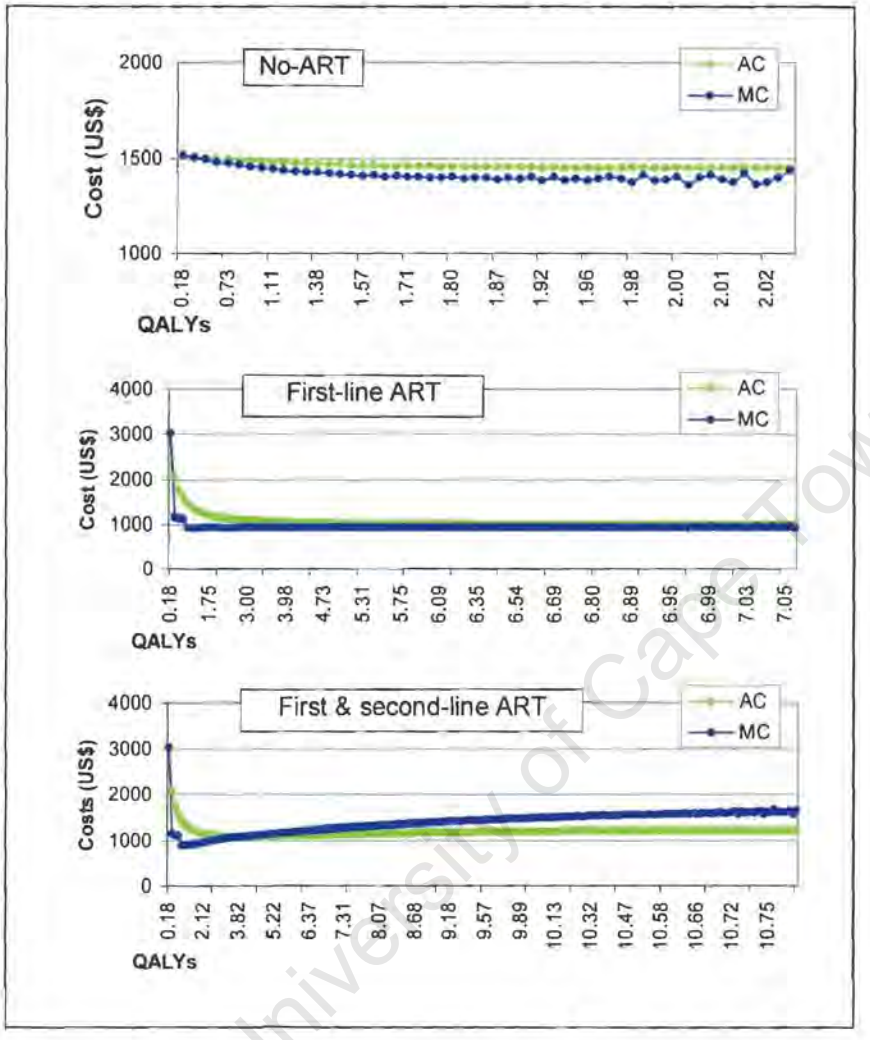
Two scenarios from the previous chapter will be excluded from this analysis. The low cost generalized scenario will not be considered because while it is likely to have an impact on outcomes, no data exist about what this impact might be. If it were considered, it would dominate certain options (having lower costs but identical outcomes) which might not be realistic. Secondly, the impact of prioritising patients based on their capacity to benefit will not be considered. In order to consider this scenario, it would be necessary to calculate the numbers of patients in need of treatment within different CD4 strata, but current demographic models do not provide these data.

2.2. Decision-making time frame

A key choice in HIV-treatment priority setting is that of the decision-making time frame. Most frequently, HIV/AIDS strategic plans tend to cover a five-year time period. South Africa's first strategic plan covered the period 2000-2005, the Operational Plan covered 2003-2008 and a new strategic plan is currently being developed to cover a four-year period from 2007 to 2011. However, this does not imply that these are the appropriate time frames for decision-making.

Patient-level models can assist in the choice of the decision-making time frame, through an assessment of the relationship between the marginal and average cost and the level of the ICER at different levels of output for each mutually exclusive intervention in the base case scenario. Figure 21 shows the average total cost (total cost divided by total QALYs) and the marginal cost (the change in total costs divided by the change in QALYs) over an individual's lifetime. Average and marginal costs are higher during the initial period on treatment and this is particularly marked in the two ART options. For No-ART and first-line ART, marginal cost approaches average cost when a very high proportion of the cohort is dead. What this means is that while all patients will have higher costs at the time of death (owing to the death transition cost), when the distribution of costs is compared across an infinite cohort, these higher costs even out because different patients will be dying at different points. However, for first and second-line ART, there is a large increase in costs as patients move on to the second-line regimen. This causes the marginal cost to exceed the average cost after 4 QALYs have been gained on treatment. The implication of this analysis is that limiting the time frame of analysis for first and second-line ART in particular will tend to overstate the cost-effectiveness of this intervention in comparison to first-line ART.

Figure 21: Average and marginal cost of interventions



This can be further unpacked by exploring the impact on results of the modelling time horizon. To do this, the base case models were run over 5 years, 10 years and until 100 per cent of the cohort was dead. Undiscounted per-patient lifetime costs ranged between US$2,516 and US$2,966 for No-ART; between US$3,473 and US$7,085 for first-line ART; and between US$3,860 and US$13,191 for first and second-line ART under different modelling time horizons (Table 34) while undiscounted QALYs ranged between 1.7 and 2.1 for No-ART; between 3.2 and 7.1 for first-line ART; and between 3.5 and 10.8 for first and second-line ART. Mean ICERs (calculated for life years and QALYs) are presented in Table 34. The discounted incremental cost per QALY gained of first-line ART versus No-ART is US$636, 736 and 795 under a 5-year, 10-

year and unlimited horizon. Conservative time horizons also tend to mask the difference in lifetime costs and QALYs as shown by the lower incremental costs and incremental QALYs under 5-year, 10-year and unlimited time horizons.

Table34: Lifetime costs, effectiveness and ICERs of interventions with alternative modelling time horizons

Treatment option	Lifetime costs	Outcomes		ICER	
		Life Years	QALYs	Life Years	QALYs
Undiscounted					
Simulation over 5 years					
No-ART	2,516	2.4	1.7		
First-line ART	3,473	3.9	3.2	638	634
First and second-line ART	3,860	4.3	3.5	968	1,334
Simulation over 10 years					
No-ART	2,881	2.8	2.0		
First-line ART	5,123	6.0	5.0	701	742
First and second-line ART	6,617	7.2	6.0	1,245	1,540
Simulation until 100% dead					
No-ART	2,966	2.9	2.1		
First-line ART	7,085	8.5	7.1	736	816
First and second-line ART	13,191	12.9	10.8	1,388	1,641
Discounted					
Simulation over 5 years					
No-ART	2,391	2.3	1.6		
First-line ART	3,275	3.7	3.0	631	636
First and second-line ART	3,620	4.0	3.3	1,150	1,327
Simulation over 10 years					
No-ART	2,686	2.6	1.85		
First-line ART	4,600	5.4	4.5	684	736
First and second-line ART	5,823	6.3	5.3	1,359	1,529
Simulation until 100% dead					
No-ART	2,743	2.7	1.9		
First-line ART	5,779	6.9	5.7	723	795
First and second-line ART	9,435	9.5	8.0	1,406	1,625

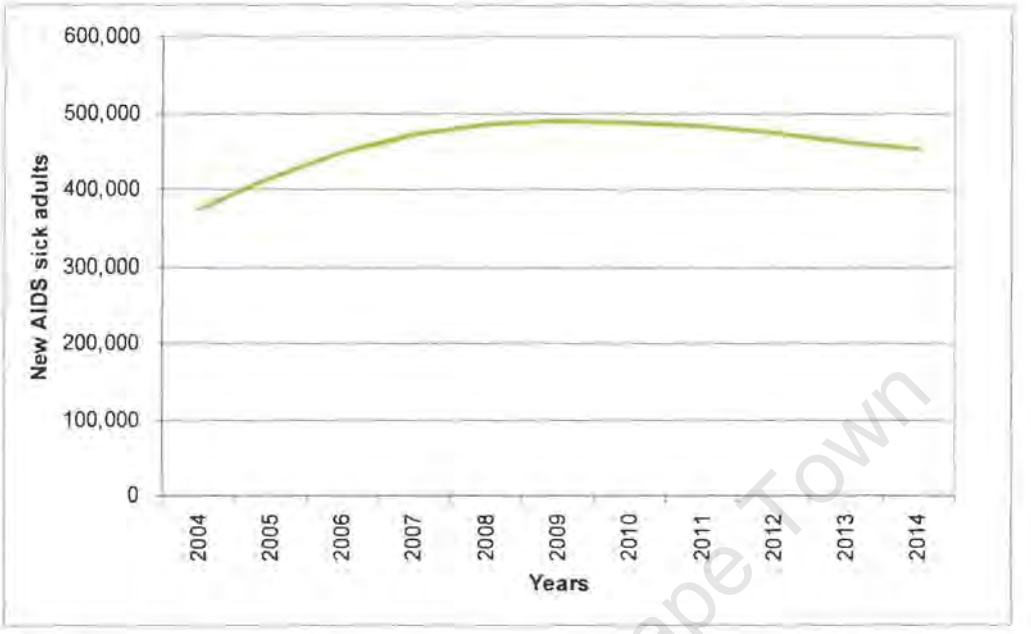
The implication is that very short time horizons tend to overstate the cost-effectiveness of the two ART strategies, and to underestimate lifetime costs and outcomes. If HIV-treatment budgets were based on shorter time horizons this could result in fewer patients being treated than had been planned. Based on this reasoning, March 2004 has been chosen for the start of the projection because although the rollout of ART was announced in late 2003, service delivery mainly began during 2004. Given the likelihood of future technological innovation in this field and the uncertain impact of the provision of ART on future incidence and prevalence of HIV/AIDS (see (Blower, Bodine et al. 2005)) projections have been limited to ten-year duration. Any change in

incidence of HIV from the start of ART in the country until 2014 would not have a significant impact on the numbers of patients developing AIDS. This is because this projection will primarily be dealing with patients who are already HIV-positive as of 2006. By limiting the projection to ten years and therefore primarily dealing with people who are already HIV-positive, this analysis avoids having to make assumptions about the impact of ART on HIV incidence. A potentially optimal strategy for strategic planning could be to continue to update HIV/AIDS strategic plans every five years, but to base these plans on a ten-year analysis of costs and outcomes.

2.3. Need

As explained in the methodology, the number of people in need of HIV-treatment is assumed to be equivalent to the number of new AIDS cases each year based on ASSA2003*lite* estimates over the period March 2004 until March 2014. It is worthwhile reiterating that because one is using new AIDS cases as the proxy for need, when one meets a certain percentage of this new need, one has met a lower percentage of total need. Figure 22 shows the number of adults newly developing AIDS, as calculated from ASSA2003*lite*.

Figure 22: New AIDS sick adults between 2004 and 2014



Source: extracted from ASSA2003*lite* AIDS and Demographic Model of the Actuarial Society of South Africa (release 24 November 2005), as downloaded in December 2005 from www.assa.org.za

2.4. Treatment targets in the Operational Plan

The government committed to the public sector rollout of ART in October 2003. By mid-November of that year, the Operational Plan detailing the implementation of the rollout was released. This document contains targets for patients entering and remaining in care as detailed in Table 35. The plan does not contain a specific target for adults, so these have been calculated based on the average proportion of new adult AIDS cases out of new AIDS sick adults and children (94 per cent) between 2004 and 2014, from ASSA2003*lite*.

Table 35: Operational Plan targets of patients starting and remaining on ART

By end of :	New patients starting ART	New adults starting ART	Patients remaining on ART	Adults remaining on ART
Mar-04	53,000	49,820	53,000	49,820
Mar-05	138,315	130,016	188,665	177,345
Mar-06	215,689	202,748	381,177	358,306
Mar-07	299,516	281,545	645,740	606,996
Mar-08	411,889	387,176	1,001,534	941,442
Mar-09	551,089	518,024	1,470,510	1,382,279

Although the Operational Plan initially envisaged placing 53,000 people on ART between November 2003 and March 2004, in his State of the Nation Address delivered after the April 2004 general elections, President Mbeki stated that this patient target would only be met by March 2005 (Mbeki 2004). The analysis of the Operational Plan in this dissertation will therefore assume that the targets of new adults starting ART in Table 35 are lagged by one year.

The Operational Plan intended to provide ART to incrementally higher proportions of new AIDS cases until 100 per cent of need was reached after a 5-year period. A comparison between these lagged patient targets, new adult AIDS cases (from ASSA2003*lite*) and the proportion of need that is met is presented in Table 36. By March 2010, the Operational Plan target would be meeting 106 per cent of need – this is because the targets were designed to meet 100 per cent of need by March 2009 (when need was higher) and were based on an earlier version of the ASSA model which produced slightly different estimates. It will therefore be assumed that the target is to meet 100 per cent of need by March 2010, which requires starting an additional 488,468 patients on ART.

Table 36: New adults starting ART, new AIDS sick adults and the proportion of met need

By end of:	New adults starting ART	New AIDS sick adults	Met need (%)
Mar-05	49,820	361,866	14%
Mar-06	130,016	406,144	32%
Mar-07	202,748	440,803	46%
Mar-08	281,545	465,758	60%
Mar-09	387,176	481,351	80%
Mar-10	518,024	488,468	106%

When monitoring the extent to which ART patient targets are being met, the most appropriate standard is the target for patients *entering* care. However, the government only releases data on the number of patients *remaining* on ART. In order to assess whether patient targets are being met, one therefore has to estimate the number of patients that would be remaining in care if the targets for patients entering care were being met.

One way of doing this is by using Markov modelling, and this approach was taken by the authors of the Operational Plan. However, their estimation of patients remaining in care assumed that death and drop out rates would be close to zero during the earlier periods on ART, while primary data suggest a higher death and drop-out rate during the first year and much lower rates during later years. Table 37 contains a comparison between the Operational Plan's target for patients remaining in care versus calculations of patients remaining in care using base case and generalized scenario survival functions. The proportion of patients remaining in care is calculated by dividing the total remaining in care in year x by the total that entered care in year x and all preceding years. The table indicates that the Operational Plan is intending much higher patient retention than has been attained in Khayelitsha.

Table 37: Comparison between Operational Plan target of patients remaining in care to calculations from the base case and generalized scenarios

By end of:	Adults remaining in care - Operational Plan	Proportion remaining in care	Adults remaining in care - base case scenario	Proportion remaining in care	Adults remaining in care - generalized scenario	Proportion remaining in care
Mar-05	49,820	100%	45,700	92%	44,327	89%
Mar-06	177,345	99%	161,275	90%	155,073	86%
Mar-07	358,306	94%	335,717	88%	320,145	84%
Mar-08	606,996	91%	572,042	86%	541,748	82%
Mar-09	941,442	90%	891,472	85%	839,514	80%
Mar-10	1,382,279	88%	1,289,227	82%	1,207,184	77%

The implication is that when one evaluates the progress of the rollout, it becomes relatively complex to ascertain whether the patient targets are being met. A paper presented at the XVI International AIDS Conference (Abdullah 2006) reported that 42,000 patients were on ART by March 2005 and 142,773 by March 2006 in South Africa. If it is assumed that 94 per cent of these are adults, then approximately 39,000 adults were on treatment by March 2005 and over 132,000 by March 2006. If one compares these data to estimates using a survival function based on base case ART, the public sector had met 82 per cent of its patient targets by March 2006 or 86 per cent if the generalized ART scenario is assumed. These percentages are higher than those estimated using the survival function assumed by the authors of the Operational Plan. See Table 38.

Table 38: Numbers of patients receiving ART in the public sector and the proportion of the Operational Plan target being met

	Adults remaining in care - actual	Per cent Operational Plan target	Per cent base case target	Per cent generalized target
Mar-05	39,060	78%	85%	88%
Mar-06	132,779	75%	82%	86%

To summarize the discussion thus far, Table 39 presents the numbers of adults entering care based on the Operational Plan targets. Beyond 2010, it has been assumed that ART will be available to all new adult AIDS cases, based on ASSA2003*lite* estimates (in other words, the

Operational Plan target for 2010 is replaced by more recent estimates of need). Just fewer than 3 million adults are projected to enter care over the ten-year period and 77 per cent could be alive and remaining in care by the end of March 2014, if survival and patient retention is as assumed in the base case scenario.

Table 39: Adults entering and remaining in care under Operational Plan targets

By end of :	Adults entering care - target	Patients remaining in care - base case	Proportion remaining in care - base case
Mar-05	49,290	45,700	93%
Mar-06	128,633	161,275	91%
Mar-07	200,591	335,717	89%
Mar-08	278,550	572,042	87%
Mar-09	383,057	891,472	86%
Mar-10	488,468	1,289,227	84%
Mar-11	488,494	1,662,543	82%
Mar-12	483,561	2,010,266	80%
Mar-13	475,540	2,329,518	78%
Mar-14	465,945	2,618,936	77%
Total	2,976,183	2,618,936	88%

While the preceding table gives a picture of the patients receiving first and second-line ART and remaining in care, it is likely that patients who do not receive ART would receive some other form of care and support. Data suggest that the quality of care that patients receive in the absence of ART is lower than base case No-ART (Cleary, Chitha et al. 2005). In what follows, it will therefore be assumed that adults with AIDS who do not receive ART receive generalized No-ART. Table 40 shows the numbers of patients receiving generalized No-ART care (the difference between the number developing AIDS and those who received first and second-line ART) and remaining in care between March 2004 and March 2014. During this period, over 1.1 million adults in need of treatment would not be able to access ART, and 0.1 per cent would be alive and remaining in care by the end of March 2014. From March 2010, no new patients would be started on generalized No-ART as all new need would be treated with ART. However, those patients who were already on generalized No-ART would continue along this treatment path - if the target is to meet 100 per cent of *new need*, one is meeting less than 100 per cent of total need. The table

however indicates that if one has been meeting all new need on a continuous basis over 4 or 5 years, meeting new need becomes very similar to meeting total need.

Table 40: Patients receiving generalized No-ART and remaining in care

By year ending :	Patients receiving generalized No-ART	Patients remaining in care - generalized No-ART	Proportion remaining in care
Mar-05	312,046	260,074	83%
Mar-06	276,128	388,624	66%
Mar-07	238,055	440,584	53%
Mar-08	184,213	432,443	43%
Mar-09	94,175	356,403	32%
Mar-10	-	208,711	19%
Mar-11	-	112,968	10%
Mar-12	-	53,898	5%
Mar-13	-	19,068	2%
Mar-14	-	2,751	0%
Total	1,104,617	2,751	0%

To conclude, this section has presented data on the numbers of adults developing AIDS between March 2004 and March 2014 and has contrasted these against the ART patient targets in the Operational Plan. These data will form the basis for calculations of total resource needs and total QALYs from alternative HIV-treatment scenarios.

2.5. HIV-treatment budget constraint

The total budget available for HIV-treatment is a crucial input into an assessment of the affordability of alternative treatment strategies. According to Hickey (2004) the 2004/05 national budget was the first that specifically made provisions to support the first year of the ART rollout. In this year, National Treasury allocated US$161 million to the budget of the Chief Directorate: HIV/AIDS and TB in the NDoH. These funds included allocations for direct spending by the NDoH, funds to be transferred to NGOs and funds to be transferred as conditional grants to the provincial health departments. The funds that are spent directly by the NDoH primarily include spending on public awareness and prevention programmes and condom distribution. Grants to NGOs include allocations for the South Africa AIDS Vaccine Initiative, HIV awareness campaigns and smaller amounts to provincial and national NGOs. Finally, around 55 per cent of

the Chief Directorate's budget is transferred to provincial health departments to finance a range of HIV/AIDS interventions including:

- Voluntary counselling and testing
- Community and home-based care and support
- Prevention of mother to child transmission of HIV programmes
- Step-down care
- Strengthening of provincial management
- Establishing regional training centres with academic institutions
- Post exposure prophylaxis for rape survivors
- Step down care
- Sex worker programmes

From 2004/05, this conditional grant from National Treasury to the Chief Directorate was increased markedly, with the intention that the increase would cover provincial ART programmes and national oversight of these treatment programmes within the NDoH. These conditional grants are intended to finance the incremental costs of the ART programme (ARVs, laboratory tests and additional clinical personnel) with the implication that fixed costs such as overheads and capital within HIV-treatment clinics and community health centres, ongoing HIV-related inpatient care and HIV-related TB treatment[16] would continue to be financed via general provincial health care budgets. Thus while a large proportion of HIV-treatment costs are likely to be covered through provincial health care budgets, there are no specific budgetary allocations to cover these costs.

Table 41 shows the estimated ART conditional grants and the total provincial health care budget for primary health care, hospital services and facility capital. The provincial estimates exclude a number of functional classifications that would not relate to health service delivery such as health sciences training and administration. Data on ART conditional grants are from Hickey (2004) and data on provincial health care budgets are from the National Treasury (2005). All data are Medium Term Expenditure Framework estimates except for the 2004/05 provincial estimates which are preliminary expenditure estimates.

[16] While TB treatment was initially provided through local governments, TB is becoming part of the provincial government responsibility and these costs are therefore reflected in provincial Medium Term Expenditure Framework (MTEF) allocations.

Table 41: Budget for ART and provincial health care budgets (in US$ millions and 2003/04 prices)

	2004/05	2005/06	2006/07	2007/08
ART conditional grants for provinces	39.68	77.05	124.68	121.05
Provincial health care budget by functions				
Hospitals	3,095.79	3,047.81	3,021.46	2,960.61
Primary health care	1,095.85	1,196.37	1,232.76	1,232.09
Capital	284.88	368.99	408.86	396.56
Total	4,476.52	4,613.17	4,663.08	4,589.26

It is widely argued that the key resource constraint to successfully scaling up HIV-treatment in developing countries is clinical personnel (Kober and Van Damme 2004). Data extracted from the South African public health personnel information system (PERSAL) provide estimates of the health personnel working in the public health care system in 2005 and 2006. As shown in Table 42, there were over 13,000 medical officers and 44,071 professional nurses in 2006.

Table 42: Numbers of health professionals working in the public health care system

	Medical doctors	Medical specialists	Total medical officers	Professional nurses
2005	8,747	3,499	12,246	43,660
2006	9,527	3,695	13,222	44,071

Source: PERSAL; http://www.hst.org.za/healthstats/index.php

Total and clinical personnel resource estimates of alternative HIV-treatment interventions will be compared against these budget constraints in later parts of this chapter.

3 Population-level costs and QALYs

This section describes the total costs (million US$) and total QALYs (million) in the base case and generalized scenarios in each HIV-treatment intervention. Projections are estimated under the assumption that 100 per cent of patients in need of ART (developing AIDS) enter the models during each cycle of the planning period. Outcomes are expressed as QALYs and results are

discounted at an annual rate of three per cent. Because calculations using life years and zero discount rates do not change the results obtained in different social choice rules, these will not be presented.

3.1. Total costs and QALYs in the base case and generalized scenarios

Figure 23 and Figure 24 present the cumulative total costs and total QALYs if each of the three HIV-treatment interventions under consideration were implemented to all in need between 2004 and 2014. By the end of March 2014, cumulative total costs in the first and second-line ART scenario would be around US$12.5 billion and total outcomes would be 12 million QALYs. First-line ART would have a cumulative cost of around US$11 billion for 10.5 million QALYs while No-ART would cost just over US$7.6 billion for 5 million QALYs. In the generalized scenario, first and second-line ART would cost around US$11 billion for 11 million QALYs, first-line ART would cost around US$9.6 billion for 9.6 million QALYs and No-ART would cost around US$5 billion for 4 million QALYs. Clinical personnel and patient specific resources (the approximate incremental costs of HIV-treatment) would amount to US$8.5 billion (base case first and second-line ART), US$7 billion (base case first-line ART), US$7.7 billion (generalized first and second-line ART) and US$6.4 billion (generalized first-line ART).

Figure 23: Total costs (US$ million) and total QALYs (million) in base case scenario

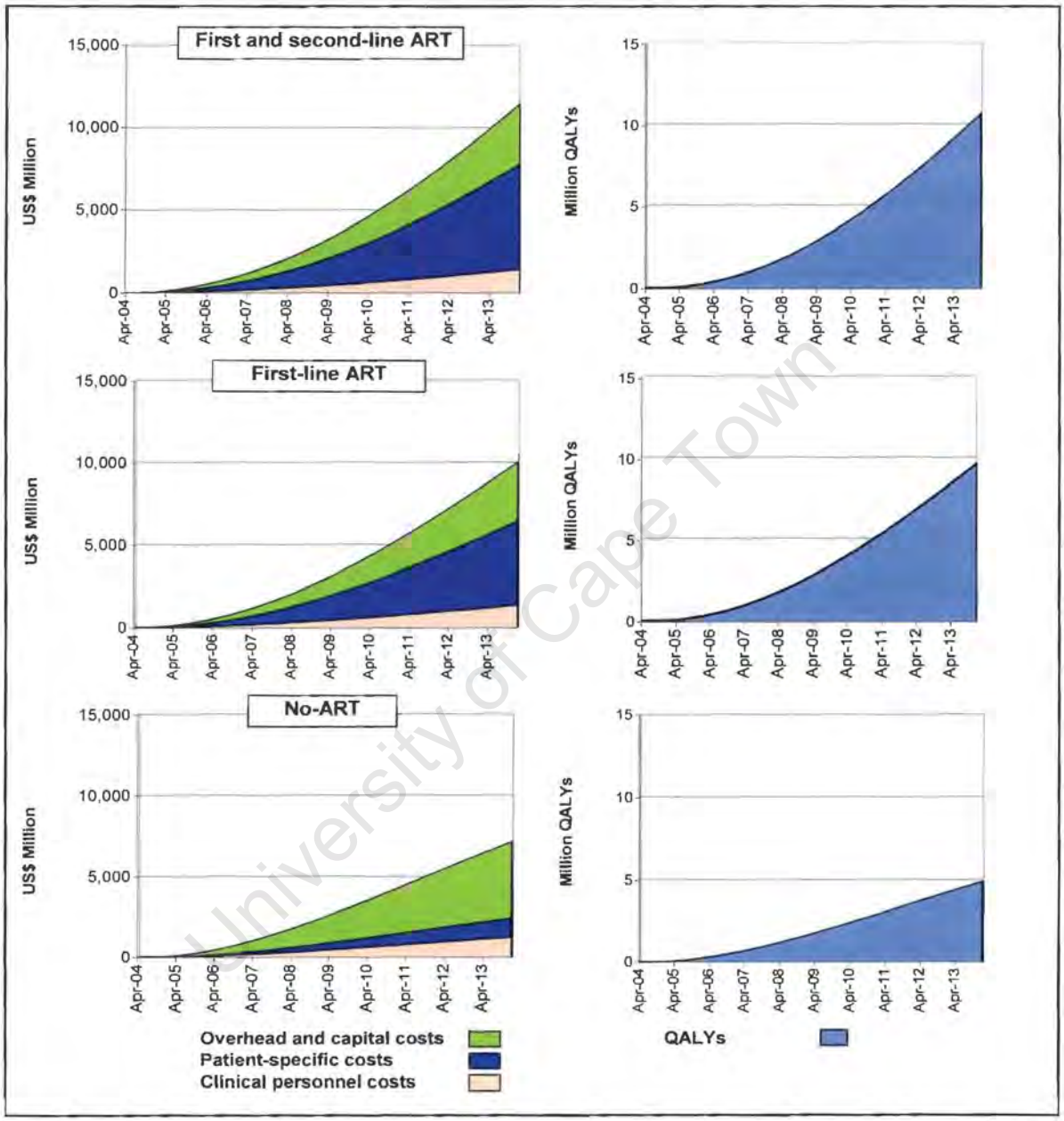
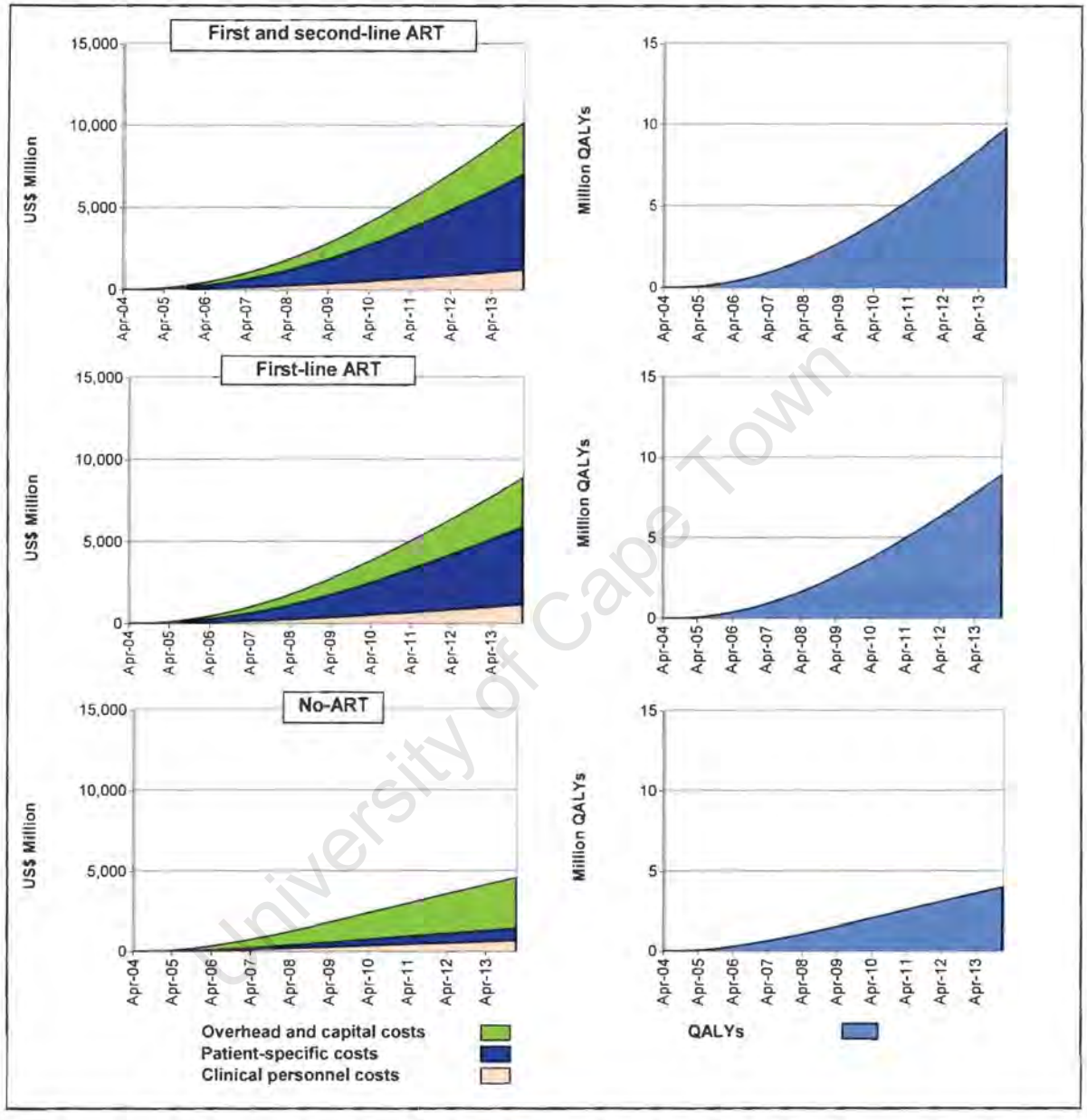


Figure 24: Total costs (US$ million) and total QALYs (million) in generalized scenario



3.2. Total clinical personnel required to provide base case and generalized HIV-treatment programmes

Given the importance of clinical personnel in the scale-up process, the number of medical officers and professional nurses required under each scenario has also been calculated. To do this, the first

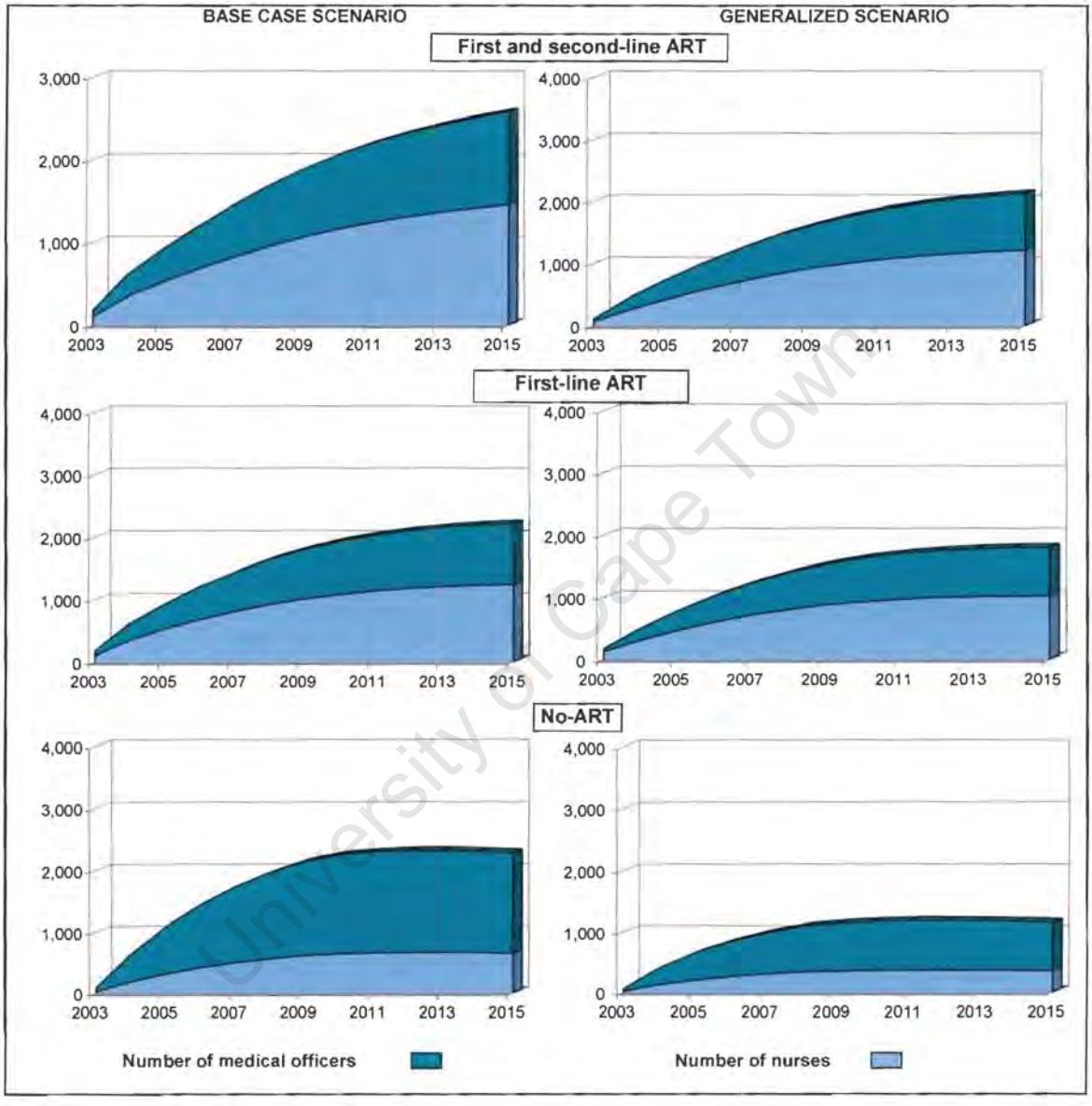
step was to group facilities into hospitals or clinics and community health centres. The second step was to calculate the total expenditure and proportion of expenditure on medical officers and nurses in these groups. The third step was to calculate the clinical staff lifetime cost (from the patient-level Markov models) incurred at hospitals and clinics or community health centres. This cost was then apportioned to nurses or medical officers based on the proportional cost of these staff in hospitals or clinics, as calculated in step two. By adding the proportion spent on nurses at clinics to the proportion spent on nurses at hospitals, it was possible to calculate the total lifetime cost spent on nurses. A similar exercise was followed for doctors. Finally, it was possible to calculate the percentage of the clinical staff lifetime cost that was spent on medical officers or nurses. These calculations are shown in Table 43.

Table 43: Breakdown of patient-level clinical staff costs (US$) in base case and generalized scenarios

	Lifetime cost - clinics	Lifetime cost - hospitals	Total lifetime cost	% of cost
Baseline scenarios				
No-ART				
Medical officer costs	63	201	264	51%
Nurse costs	155	97	252	49%
Total	218	298	516	100%
First-line ART				
Medical officer costs	525	157	682	73%
Nurse costs	173	76	249	27%
Total	698	233	931	100%
First and second-line ART				
Medical officer costs	774	186	960	74%
Nurse costs	255	90	345	26%
Total	1,029	276	1,305	100%
Generalized scenarios				
No-ART				
Medical officer costs	27	114	141	54%
Nurse costs	66	55	121	46%
Total	93	169	262	100%
First-line ART				
Medical officer costs	438	93	531	74%
Nurse costs	144	45	189	26%
Total	582	138	720	100%
First and second-line ART				
Medical officer costs	608	105	713	74%
Nurse costs	200	51	251	26%
Total	808	156	964	100%

In the social choice models, these proportions were used to calculate the total cost relating to nurses and doctors by proportioning the clinical personnel costs (see Figure 23 and Figure 24) as indicated in Table 44. By dividing the total medical officer or nurse cost by the average annual cost of employment of each staff category, it was possible to estimate the number of medical officers and nurses required to manage each HIV treatment intervention and scenario during each year of the projection. These results are shown in Figure 25. ART treatment options are much more medical officer intensive than the No-ART options. By the end of March 2014, 1,625 medical officers and 1,262 nurses would be required under base case first and second-line ART while 737 medical officers and 1,858 nurses would be required by base case No-ART. In the generalized scenario, 1,400 medical officers and 1,070 nurses, or 382 medical officers and 868 nurses would be required in the generalized first and second-line ART or No-ART options respectively.

Figure 25: Number of medical officers and professional nurses required to manage HIV-treatment programmes each year



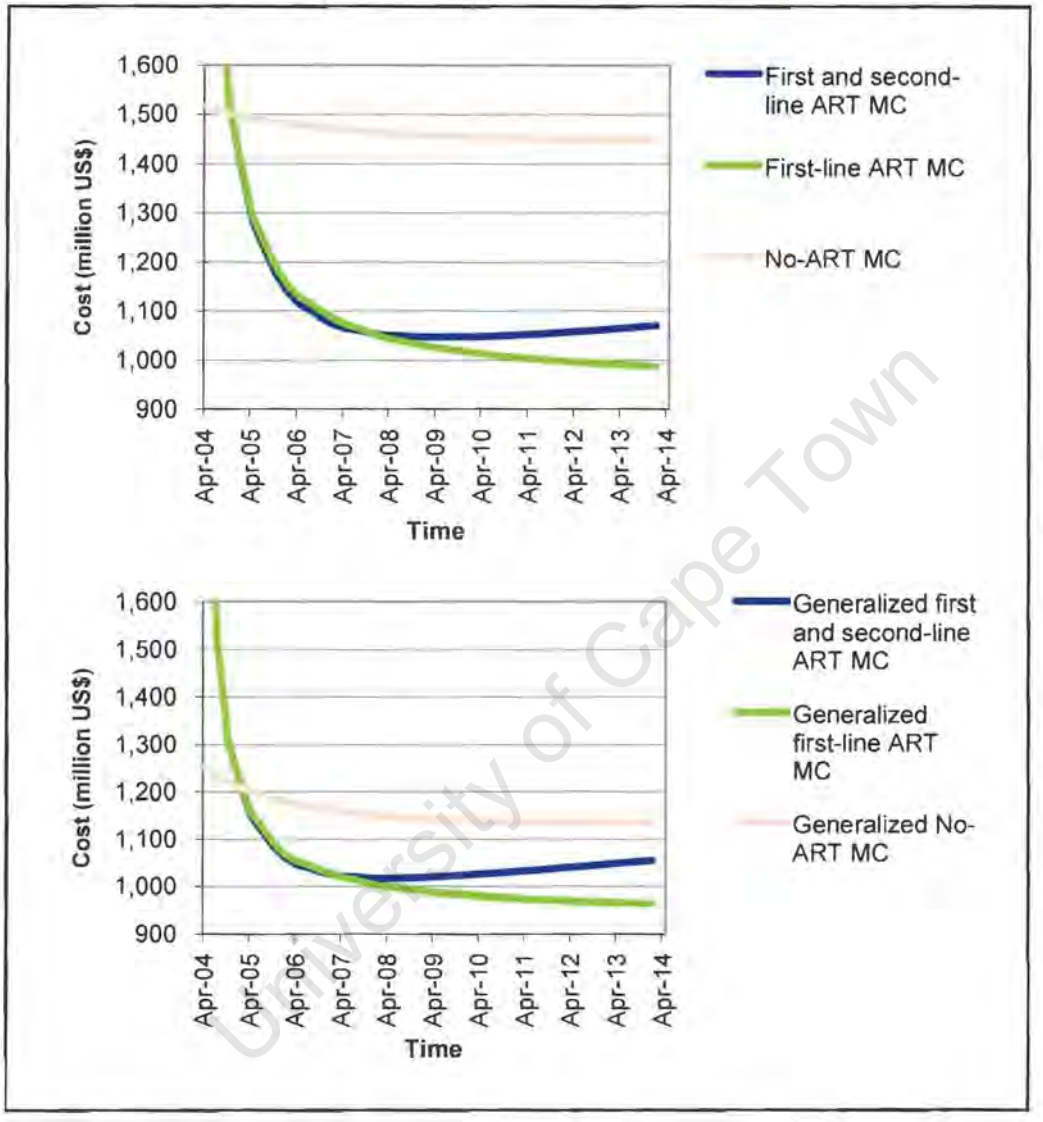
3.3. Validating the decision-making time frame

There are two ways in which the time horizon has an influence on the modelling approach. In patient-level models, the modelling time horizon has been shown to have an impact on lifetime costs, outcomes and the level of the ICER (see section 2.2). At the population level, a decision-

making time horizon might also be influential because countries need to consider whether they can sustain HIV-treatment programmes once these have been initiated. While a ten-year time-frame has been adopted for patients *entering* the projection, this time frame will not capture the full costs of treatment because the majority would still be alive and remaining in care (see Table 5 for example). The further one looks into the future, the higher the potential resource-level required owing to increasing numbers of patients being enrolled and maintained in care, but the more likely that projections become irrelevant owing to changes in technology (e.g. new medicines or therapeutic vaccines).

Economic thinking suggests that the chosen treatment programme should have the lowest marginal cost (change in cost divided by the change in QALYs). However, if the marginal cost differs over the time frame, then different HIV-treatment programmes might seem more cost-effective at different points in time. To test whether this might be the case, the marginal cost of each programme delivered to 100 per cent of patients in need was calculated for each cycle (three-month period) of the projection. This involved dividing the change in total costs from one cycle to another by the change in total outcomes in the same period. The results are presented in Figure 26. In the earliest time periods, the marginal costs of the first-line ART and first and second-line ART options are similar. However, as time passes and more patients transition to second-line, the marginal cost of the two first and second-line options becomes higher than the two first-line options. The implication is that very short decision-making time-frames might be misleading. This analysis has therefore validated the choice of a ten-year decision-making time frame that has been proposed.

Figure 26: Evolution of the marginal cost (discounted at 3 per cent) over time in base case and generalized scenarios



3.4. Social choice rules

This section compares the QALYs gained and percentage of need that can be reached under health maximisation, the decent minimum or equal health social choice rules in the base case and generalized scenario. Base case and generalized scenarios are not assumed to be in competition with each other – in other words the generalized scenario interventions are only in competition with each other, but not with the base case scenarios. This is because it would be unfair to design

policy that specifically places some patients on a generalized option and others on a base case option given the different quality of care.

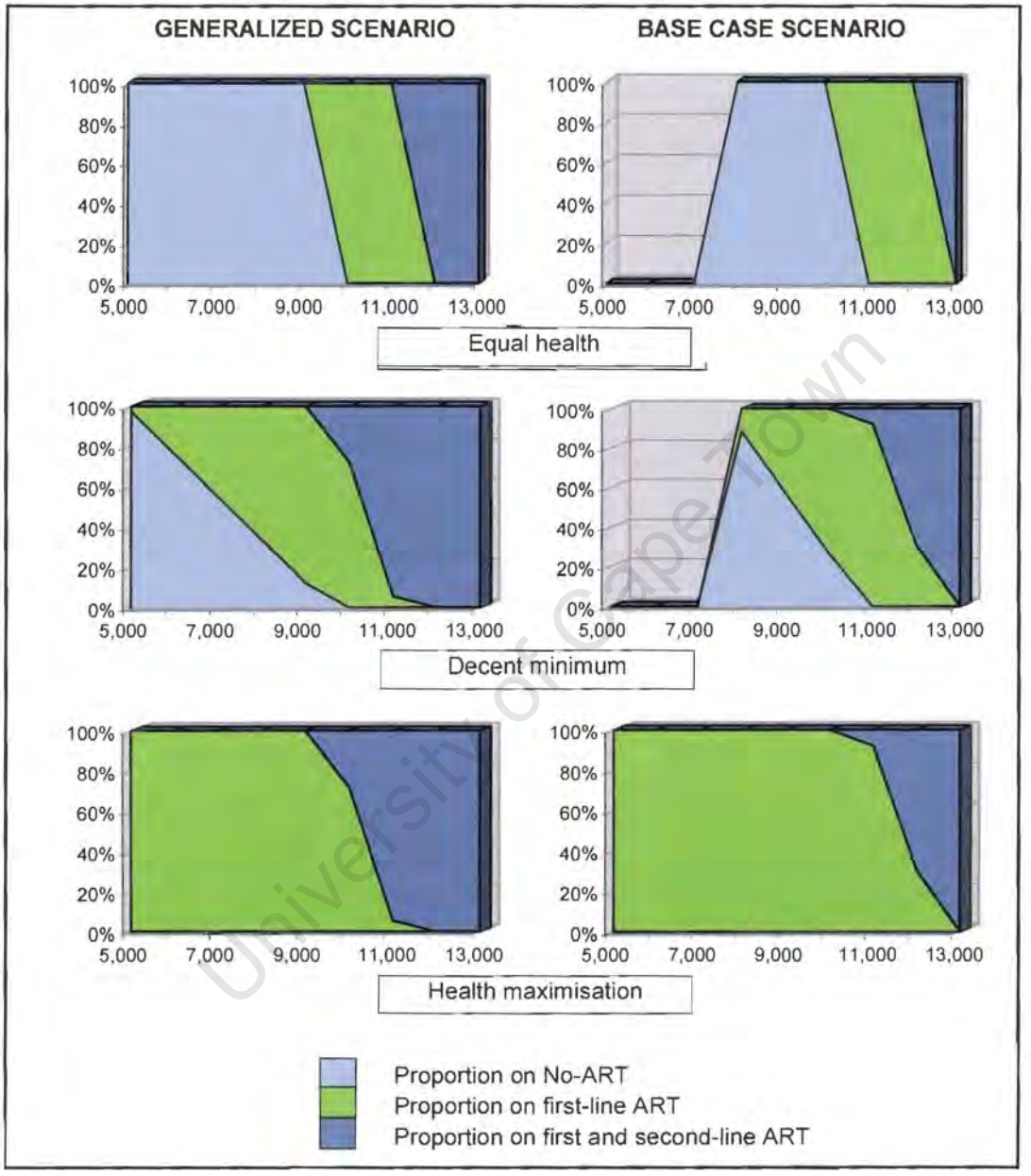
Social choice rule calculations are based on the total discounted costs and QALYs associated with each HIV-treatment intervention between March 2004 and March 2014. The health maximisation social choice rule chooses the scenario or combination of scenarios that produces maximum QALYs at a given budget constraint irrespective of the proportion of need that is met, the "decent minimum" maximises outcomes subject to all patients receiving some form of treatment, while equal health places all patients on the same treatment intervention. Table 44 shows the mix of treatment programmes and total QALYs that are produced at different budget constraints. For the two more equitable social choice rules (equal health and the decent minimum) the QALY loss associated with being equitable is calculated by comparing each of these against health maximisation. The cost of health maximisation is expressed in terms of the percentage of need that is unmet. Results are presented for budgets ranging from US$5 billion to US$13 billion. The former is slightly higher than the cheapest treatment strategy (generalized No-ART) while the latter is slightly higher than the most expensive option (base case first and second-line ART).

In the base case scenario, neither of the more equitable rules places any patients on treatment until the budget is approximately US$8 billion, but the health maximisation rule is able to place 64 per cent of patients on first-line ART for a QALY gain of 6.8 million. At lower budgets, both equitable strategies place patients on No-ART, which is never a policy choice under health maximisation. Results from social choice rules converge as the budget increases. Once US$11 billion is available, 100 per cent of need is met under all social choice rules, and the decent minimum and health maximisation produce the same results. Similar patterns are found in the generalized scenario, although note that at lower budgets the generalized scenario is both more equitable and more technically efficient than the base case scenario. The proportion of patients on different HIV-treatment programmes as the budget increases is depicted in Figure 27.

Table 44: Total QALYs and mix of programmes at different budget constraints in the social choice rules

Budget (US$ million)	EQUAL HEALTH					DECENT MINIMUM					HEALTH MAXIMISATION				
	Proportion on:			Total QALYs (millions)	QALY cost (millions)	*Proportion on:*			Total QALYs (millions)	QALY cost (millions)	*Proportion on:*			Total QALYs (millions)	Unmet need
	No-ART	First-line ART	First & second-line ART			No-ART	First-line ART	First & second-line ART			No-ART	First-line ART	First & second-line ART		
	Base case scenario					**Base case scenario**					**Base case scenario**				
5,000	-				-4.8	-	-	-	-	-4.8	-	46%	-	4.8	54%
6,000	-				-5.8	-	-	-	-	-5.8	-	55%	-	5.8	45%
7,000	-				-6.8	-	-	-	-	-6.8	-	64%	-	6.8	36%
8,000	100%			5.2	-2.5	88%	12%	-	5.8	-1.9	-	74%	-	7.8	26%
9,000	100%			5.2	-3.5	58%	42%	-	7.5	-1.2	-	83%	-	8.7	17%
10,000	100%			5.2	-4.5	27%	73%	-	9.1	-0.6	-	92%	-	9.7	8%
11,000		100%		10.5	-0.1	0%	92%	8%	10.6	0.0	-	92%	8%	10.6	0%
12,000		100%		10.5	-0.8	0%	31%	69%	11.3	-0.0	-	31%	69%	11.3	0%
13,000			100%	11.7	-	-	0%	100%	11.7	-	-	0%	100%	11.7	-
	Generalized scenario					**Generalized scenario**					**Generalized scenario**				
5,000	100%			4.2	-0.8	97%	3%	-	4.4	-0.6	-	52%	-	5.0	48%
6,000	100%			4.2	-1.8	76%	24%	-	5.5	-0.5	-	63%	-	6.0	37%
7,000	100%			4.2	-2.8	55%	45%	-	6.7	-0.4	-	73%	-	7.0	27%
8,000	100%			4.2	-3.8	34%	66%	-	7.8	-0.2	-	84%	-	8.1	16%
9,000	100%			4.2	-4.8	12%	88%	-	9.0	-0.1	-	94%	-	9.1	6%
10,000		100%		9.6	-0.3	0%	72%	28%	9.9	0.0	-	72%	28%	9.9	0%
11,000		100%		9.6	-1.0	0%	6%	94%	10.6	-	-	6%	94%	10.6	0%
12,000			100%	10.7	-	-	-	100%	10.7	-	-	-	100%	10.7	-
13,000			100%	10.7	-	-	-	100%	10.7	-	-	-	100%	10.7	-

Figure 27: Proportions of patients on alternative HIV-treatment programmes at different budget constraints



There are a number of implications of this analysis:

- At lower budgets, in particular in the base case scenario, the equitable social choice rules present a potentially unacceptable reduction in the health gains of HIV-positive people

because they allow no access to antiretroviral treatment. However, the benefit of these solutions is that all HIV-positive people will receive treatment for opportunistic infections and palliative care.

- All social choice rules favour the use of first-line ART but once again it might be unacceptable to limit access to an additional effective regimen. This would be particularly controversial for patients who rapidly fail the first-line regimen owing to side-effects.
- The generalized scenario is both more equitable and more technically efficient than the base case scenario at lower budget constraints.

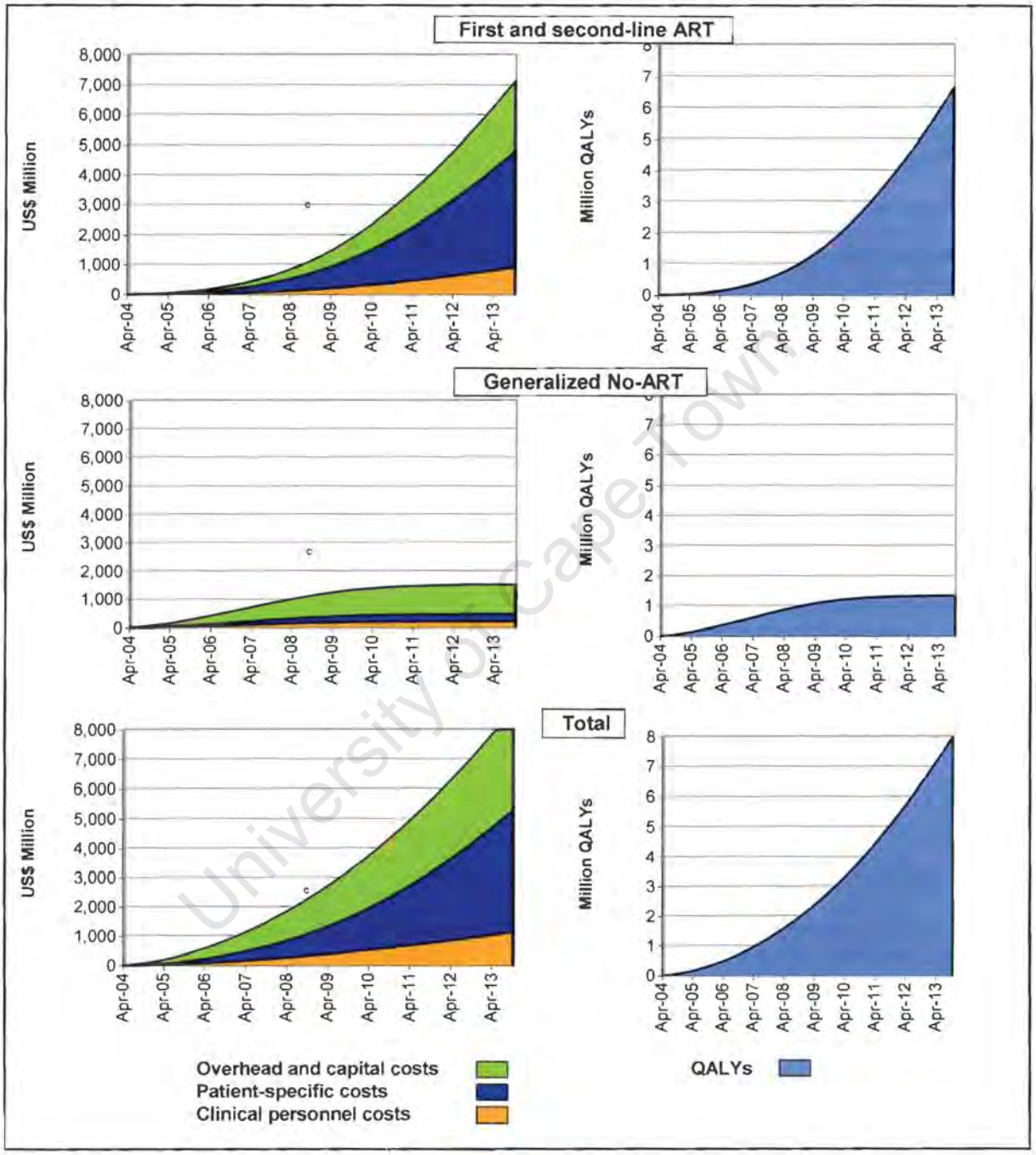
4 Analysis of the Operational Plan

This section analyses the government's planned scale-up of ART from an equity, efficiency and affordability perspective. In analysing the Operational Plan, it has been assumed that the public sector rollout of ART proceeds according to the planned treatment targets and that the remaining patients in need receive generalized No-ART care. Base case first and second-line ART has been assumed because the Operational Plan describes a relatively resource intensive model of care and generalized No-ART has been assumed given that data show that patients receive a less resource intensive form or this care than that assumed in base case No-ART (Cleary, Chitha et al. 2005).

4.1. Total costs and QALYs in the Operational Plan

Figure 28 presents the discounted cumulative costs and QALYs of these two programmes. By the end of March 2014, the ART programme would cost US$7,551 million at an output of 7 million QALYs, generalized No-ART would cost US$1,524 million for 1.3 million QALYs and the total cost of this comprehensive treatment plan would be US$9,075 million for 8.4 million QALYs.

Figure 28: Estimated costs and QALYs of implementing the Operational Plan



The affordability of the programme in the medium term can be assessed by comparing the costs between March 2004 and March 2008 against the budget allocated over this period. The

incremental costs of first and second-line ART (clinical personnel and patient-specific) are compared against the ART conditional grants. This is because, as explained above, only these components are financed via this mechanism. The full costs of generalized No-ART and the overhead and capital costs of ART are compared against provincial health care budgets (excluding conditional grants). Costs and budgets are expressed in 2003/04 prices and discounted at an annual rate of three per cent. As shown in Table 45, the incremental costs of the ART programme would initially be covered by the conditional grants, but by 2006/07, these allocations become inadequate. The costs of generalized No-ART and the ART overhead and capital costs amount to 9 per cent of the provincial health care budget by 2007/08. If the growth of these budgets keeps pace with real budgetary growth rates seen in recent years[17], it is possible that HIV-treatment could amount to 40 per cent of combined conditional grants and provincial health care budgets by 2014.

Table 45: Comparison of HIV-treatment costs to conditional grants and health care budgets (US$ million)

	2004/05	2005/06	2006/07	2007/08
ART conditional grants for provinces	39.68	77.05	124.68	121.05
ART annual clinical personnel cost	4.65	16.15	32.26	52.43
ART annual patient-specific cost	10.61	44.64	103.14	184.79
Total	**15.26**	**60.79**	**135.40**	**237.22**
Proportion of ART conditional grants	*38%*	*79%*	*109%*	*196%*
Provincial health care budget	4,476.52	4,613.17	4,663.08	4,589.26
ART annual overhead and capital cost	12.62	43.06	84.95	137.03
No-ART annual clinical personnel cost	15.78	35.22	42.53	42.67
No-ART annual patient-specific cost	18.24	40.93	49.64	49.96
No-ART annual overhead and capital cost	75.01	167.39	202.13	202.80
Total	**121.65**	**286.60**	**379.24**	**432.46**
Proportion of provincial health care budgets	*3%*	*6%*	*8%*	*9%*

While total resources are clearly a constraint to scaling up HIV-treatment programmes, it is also argued that clinical personnel constraints might be even more important (Kober and Van Damme

[17] The recent growth in real per capita public health care budgets follows a period of stagnation at the end of the 1990s in government health care financing (McIntyre, Gilson et al. 2006). In estimating the budget for 2013/14, it has been assumed that the recent rate of increase would be maintained.

2004). The NDoH has estimated that 725 medical officers and 2,175 professional nurses would be required to deliver comprehensive HIV-treatment between April 2005 (lagged to April 2006) and April 2008 (lagged to April 2009) (Operational Plan for Comprehensive HIV and AIDS Care 2003), each comprising 5 per cent of these cadres employed in the public sector in 2006. In comparison, Table 46 shows estimates of clinical personnel required based on the data in this thesis. These indicate that 819 medical officers and 1,075 professional nurses would be required in the base case scenario and 607 medical officers and 660 professional nurses in the generalized scenario by April 2009. Estimates of medical officers in the base case scenario are higher than those in the Operational Plan, but estimates of professional nurses are lower in both scenarios. By April 2014, base case first and second-line ART and No-ART would require 10 per cent and 2 per cent respectively of the medical officers and professional nurses working in the public sector in 2006. In comparison to the Operational Plan, the estimates from this thesis indicate that current delivery of HIV-treatment is more medical officer intensive than has been envisaged in the Operational Plan.

Table 46: Clinical personnel required between April 2006 and April 2009 to deliver base case or generalized HIV-treatment programmes

	BASE CASE SCENARIOS		GENERALIZED SCENARIOS	
	Medical officers	Professional nurses	Medical officers	Professional nurses
By April 2006				
First & second-line ART	139	108	107	82
No-ART	264	664	160	365
Total	**299**	**473**	**107**	**82**
Per cent of 2006 workforce	**2%**	**1%**	**1%**	**0%**
By April 2007				
First & second-line ART	245	190	198	151
No-ART	302	762	175	399
Total	**420**	**589**	**198**	**151**
Per cent of 2006 workforce	**3%**	**1%**	**1%**	**0%**
By April 2008				
First & second-line ART	386	300	319	243
No-ART	299	755	166	377
Total	**552**	**677**	**319**	**243**
Per cent of 2006 workforce	**4%**	**2%**	**2%**	**1%**
By April 2009				
First & second-line ART	568	441	476	362
No-ART	251	634	131	297
Total	**819**	**1,075**	**607**	**660**
Per cent of 2006 workforce	**6%**	**2%**	**5%**	**1%**

While these estimates of clinical personnel requirements might seem feasible, there are a number of factors that should be borne in mind. Firstly, according to a Human Sciences Research Council study, there were 4,222 and 32,734 vacant medical officer and nurse[18] posts in the public sector in 2002 (Hall and Erasmus 2005). Across all categories of health professionals, 27.2 per cent of posts were vacant (Day and Gray 2005). In addition, HIV/AIDS is expected to have an impact on the supply of health professionals, firstly because in South Africa in 2002, 16.3 per cent of health workers were HIV-infected (Shisana, Hall et al. 2002) and secondly because perceptions of risk due to HIV exposure could cause resignations or reduce incentives for young people to choose a career in the health sector (Over 2004). International migration of health professionals from South Africa is also increasingly of issue. Data reported in Schneider, Blaauw et al. (2006) indicate that more than 500 medical officers and 1,000 nurses from South Africa register annually with the United Kingdom General Medical Council. Aggregate levels of health professionals at the national level also tend to hide the disparities within South Africa; the internal brain drain from the public to the private sector and from rural to urban areas is particularly problematic (McIntyre, Gilson et al. 2006; Schneider, Blaauw et al. 2006).

Besides the various uncertainties inherent in predicting the need for health care workers in the public sector in general, it should also be noted that the estimates of clinical personnel requirements in Table 46 are heavily influenced by data from the Khayelitsha programme which had 1.3 medical officers for each professional nurse while the TC Newman outpatient department ART programme has 3.2 medical officers to each professional nurse, making that programme markedly more medical officer intensive. While norms of staffing in ART sites are still under development in the Western Cape, a potential normative staffing profile would be 0.7 medical officers and 2.1 professional nurses for every 1,000 patients (Dr Andrew Boulle, Public Health Specialist, School of Public Health and Family Medicine, University of Cape Town, personal communication). Under this staffing norm, similar numbers of medical officers would be required but far more professional nurses than presented in Table 46. Given the wide range of potential staffing profiles that exist in HIV-treatment services, the data presented in Table 46 should be treated with caution.

[18] Including all nurse posts, not only professional nurses.

4.2. Social choice rules and the Operational Plan

Using mathematical programming, this section calculates the percentage of need that can be met and QALYs that can be gained under different social choice rules between 2004 and 2014 at a budget constraint of US$9,075 million, which is the estimated total cost of meeting the Operational Plan targets for base case first and second-line ART and generalized No-ART, as shown in Figure 29. Although it was argued that South Africa is likely to be implementing base case ART together with generalized No-ART, in the social choice rule calculations it is assumed that the country can either implement a base case or a generalized approach but not a mix of the two because of considerations of fairness between patients. In Table 47, the change in QALYs and/or change in coverage between the base case and generalized scenarios for each of the social choice rules versus the Operational Plan is shown. In the first part of the table, first-line ART is not included as a policy option, but this constraint is relaxed in the bottom half. Where first-line ART is not a policy choice, the social choice rule results will obviously differ from those presented in earlier sections.

In the Operational Plan scenario, 17 per cent of patients receive generalized No-ART and the remainder receive first and second-line ART. Recall that in the Operational Plan scenario, the scaling-up of access to first and second-line ART happens incrementally over the years between 2004 and 2010, with 100 per cent of *new need* met through ART from this time until the end of the projection in 2014. This scenario gains 8.4 million QALYs.

In the base case scenario (with costs and outcomes derived from the Khayelitsha pilots) health maximisation increases QALYs in comparison to the Operational Plan (by 0.11 million), but a lower level of need would be met. While the decent minimum meets 100 per cent of need, far fewer patients receive first and second-line ART and the cost in terms of foregone QALYs in comparison to the Operational Plan is 1.23 million. If generalized treatment scenarios are assumed, both health maximisation and the decent minimum gain QALYs in comparison to the Operational Plan, although fewer patients receive first and second-line ART[19].

[19] The Operational Plan meets a high proportion of need through first and second-line ART but gains fewer QALYs than the base case health maximization approach, for example. This apparent anomaly relates to the timing of treatment, which is exacerbated by discounting results. In the Operational Plan, treatment is scaled-up over time, while all other scenarios assume an instant scale-up from the first year of the projection. What this means is that while ultimately 83 per cent of QALYs are gained through ART in the

If first-line ART is included as a policy choice (as shown in the second half of the table) no social choice rules choose first and second-line ART at a budget of US$9 billion. Health maximisation favours first-line ART exclusively while the decent minimum proposes a combination of No-ART and first-line ART. If the set of generalized treatment strategies is favoured, it is possible to gain as many as 0.7 million QALYs through the decent minimum in comparison to the Operational Plan. The highest QALY gain comes through maximising health under the generalized scenario but five per cent of need is unmet.

The equal health social choice rule proposes that all patients receive No-ART at a budget constraint of US$9 billion. Given the emphasis that civil society is placing on access to ART, equal health is likely to be considered to be too costly in terms of its impact on individual health gains. In addition, it is likely that No-ART care provides too little motivation for patients to overcome access barriers such as stigma. Thus it is ironic that the most "equitable" solution at a budget of US$9 billion could be health maximisation which reaches 95 per cent of patients with generalized first-line ART.

Operational Plan, these are accrued later in time and are discounted at a higher rate than is the case under the other scenarios.

Table 47: Comparison of the Operational Plan to social choice rules

	Per cent No-ART	Per cent first-line ART	Per cent first and second-line ART	Total QALYs (millions)	Unmet need (%)	QALY gain or loss (millions)
Operational Plan	17%	-	83%	8.4	0%	-
Base case scenario						
Health maximisation	-	-	73%	8.5	27%	0.11
Decent minimum	70%	-	30%	7.1	0%	-1.23
Equal health	100%			5.2	0%	-3.15
Generalized scenario						
Health maximisation	-	-	82%	8.7	18%	0.36
Decent minimum	32%	-	68%	8.6	0%	0.21
Equal health	100%			4.2	0%	-4.13
IF FIRST-LINE ART WERE A POLICY OPTION						
Base case scenario						
Health maximisation	-	84%	-	8.8	16%	0.42
Decent minimum	55%	45%	-	7.6	0%	-0.77
Equal health	100%			5.2		-3.15
Generalized scenario						
Health maximisation	-	95%	-	9.1	5%	0.77
Decent minimum	11%	89%	-	9.1	0%	0.70
Equal health	100%			4.2		-4.13

The implication of this analysis is that if the country wishes to place a higher proportion of patients on ART and gain QALYs, thereby improving both equity and efficiency, consideration should be given to implementing less resource intensive models of ART delivery and/or limiting ART regimens to first-line only. This strategy is in line with the WHO's policy of a public health approach to scaling up ART in resource limited settings (Gilks, Crowley et al. 2006) and is similar to the approach adopted in Malawi's scale-up of ART (Harries, Schouten et al. 2006). While many would favour this strategy as it leads to higher patient coverage, others argue against compromising the quality of ART care (recall that the generalized scenarios assume that patients have no access to hospital outpatient care, to tertiary inpatient care and are followed-up less frequently than has been the case in Khayelitsha). Proponents of maintaining quality of care would probably see the current South African strategy in a favourable light.

The assumption that the patients in need of care who are not accessing ART receive generalized No-ART care deserves additional scrutiny. The provision of even generalized No-ART care requires diagnosis and a chronic disease approach to patient management to ensure that appropriate prophylactic medication and treatment of acute infections is provided. Given the current emphasis on scaling up ART in the public health system, the development of a chronic

disease model of No-ART care has been neglected. It is also likely that some patients with HIV are unaware of their HIV-status - some might remain asymptomatic while their immune system becomes compromised and die suddenly from an acute infection. Others might be symptomatic but might perceive the barriers to entry into care to be too high. Thus while the current Operational Plan is not explicitly focusing solely on technical efficiency (health maximisation) it implicitly takes this approach by providing first and second-line ART to a proportion of patients in need and potentially less than generalized No-ART to the remainder.

5 Summary

This chapter has extended the results from patient-level models to calculate the total costs and QALYs associated with alternative treatment interventions and those required to implement the Operational Plan's treatment strategies and patient targets.

The generalized scenario performs better than base case in terms of the proportion of need that can be met and the amount of QALYs that can be gained (technical efficiency) unless the highest budgets are available. Within scenarios, first-line ART allows fairly high patient-level benefits to be achieved and at the same time secures a wider sharing of the benefits of treatment than first and second-line ART. However, implementing first-line ART as a policy option might not be acceptable to society given that this would imply a downgrading of the individual HIV-treatment benefits that are currently proposed through the Operational Plan. This might be acceptable in the short-run if higher numbers of patients in need were reached and if a commitment were made to progressively extend access to first and second-line ART over the medium term.

Although the approach taken in this chapter is consequentialist, these findings could be considered to be inputs into a fair process concerning HIV-treatment decision-making. The following chapter in this dissertation revisits the distributive framework that has been proposed in this thesis, including evidence from patient-level and population-level modelling, in order to sketch the outlines of a procedurally just framework for HIV-treatment priority setting.

Chapter 8: Discussion - the distributive framework re-examined

1 Introduction

This thesis has advocated an approach to HIV-treatment priority setting that simultaneously considers the costs of scaling up, efficiency, equity and the equity/efficiency trade-off. This has been accomplished by extrapolating extensively justified patient-level data to the population-level. Once armed with the total costs and total health gains of alternative interventions, the social choice rule, which acts as the set of principles or values on which distributions are based, is used to determine the health gain, proportion of need that is met and the proportion of patients receiving alternative treatment strategies at a given HIV-treatment budget.

Because reasonable people are likely to disagree on the choice of the value base and the extent to which the health care budget should be allocated to HIV-treatment, this chapter considers how decision-making could be undertaken within what Daniels (2004) terms a "fair process". The elements of this fair process will be discussed in section 2 of this chapter. Briefly, however, if there are no independent criteria that suggest that (a) health maximisation (for example) is the preferred social choice rule and (b) US$12 billion is the appropriate budget, then procedural justice can help to resolve disagreements in a way that achieves legitimacy (Daniels 2004; Wailoo and Anand 2005).

Although the participants at a "Consultation on ethics and equitable access to treatment and care for HIV/AIDS" have argued that universal[20] access "is the only truly ethical outcome" (WHO and UNAIDS 2004 p. 8), one has to bear in mind that once some level of equal access has been achieved, ethical debate could focus, for example, on whether a more resource intensive approach should be taken, whether third-line or salvage therapy should be offered and whether ART should be extended to patients with CD4>200 cells/μl. Thus while some of the technical calculations presented in this thesis will become outdated as new treatment technologies become available, the framework will continue to have relevance when these new technologies are considered.

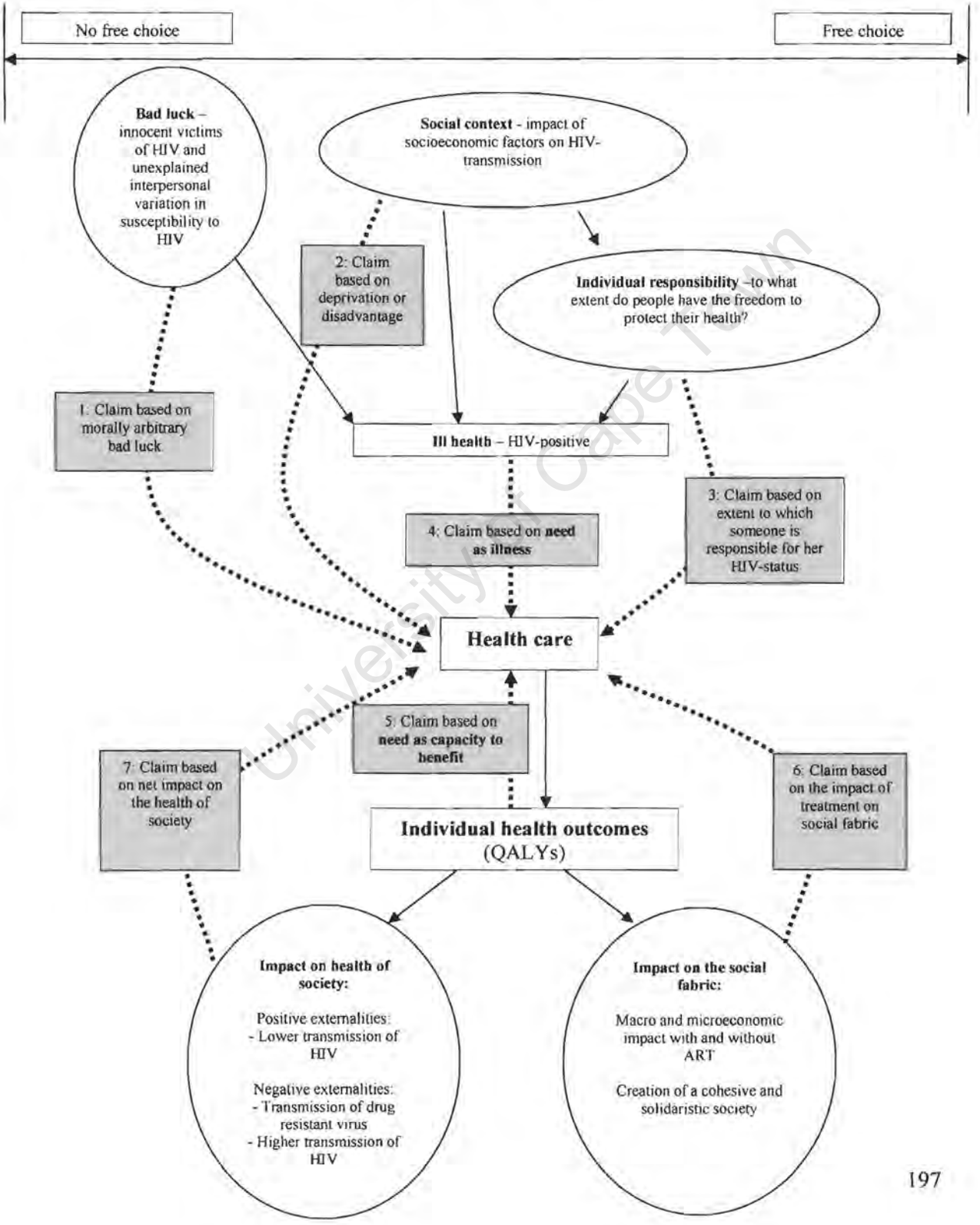
[20] It is acknowledged that the term "universal access" is misused in this context. Presumably, this means equal access for all patients with a defined level of need.

The following sections in this chapter will revisit a number of the debates that have been introduced in earlier chapters and in particular will cover the following three questions:

- Whether a resource-intensive (base case) or more minimalist (generalized) approach to treatment should be taken
- Whether No-ART, first-line only or first and second-line should be offered
- Whether and/or how medical eligibility criteria for ART should be revised

While the questions of what health care and hence what health benefit to distribute to HIV-positive adults are the central questions of this thesis, it is also important to broaden the discussion to consider the personal characteristics of HIV-positive people who are in need of treatment and the broader benefits that could accrue to society. The discussion will follow the structure of Figure 29 which has been reproduced overleaf from Chapter 3. The views of participants in a fair process regarding the importance of the claims of HIV-positive people on the good (the question of to whom the good should be distributed) will determine the existence of and strength of calls for increased HIV-treatment resources and might also relate to the value base that is chosen. This in turn will determine the impact that HIV-treatment will have on the health of society (a balancing of the positive and negative externalities associated with treatment) and the social fabric (claims 6 and 7). Depending on the strength of these claims, a decision about what health or health care to distribute could be revised. In this manner, claims are considered, reconsidered and integrated into the final decision.

Figure 29: A framework for considering to whom the good should be distributed. Solid arrows show the causes of illness and consequences of health care. Dotted arrows show claims on the good. Reproduced from Chapter 3.



2 Criteria for procedurally just decisions

In an earlier chapter, it was argued that procedural justice focuses on the fairness of the process through which a distribution is achieved (Rawls 1971). It is argued that the distribution of the good to HIV-positive South Africans is a problem of pure procedural justice. While there might be no independent criterion or agreement on the social choice rule, if a fair process is followed, the resulting outcomes might be fair.

Economists traditionally argue that procedures are only valuable for their instrumental role in promoting better outcomes. According to Wailoo and Anand (2005), this reflects the notion of perfect procedural justice. For example, in the fair division of a cake, the procedure of cutting the cake is valuable if it ensures the outcome of a close to equal division. On the other hand, in pure procedural justice, while procedures continue to have instrumental value, they can also have inherent or intrinsic value. The inherent value of procedures is also suggested by those who advocate for a "communitarian claims" approach where it is argued that the community finds value in the process of being involved in decision-making (Mooney and Jan 1997; Mooney 1998; Mooney, Jan et al. 2002; Mooney 2005).

Daniels (2004) argues that the central requirements of fair process are:

- Publicity: the process must be transparent and involve publicly available rationales for the priorities that are set. This has the added benefit of encouraging good governance.
- Relevance: stakeholders who are affected by the decisions should agree that they rest on reasons, principles and evidence that they view as relevant to making fair decisions about priorities. This has the added benefit of assuring stakeholders that their voice has been heard.
- Revisability and appeals: decisions can be revisited and revised in light of new evidence and arguments. This appeals process provides protection to those who have legitimate reasons for being an exception to adopted policies.
- Enforcement or regulation: a mechanism is in place to ensure that the previous three conditions are met.

The goal of securing adequate publicity is ambitious (Daniels 2004). In most countries, resource allocation and priority setting decisions are taken behind closed doors. If the condition of publicity is to be met, the full rationales, resultant recommendations and any complaints or disputes related to these decisions would need to be in the public domain in a format that was comprehensible to a lay audience. There are a number of intrinsic and instrumental values to publicity. Firstly, it gives legitimacy to decisions that are taken (Wailoo and Anand 2005) and gives the public greater confidence in the process and the outcomes (Daniels 2004). Secondly, people value knowing why decisions that affect their lives have been taken in the way they have been taken (Litva, Coast et al. 2002; Daniels 2004; Wailoo and Anand 2005). Thirdly, a form of precedence emerges which assists in consistency over time. This has intrinsic value from an equity perspective as it ensures that like cases are treated in a like manner (Wailoo and Anand 2005). Consistency also has instrumental value since the setting of precedence assists in future decision-making, thereby improving the quality of decision-making over time.

The relevance condition, on the other hand, suggests that there should be restrictions on the kinds of rationales that are permitted to serve as a basis for decision-making and that these restrictions should be agreed by key stakeholders. These "reasonable rationales" could be set within a communitarian claims process, where the community sets the "structures, principles or rules on which to base the social welfare function…and hence the basis for priority setting in health care" (Mooney 1998 p. 1173). In other words, the community could be consulted about the social choice rule as well as the personal characteristics of people that could serve to justify additional claims or limitations of claims on the good (see claims 1 to 3 in Figure 29). The latter deals with issues of vertical equity - the unequal but equitable treatment of unequals (Mooney 2003). It might be decided that HIV-positive people have a disproportionate claim on health care resources given the social context of sufferers and the impact on the social fabric and the health of society that treatment affords. Communitarian claims is not about replacing the bureaucrat – instead the community would play a role in establishing the value base of the health care system and the bureaucrat would have a role in ensuring that the system is managed according to these values (Black and Mooney 2002). Research has shown that the public finds intrinsic value in having a voice in decisions; allowing the community to set reasonable rationales provides one avenue for this voice (Litva, Coast et al. 2002; Wiseman, Mooney et al. 2003; Wailoo and Anand 2005).

The third requirement for fair process is revisability and due process. A distributive decision will tend to be more acceptable if there are mechanisms which allow decisions to be challenged and

reversed if required (Wailoo and Anand 2005). This requirement also allows for the improvement and revisiting of policy over time as resource constraints and technologies change. Revisability and due process is strongly related to the publicity condition because the transparency of the original decision both in terms of rationales and ultimate recommendations facilitates the identification of mistakes. It also provides an avenue for parties affected by decisions to appeal. Even if unsuccessful, the appeals process gives voice to stakeholders who might not have been included in the original decision and adds to the body of precedence thereby improving the quality of future decision making (Daniels 2004).

The final requirement for fair process is regulation and enforcement. A mechanism needs to be created to ensure that the fair process complies with the publicity, relevant reasons and revisability and appeals requirements that are outlined above (Daniels 2004).

Steps towards the implementation of fair process would include clarifying institutional levels of decision-making, developing structures to address decisions at each level, training to develop competence in fair process, learning from experience, improving the process through training and research and developing mechanisms for enforcement (Daniels 2004). Although the development of fair process should not stall the scaling up of treatment, as fair process is developed it could have additional benefits through serving as a model for other decision-making in the health care system, improving accountability and empowering communities (London 2003). Rather than framing the poor and marginalized as candidates for redistributive policy by a benevolent state (McIntyre and Gilson 2002), this approach encourages active community participation in resource allocation in health which could improve the ability of civil society to hold governments and donors to account, with both instrumental and intrinsic value.

3 To whom should the good be distributed? (claims 1-3)

The question of to whom the good should be distributed considers whether the personal characteristics of HIV-positive people should be a basis for additional claims or for limitations of claims on the good, as illustrated through claims 1-3 in Figure 29. Because HIV is a preventable sexually transmitted disease, HIV-positive people have traditionally been subject to high levels of stigma and discrimination (Rankin, Brennan et al. 2005). It is argued that stigma is a tool used by

cultures to exclude those felt to have broken existing rules; the dominant stereotype of HIV-positive people is therefore one that casts them as immoral (Furber, Hodgson et al. 2004). This inevitably leads to discussion of personal responsibility in HIV-acquisition.

According to this concept, a person's claim on the good could differ if the causes of her illness were exogenous as opposed to determined by personal risky behaviour (Edgar, Salek et al. 1998; Olsen, Richardson et al. 2003), as illustrated by the free choice continuum in Figure 29. The oval in the top left-hand corner relates to having no free choice in acquiring HIV, and therefore claim 1 on the good relates to bad lack, but it can also relate to compensation for people who acquired HIV through mother to child transmission, sexual violence, blood transfusion and accidental occupational exposure[21]. HIV-positive people who fall into this group could be argued to have additional claims on the good for reasons of desert (Macklin 2004). On the other end of the free choice continuum, it is presumably possible to have full responsibility in HIV acquisition, although it is difficult to imagine how unless someone purposefully exposes him or herself to HIV on a repeated basis. It is important to recognize that responsibility for HIV status is mediated via social context. What this means is that any claim that someone is less deserving of HIV-treatment owing to personal responsibility in acquiring HIV would have to consider the social context within which infection occurred. Thus the claim based on disadvantage would complement the claim based on compensation for morally arbitrary bad luck and would offset any reduction in claims owing to personal responsibility for HIV-status.

Given the importance of the claim based on disadvantage in mitigating any reduction in claims based on personal responsibility, it is worthwhile reiterating the importance of the social context in HIV infection in South Africa. HIV is a viral disease caused by a retrovirus that attacks the body's immune system. Not all viruses become epidemics – to do this, the virus needs a niche or social context within which it can thrive. While HIV/AIDS does not only affect the poor, poverty is argued to be the main aspect of its social context (Van Niekerk 2001). Globally, the HIV epidemic is predominantly situated within relatively poor countries. Sub-Saharan Africa has only 10 per cent of the world's population, but over 60 per cent of the world's HIV-infected people (25.8 million) (UNAIDS and WHO 2005). A number of social factors are argued to place South Africa at a high risk of an HIV epidemic, including inequalities in income and levels of

[21] One example would be needle stick injuries where a health professional accidentally pricks herself with a needle that has been used on a patient.

employment, mobility, and violence (Fassin and Schneider 2003). Although South Africa is an upper middle income country, it also has high rates of unemployment, abject poverty among more than 50 per cent of the population, sharp inequalities in the distribution of income, property and opportunities and high levels of crime and violence (Terreblanche 2002). Particularly striking is the strong racial bias in these inequalities[22]. For example, in 1995, the per capita income of whites was 7.4 times higher than that of Africans. While 6.7 per cent of whites were unemployed[23], the corresponding figure was 46 per cent for Africans. By 1993, estimated real per capita social spending on Africans was just over half the level of social spending on whites (Terreblanche 2002). Health policy, like other government social policy, served the objective of maintaining economic and political power for whites (McIntyre and Gilson 2002; McIntyre, Gilson et al. 2006). In the early 1990s this bias translated into infant mortality rates that were nearly 11.5 times higher for Africans than for whites and maternal mortality rates that were 31 times higher for African than for white women (McIntyre and Gilson 2002)

Inequalities in income and employment are argued to increase a person's vulnerability to HIV infection via higher exposure through risky sexual behaviour, diminished access to health information and preventive devices, higher frequency of sexually transmitted infections which increase vulnerability to HIV, absent or delayed diagnosis and treatment of these infections, and less concern about one's health and the future owing to the difficulties of the present (Johnson and Budlender 2002). Data from a 2002 household survey provide some support for this hypothesis in that prevalence was 40.2 per cent for those who had a genital ulcer within the last 3 months (Shisana and Simbayi 2002). Besides sexually transmitted diseases, it is also argued that malnutrition and certain types of parasites prevalent in Africa increase susceptibility to HIV infection (Stillwaggon 2002).

The second factor that is argued to be associated with vulnerability to HIV is the mobility of the population (Lurie 2000; Fassin and Schneider 2003). This mobility has historical importance in South Africa. During the period of colonial rule, labour was scarcer than land, with the result that the white landowning elite forced the indigenous population into slavery, serfdom and other

---

[22] The use of the terms "African", "coloured", "Indian" and "white" reflects the stratification of the population into race groups in terms of the former Population Registration Act. The term "blacks" refers to Africans, coloureds and Indians. While it is necessary to maintain these terms when discussing the legacy of Apartheid and resultant socioeconomic and health inequalities, this does not imply any legitimacy in these terms.

[23] Unemployment is according to the expanded definition of the labour force outside the formal sector.

repressed forms of labour. In many cases, it was only possible to acquire labour by depriving indigenous people of their land. However, by the late 19th and early 20th century, almost all arable land in South Africa was occupied by white landowners and smaller farms became economically unviable. For the first time, white landowners entered the job market where they were in direct competition with African labour. This situation eventually led to the enactment of a number of laws aimed at keeping Africans subjugated as a subservient labour force, thereby protecting the elite status of the white population group. An attempt at separate development of each race group also implied that while a certain supply of African labour in urban areas was desired by the Apartheid government, generalized urbanisation was not. A small minority of African people were therefore allowed to live permanently in urban areas, but the vast majority were confined to the "homeland[24]" areas. Overcrowding and deteriorating conditions in these areas made it urgently necessary for people to seek work in urban areas, a process which was tightly controlled under a system of influx laws (Terreblanche 2002).

An example of this system that is particularly pertinent to the discussion of HIV transmission is the migrant labour system in the mines. Levels of HIV infection are amongst the highest in mine workers. For example, the gold mines employ around 350,000 workers, 90 per cent of whom are migrants either from the ex-homeland and rural areas in South Africa or from other countries such as Lesotho, Botswana and Mozambique. The vast majority of workers are housed in single sex hostels with up to 18 people sharing a room. According to informants in one study, drinking and sex appeared to be some of the few activities available on a day to day basis that could divert attention from the stressful, dangerous and physically taxing nature of work underground in the mines (Campbell 1997). Mine workers who acquire HIV and other sexually transmitted infections while working on the mines potentially transmit these infections to their wives when they return to their homes during holidays, thereby facilitating the spread of HIV between different communities.

The third aspect of the social context that is argued to increase South Africa's risk of an HIV epidemic is high levels of sexual and physical violence (Wood, Maforah et al. 1998). According to Terreblanche (2002), high levels of violence and crime in South Africa are rooted in the history of repression, discrimination, political struggle and chronic community poverty. While levels of

[24] In terms of the 1913 "Natives Land Act", Africans were confined to living in ten homelands which were scattered throughout South Africa. These areas comprised less than 14 per cent of the total surface area of South Africa.

crime in general are high in South Africa, poor women and children (the majority of whom are African) are particularly vulnerable to violent crime which is often perpetuated within households (Terreblanche 2002). It is hypothesised that there are at least four mechanisms through which violence can increase exposure to HIV. Violence may increase a woman's risk of HIV infection through forced/coercive sexual intercourse and by limiting women's ability to negotiate the use of HIV prevention technologies such as condoms. Physical and sexual abuse during childhood has also been associated with high sexual risk-taking behaviour in adolescence and adulthood. Also, women who are HIV-positive and disclose their status may be at increased risk of violence (Maman, Campbell et al. 2000). In a study in an antenatal clinic in South Africa, it was found that physical intimate partner violence was associated with increased odds of HIV infection (Dunkle, Jewkes et al. 2004).

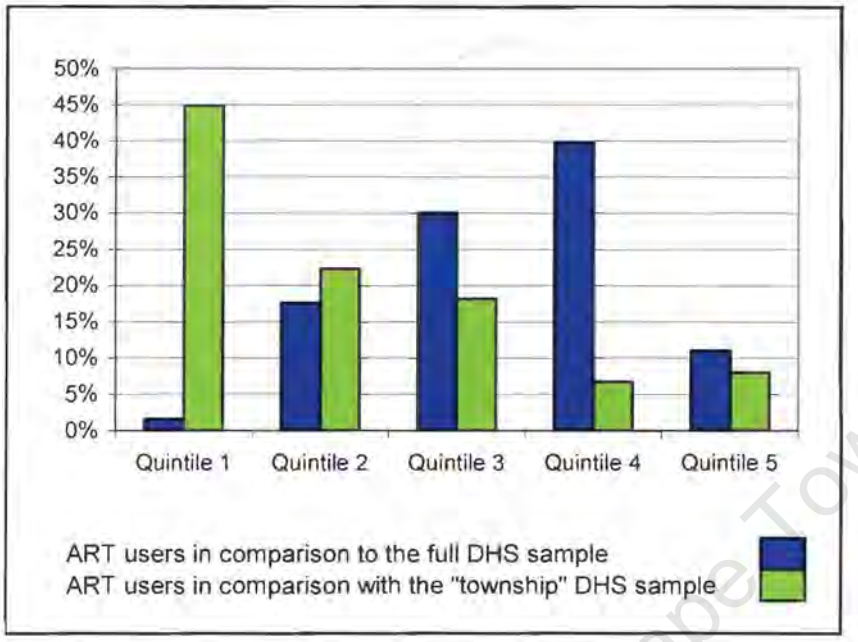
The strong race dimension of inequalities in income, employment, exposure to physical and sexual violence and the need to migrate from one's home in search of work is reflected in recent HIV-prevalence statistics from a nationally representative survey. These data indicate that Africans have by far the highest prevalence (19.9 per cent), followed by coloureds (3.2 per cent), Indians (1.0 per cent) and Whites (0.5 per cent) (Shisana, Rehle et al. 2005). In certain mining communities, HIV-prevalence is over 20 per cent for women who have had one lifetime sexual partner and is up to 50 per cent for women who have had three sexual partners (Williams, Gilgen et al. 2000; in Fassin and Schneider 2003). As argued by Fassin and Schneider (2003 p. 496), "in this instance, social context has a far greater bearing on risk of infection than individual sexual behaviour". This point once again highlights the importance of disadvantage in mitigating claim reductions based on personal responsibility.

While HIV has clear race dimensions in South Africa, there is growing evidence to suggest that income inequalities are increasingly taking on class as opposed to race dimensions - within-race income inequalities are rising in particular within the African group (Terreblanche 2002). The claim based on disadvantage suggests the need for claims preferentially to benefit the poor even within the historically disadvantaged African group. While there are only limited data regarding the socioeconomic status of HIV-positive South Africans, one question in an earlier version of the above-mentioned survey reported that prevalence was 13.9 per cent, 14 per cent, 6.5 per cent, and 5.0 per cent respectively for respondents who had "not enough money for food", "enough for basics, short for other", "enough for most important things", and "some money for extras" (Shisana and Simbayi 2002).

The race dimension of HIV in South Africa is further supported by evidence showing that Africans are the main beneficiaries of HIV-treatment services. In Aid for AIDS, a private sector HIV/AIDS disease management programme, a study on a sub-sample of beneficiaries (n=6,288) indicated that 97 per cent were African (Nachega, Hislop et al. 2006). Similarly, in the public sector, while data were not presented on race, at least 91 per cent of the sampled recipients (n=749) of ART services in 5 clinics in the Western Cape were African, as proxied by home language (Pienaar, Myer et al. 2006). Data on the socioeconomic status of these patients also indicate that the service has the potential preferentially to benefit the poor. Socioeconomic status[25] of ART users in this instance is compared against the asset quintiles of South Africa as a whole and against the asset quintiles of a sample of township residents, following a methodology described in Thiede, Palmer et al. (2005). Results suggest that these users of ART services have a relatively high socio-economic status when compared to the population as a whole, but are poor in comparison to their communities with over 50 per cent of users represented by the poorest two quintiles. Thus while Western Cape ART users could be wealthy by the standards of the country, these data indicate that ART services are accessible to the poorest in urban township communities. Whether or not this finding is a function of the far greater numbers of poor HIV-positive people in these communities is unknown.

[25] Socioeconomic status has been assessed using the asset index methodology which has been described in Chapter 4 and previously used to assess non-HIV related mortality.

Figure 30: Asset quintiles of ART users in the Western Cape



To conclude, given that a particular social context is argued to encourage HIV-transmission and that this social context is argued to be the legacy of Apartheid (Terreblanche 2002), a focus on individual responsibility for HIV-status runs a risk of penalizing people who have become ill through no fault of their own. Such a stance could increase levels of stigma and discrimination against HIV-positive people and reduce levels of social solidarity (see additional discussion under claim 8 in section 6). Because HIV finds fertile soil amongst the most historically disadvantaged race group in South Africa one could argue that there is a social responsibility to provide additional access to HIV-treatment to these groups for reasons of vertical equity. According to Mooney (1996 p. 102): "if, as is normally the case, ill health is not randomly distributed across different groups in society, might that society not want to give preference, on vertical equity grounds, for health gains to those groups in that society who are on average in poor health?" Participants in a fair process might therefore be concerned that HIV-positive people be given preferential access to health care services (and to other social services given the relationship between health and other social inequalities) to counter the legacy of disadvantage that has potentially made them vulnerable to HIV-infection.

4 What health care and health benefit?

The key empirical contribution of this thesis has been an assessment of technical efficiency and equity in HIV-treatment at the population-level – the question of what health care should be distributed and what health benefit could be achieved. In assessing these issues, the thesis has taken the approach of quantifying health care benefits solely in terms of health via the QALY model. Given that the focus of this dissertation has been one of maximisation or equalisation within the health care budget constraint, it has also been necessary to focus solely on the opportunity cost of health care resources. While this approach is theoretically correct, it is also narrow. Section 4.1 therefore revisits weaknesses in the QALY approach and discusses other benefits from health care. Section 4.2 reviews the data that have been presented regarding the question of whether a base case or less resource-intensive (generalized) approach to care should be adopted. Section 4.3 discusses the three mutually exclusive HIV-treatment strategies and Section 4.4 revisits the debate about whether the medical eligibility criteria for treatment initiation should be revised.

In a more practical context, the following comments might apply to the different social choice rules:

- Those who favour equal health in HIV-treatment would be likely to be in support of focusing on developing the infrastructure and capacity of traditionally poorly served areas such as rural and ex-homeland areas.
- Those who favour health maximisation would be likely to accept that access to treatment is better in areas where existing capacity is the greatest and hence where the largest numbers of patients can be enrolled quickly into care.
- Those who favour the decent minimum would be likely to prefer a balance between utilising existing capacity and creating additional capacity thereby trading off between the maximisation of outcomes and increasing geographical coverage.

4.1. QALYs and other health care benefits

This thesis has defined health from the individual's point of view as an absence of illness or injury, quantified through the QALY approach. While this approach is defended given the need to assess opportunity costs in HIV-treatment, as reflected by the alternative health care benefits that are achievable, it is also relevant to consider the benefits of health care that are not captured within the QALY approach.

Ryan (1999) has shown through the example of in vitro fertilization that the benefits of health care are broader than health, and can include process attributes such as "attitudes of staff towards the patient". While there is clearly more that is utility generating from health care than health gains, the importance of non health attributes (both process and outcome) will depend on the intervention or disease. Thus while the individual utility (or disutility) that is gained from HIV diagnosis is more likely to be related to the information that is received during the test, in the context of HIV-positive people facing imminent death, it is likely that treatment is particularly valued in terms of reductions in morbidity and mortality. However, this dominance of health-related utility in HIV-treatment might change over the duration of the patient's lifetime. While initial utility might largely relate to health gains, as health improves and duration on treatment increases, non-health outcomes and process attributes might become more important. These could include the attitudes of health care staff, continuity of contact with the same staff given the long-term nature of treatment, autonomy and access to information. These elements of health care benefit are not captured in the QALY.

A further shortcoming of the QALY approach used in this thesis is that it is assumed that a QALY has a constant marginal value, when this is unlikely to be the case. With application to HIV-treatment, this assumption means that one is indifferent between 1 QALY gained on No-ART followed by death and 1 QALY gained within the context of more substantial health gains on ART. To illustrate this point, imagine that a respondent has a value base that is best reflected by the decent minimum social choice rule and that the HIV-treatment budget has an upper constraint of US$9 billion. These two choices point towards around half of the patients in need receiving No-ART and the rest receiving first-line ART. While it is possible within a procedurally just approach for participants to disagree about outcomes, in this case it is also possible that disagreement about outcomes could be explained by increasing marginal value of health, especially in the context of interventions such as No-ART where individual health gains

are small. If the marginal value of a QALY gained on No-ART is lower than has been assumed, this would mean that No-ART is relatively less cost-effective, while first-line ART and first and second-line ART are relatively more cost-effective. This shortcoming of the QALY approach would be particularly problematic if one were given no information about the health care interventions that were under discussion. It is therefore important that participants in a fair process be given information about both the individual and population health gains from alternative interventions, and that any decision-making is iterative to allow for adequate reflection and revision if necessary.

4.2. Base case or generalized treatment

While the base case scenario is based on utilisation, unit cost and outcome data that are derived from the long-term follow-up of a cohort of patients in the three Khayelitsha HIV-clinics where ART delivery was piloted prior to the initiation of this service in the country, the generalized scenario is based on a number of assumptions designed to estimate the lifetime costs and outcomes of a less resource intensive approach to treatment. Three key adjustments were made. Firstly, ART visit utilisation was assumed to occur at the frequency recommended in the National Antiretroviral Treatment Guidelines (2004). This amounts to 6 visits instead of between 9 and 10 during the pre-ART period and first 3 months on treatment in Khayelitsha; 3 visits as opposed to 4 in months 3 to 6; and 6 visits as opposed to 7.2 in months 6 to 12. Thereafter, visits are assumed to continue at a rate of approximately 1 per month. No-ART visits were based on a large local natural history cohort (Badri, Cleary et al. 2006), and were considerably reduced from 3.4 to 1.9 in the CD4<50 cells/μl group, and from 2.6 to 1.4 per quarter in the CD4 50-199 cells/μl group.

The second key adjustment is to unit costs. Inpatient care was restricted to secondary and district levels (approximately 30 per cent of inpatient care was at the tertiary level in Khayelitsha) and No-ART outpatient care was restricted to clinics and community health centres. This reflects the type of care that exists in most of South Africa outside the larger cities. The average cost per inpatient day was reduced from US$161.67 to US$118.56 and the average No-ART visit cost was reduced from US$19.62 to US$17.24. No adjustments were made to ARV regimens or the frequency of laboratory investigations – these continue to reflect the recommendations in the national protocol (Department of Health 2004).

The third adjustment, which has the largest impact, is to clinical outcomes. Death transition probabilities have all been increased by approximately 40 per cent to reflect clinical outcomes at one year in a number of developing country ART cohorts (ART-LINC and ART-CC 2006) and a review of developing country natural history cohorts (Schneider, Zwahlen et al. 2004). These reduced outcomes could reflect the potentially lower adherence and more rapid development of drug resistance that could be an implication of less resource-intensive models of care (Mugyeni 2004)[26].

Whereas in the base case scenario, discounted lifetime costs are US$2,966 for No-ART, US$5,779 for first-line ART and US$9,435 for first and second-line ART, these are reduced by approximately US$1,150 for No-ART, US$1,500 for first-line ART and US$4,000 for first and second-line ART in the generalized scenario. Similarly, while outcomes are 2.9 life years for No-ART, 8.5 life years for first-line ART and 12.9 life years for first and second-line ART at a zero annual discount rate in the base case scenario (2.1, 7.1 and 10.8 QALYs respectively), No-ART life expectancy was decreased by 0.6, first-line ART life-expectancy was decreased by 1.6 and first and second-line ART life-expectancy was decreased by 3.2 years in the generalized scenario. ICERs in the base case and generalized scenarios are similar. Given that most of the reduction in lifetime costs in the generalized scenario is related to lower life-expectancy, it is instructive to note that the ICERs in the generalized scenario are lower if no adjustment is made to outcomes. This indicates that the generalized scenario is likely to have conservatively underestimated cost-effectiveness.

When the costs of scaling up are assessed, the generalized scenario achieves both higher QALYs and higher coverage than the base case scenario under most social choice rules. Under health maximisation at budgets lower than US$11 billion, the generalized scenario has higher QALYs and covers a greater proportion of need than base case. For budgets of over US$11 billion, the base case scenario is superior owing to the higher outcomes that are achieved through this approach. Under the decent minimum, no treatment would be offered in the base case scenario if the budget is less than US$7 billion, while solutions are possible in the generalized scenario at budgets between US$5 and 7 billion. The generalized scenario continues to be superior to base case in terms of higher QALY gains and a higher proportion of patients receiving either first-line

[26] A study on adherence in Khayelitsha showed that a high proportion of patients achieved suppression of viral replication – virus was <400 copies/ml in 89, 84 and 70 per cent of patients at 6, 12 and 24 months respectively (Coetzee, Boulle et al. 2004).

ART or first and second-line ART until the budget exceeds US$11 billion. Under equal health, the generalized scenario outperforms base case because, as with the decent minimum, it offers a solution when the budget is between US$5 and 7 billion. For budgets between US$8 and 9 billion, base case has higher QALYs than generalized, which is mainly owing to the step-wise nature of the equal health social choice rule, where no patients can be moved onto a more effective treatment strategy until sufficient budget is accumulated to allow all patients to be moved simultaneously. However, the budget required to move to a more effective strategy is accumulated sooner in the generalized scenario than in base case, which means that once a budget of US$10 billion is reached, all patients are put onto first-line ART with a QALY benefit of 4.4 million in comparison to base case.

In many respects, this generalized scenario could be argued to enhance the generalization of results from a pilot setting to the general HIV population as opposed to being a reflection of a less resource intensive model of care. When the latter is debated in the literature, the discussion is of a model of care that is far less costly than the generalized scenario. In terms of utilisation, the debate concerns whether visits to a doctor can be reduced from once every three months (which is the norm in developed countries and is recommended in South Africa[27]) to once every six months or once per annum or even whether routine consultations with a doctor are necessary, especially during the maintenance period on treatment (Jaffar, Govender et al. 2005; Harries, Schouten et al. 2006). Similarly, it is debated whether treatment can be delivered through lower cadres of health professionals such as nurses or even community health workers given the scarcity of doctors in many countries (Farmer, Leandre et al. 2001; Calmy, Klement et al. 2004; Jaffar, Govender et al. 2005; Stewart and Loveday 2005; Harries, Schouten et al. 2006).

Many countries use a fixed dose combination of stavudine, lamivudine and nevirapine in their first-line regimen at a cost of around US$140 per annum (MSF 2002) as opposed to the more complex non fixed-dose combinations that are used in South Africa which have an average annual cost of US$291[28]. It is argued that this easy-to-use first-line regimen has been fundamental to the development of large-scale ART programme with few resources (Calmy, Klement et al. 2004; Ferradini, Jeannin et al. 2006). Although this fixed-dose regimen is pre-

[27] Visits are to a combined doctor/nurse team in Khayelitsha – the monthly visits mentioned in an earlier section are therefore not necessarily doctor visits.
[28] First-line regimens are stavudine, lamivudine and nevirapine (non fixed-dose version) or stavudine, lamivudine and efavirenz.

qualified by the World Health Organisation (WHO 2003) and has been shown to have excellent safety and efficacy (Laurent, Kouanfack et al. 2004), South Africa continues to use the less patient-friendly and more expensive non fixed-dose versions. There is also debate about the type and frequency of laboratory investigations that are necessary (Calmy, Klement et al. 2004; Jaffar, Govender et al. 2005). WHO (2002) ART guidelines distinguish between four categories of laboratory investigations: absolute minimum, basic recommended, desirable and optional. South Africa's recommendations fall within the optional category. Treatment initiation in Haiti was successfully based on clinical criteria during the period 1998 to 2002 when CD4 testing was not available (Farmer, Leandre et al. 2001; Koenig, Leandre et al. 2004). In Malawi, where less than 10 per cent of patients initiate care based on CD4 criteria, it is argued that the introduction of CD4 and viral load laboratory tests could undermine the country's ART programme through the introduction of unnecessary complications (Harries, Schouten et al. 2006).

Additional adjustments to costs were therefore made in a "low cost generalized scenario". These included assuming that all patients would receive non fixed-dose combinations of stavudine, lamivudine and nevirapine, which reduces annual first-line ARV costs from US$291 to US$162 (additional cost reductions of US$22 per annum could be achieved if the fixed-dose combination was used). It was also assumed that no viral load testing would be undertaken, which reduces costs by between US$60 and US$70 per annum[29]. In sum, this scenario reduced undiscounted lifetime costs by an additional US$1,400 for first-line ART and by US$1,600 for first and second-line ART in comparison to the generalized scenario. The discounted ICER per QALY gained for low cost first-line ART was US$561 in comparison to US$921 in the generalized scenario, and the ICER for first and second-line ART was US$1,541 in comparison to US$1,635.

To conclude, it has been shown that the implementation of a less resource-intensive model of HIV-treatment has the potential to enhance equity and efficiency in HIV-treatment at lower budget constraints[30]. A key argument against these less resource-intensive approaches, however, is that they could accelerate the transmission of resistant virus strains from poorly adherent patients to other members of the population, which could result in a negative externality for population health. This issue is explored in detail in section 5.

---

[29] All other laboratory investigations are maintained according to the national protocol. This means that the low cost generalized approach is still more resource-intensive than many treatment programmes in other African countries.

[30] As usual, if higher budgets are available then health is maximized through more resource-intensive approaches.

4.3. No-ART, first-line ART or first and second-line ART

In the approach advocated in this thesis, the proportion of patients receiving different treatment strategies is a function of the feasible treatment strategies, the budget constraint and the value base. In other words, numbers of patients receiving alternative treatment strategies are not explicitly chosen, but are instead the solution of prior choices. This is true except if the set of feasible treatment strategies is changed. While changing the treatment policy to include or exclude certain strategies would not influence the overall approach described for decision-making, the empirical results would need to be recalculated. This section discusses the interplay between value base, budget constraint and the resultant QALYs that can be gained, proportion of need that can be reached and the percentage of patients receiving No-ART, first-line ART and first and second-line ART.

If a health maximisation social choice rule is adopted, No-ART treatment, although included in the feasible set, would never be a policy choice. At budgets less than US$11 billion, up to 92 per cent of patients receive first-line ART. For budgets between US$11 and US$12 billion, a strategy that mixes first-line ART with first and second-line ART maximises QALYs. If the budget is greater than US$13 billion, first and second-line ART is the dominant policy. Similar results are found in the generalized scenario, although first and second-line ART becomes the dominant strategy at a lower budget. To summarize, under a health maximisation social choice rule, No-ART is never a policy choice and first-line ART provides the largest health gains over most budget ranges, unless the highest budgets are available.

Under the decent minimum, a strategy of No-ART is combined with first-line ART when budgets are between US$8 and US$10 billion (or US$5 and US$9 billion in the generalized scenario) which leads to a QALY loss of between 0.6 and 1.9 million in comparison to the health maximisation strategy. Once budgets are US$11 billion (or US$10 billion in the generalized scenario), the decent minimum has the same results as health maximisation.

Equal health is the most austere social choice rule in terms of QALY gains, but has the benefit of ensuring that all patients receive the same form of treatment. It is associated with a step-wise approach to scaling up treatment - patients are maintained on an equally effective treatment until the budget is sufficient to allow the entire group to be moved to a more effective treatment. The

implication is that QALY losses in comparison to health maximisation accumulate over certain ranges of the budget, whereas in the decent minimum, QALY losses decrease as the budget increases.

Current government HIV-treatment policy includes first and second-line ART for a proportion of the HIV-positive population in need. Although policy states that patients who are unable to access ART should receive No-ART care, the likely uptake of this treatment is low. It was therefore assumed for the purposes of costing the Operational Plan that generalized No-ART care would be used by patients who were not able to use ART. Results indicate that 8.4 million QALYs could be gained through this treatment programme between March 2004 and March 2014, and that 83 per cent of these would be gained through first and second-line ART. The total cost of the programme is estimated to be US$9 billion.

If one assumes that US$9 billion is the government's treatment budget over this period, mathematical programming can be used to assess whether improvements in efficiency and equity could be attained in comparison to the government's current approach. While the assumption that the budget is US$9 billion is clearly heroic, the general conclusions and line of argument would not be altered by the use of a different budget, particularly if this were lower.

If first-line ART is not a policy choice, then QALYs can be gained in comparison to the Operational Plan through taking a health maximisation approach to treatment under either base case or generalized scenarios, but because a health maximisation strategy does not include No-ART care, a proportion of patients would be untreated. Alternatively, if a less resource intensive approach is taken (generalized scenario), then QALYs can be gained through the decent minimum, but a lower percentage of patients will be able to access first and second-line ART. This is because the decent minimum does not mix base case and generalized scenarios, while the assumption in costing the Operational Plan is that No-ART care in reality would be more likely to resemble generalized No-ART. Implementing an equal health approach is unsurprisingly associated with substantial QALY losses in comparison to the Operational Plan.

If first-line ART is a policy option, then more possibilities are available for improving coverage of patients with ART and for gaining QALYs. In particular, a generalized health maximisation strategy gains 0.77 million QALYs and provides first-line ART to 95 per cent of patients in need. Again in the generalized scenario, the decent minimum gains 0.7 million QALYs and provides

first-line ART to 89 per cent of patients. As before, equal health is associated with QALY losses because the budget is too low for all patients to receive first-line ART.

While adopting a first-line only strategy might shift South Africa on to a higher welfare contour, because this would violate the no-loser constraint, there would be likely to be some opposition to this change in treatment policy. What this means is that patients who have been able to access ART to date would suffer the curtailment of individual health gains in favour of a more equal sharing of health care benefits. While it is ultimately the task of participants in a fair process to consider whether a first-line strategy would be fairer, one might be tempted to concur with Harries, Schouten et al. (2006) who argue that when the majority of patients do well on the first-line regimen, and when unmet need remains large, then priority should be given to the provision of first-line treatment to those who are not receiving ART rather than offering better care for a minority who are already on ART.

It is also possible that when government established its ART strategic plan, it assumed that first and second-line ART would be less effective than appears to be the case. For example, in the costing of this plan it was assumed that first and second-line ART would gain 5.5 life years (Joint Health and Treasury task team 2003). Similarly, in the costing of the World Health Organisation's "3 by 5" it was assumed that between 5 and 7 life years would be gained depending on the level of development of the country (Gutierrez, Johns et al. 2004). In contrast, the data from this thesis have indicated that the outcomes from first-line ART could be 8.5 life years while first and second-line ART could be 12.9 life years under base case assumptions[31], and even this could be conservative when compared to the 31.1 years reported in patients of a similar age in a large European HIV cohort study (van Sighem, van de Wiel et al. 2003). The implication is that government committed itself to therapy that is potentially far more effective than envisaged and therefore far more costly given that 65 per cent of first and second-line costs are associated with the recurrent utilisation of medicines and laboratory investigations.

To summarize, first-line ART is favoured by more social choice rules over a greater range of budgets than No-ART or first and second-line ART. This result is strengthened within the context of less resource intensive models of care. Limiting ART policy to first-line only would need to be

[31] In the generalized scenario, outcomes are 6.9 and 9.7 life years for first-line ART and first and second-line ART respectively.

undertaken with due consideration for the South African constitution's commitment to the progressive realisation of socioeconomic rights[32]. What this means is that if equitable access to first-line ART were achieved, the next step would be to investigate extending access to second-line. This would become especially relevant if second-line regimens became less complex. Currently, barriers to second-line ARVs include the higher cost (over three times more expensive than first-line regimens), higher pill burden (approximately 9 pills per day as opposed to between 5 and 6[33]) and refrigeration requirements which necessitate a cold chain in transportation and storage (Calmy, Klement et al. 2004). If attempts are successful to improve the ease of administration of these regimens, reconsideration of a first-line only policy would become even more relevant.

Before concluding, it is necessary to revisit the use of a No-ART treatment strategy. The costs of base case No-ART include between 10 and 14 clinic, community health centre or hospital outpatient department visits per annum; between 0.3 and 0.4 cases of treated tuberculosis; and between 1.2 and 2.8 ongoing days in hospital with an additional 5.7 at the time of death. Besides the use of medicines for the treatment of opportunistic infections, prophylactic medicines are also included. This care is associated with 2.9 life years at a zero annual discount rate. These outcomes, which are similar to the outcomes from developed world cohorts during the pre-ART era (Schneider, Zwahlen et al. 2004), have been derived from a large natural history cohort receiving treatment at one tertiary and one secondary hospital in Cape Town (Badri, Bekker et al. 2004). In contrast, developing world cohorts have average outcomes of 2.3 life years at a zero annual discount rate (Schneider, Zwahlen et al. 2004). The difference between the Cape Town cohort and other developing country cohorts is likely to relate to the additional access to care that the former received. Therefore, the 2.3 life years in the developing world could reflect survival in the absence of treatment and prophylaxis, while the additional No-ART care described above is associated with a minimal survival gain of 6 months. Participants in a fair process might therefore argue that a No-ART strategy is a better reflection of "do-nothing" in terms of outcomes, while it is associated with fairly substantial lifetime costs (undiscounted) of around US$3,000 per patient. In addition, if the QALYs gained through HIV-treatment are associated with increasing marginal value at the population level, then a No-ART QALY has a lower value than an ART QALY. If

[32] According to London, (2003 p. 10) "progressive realization balances the recognition of the limitations of existing resource constraints, with the obligation on the state to increase, over time, its legislative and financial commitments to meet the socio-economic entitlements of the most vulnerable".
[33] If the fixed-dose combination of stavudine, lamivudine and nevirapine were used in first-line, the pill burden would be 2 per day.

participants in a fair process decide that No-ART is not a reasonable treatment policy because it is associated with the provision of ineffective yet costly care, this implies that higher budgets need to be available before equal health or decent minimum solutions become possible; the overall approach to decision-making advocated in this thesis is however unaffected.

To conclude, this section has illustrated how the choice of the budget and the value base determines the treatment strategy. As before, these choices should be iterative. If upon reflection the outcomes or coverage are felt to be inadequate, the implication is that either the budget is too low, or that the chosen value base needs adjustment. If the less effective treatment strategies are felt to be inappropriate policy choices, then higher budgets need to be available to operationalise the more equitable social choice rules. By focusing on setting the value base and the budget constraint, decision-making around HIV-treatment strategies could be a far more straightforward process.

4.4. The when-to-start debate (claims 4-5)

In South Africa, a patient is medically eligible for ART if she or he has an AIDS diagnosis at any CD4 level or a CD4 count of less than 200 cells/μl at any WHO stage. While these medical eligibility criteria are usually thought of as merely technical, in reality they reflect a trade-off between maximising individual patient benefits and increasing coverage. Because of this trade-off, patients are eligible for ART at a point where their capacity to benefit is slightly reduced owing to their advanced stage of illness.

In the developed world, a recognition that the long-term use of ART could be associated with severe toxicities initially lead to a shift in clinical opinion from a 'hit-early hit-hard' strategy where ART was initiated at high CD4 levels, to a more conservative strategy of deferring treatment until a patient's CD4 count was less than 200 cells/μl (Ho 1995; Lane and Neaton 2003). However, the case for earlier treatment has recently been re-examined (Holmberg, Palella et al. 2004). The CD4 count threshold at therapy initiation is an important determinant of both clinical benefit (Ho 1995; Hogg, Yip et al. 2001; Sterling, Chaisson et al. 2001; Lane and Neaton 2003; Phillips, Lepri et al. 2003; Holmberg, Palella et al. 2004) and cost-effectiveness of ART (Freedberg, Losina et al. 2001; Schackman, Freedberg et al. 2002; Bachman 2006; Badri, Cleary et al. 2006). Both in developed countries (Hogg, Yip et al. 2001; Schackman, Freedberg et al. 2002; Lane and Neaton 2003; Phillips, Lepri et al. 2003; Sterling, Chaisson et al. 2003) and in

developing countries (Bachman 2006; Badri, Cleary et al. 2006), studies indicate that starting ART at CD4 cell counts higher than 200 cells/μl could impact favourably on outcomes. A recent review concluded that earlier initiation of ART could be associated with better immune improvement, less drug related toxicity and reduced HIV transmission (Lane and Neaton 2003). However, current guidelines recommend initiating ART at a variety of CD4 cell count thresholds for patients with high viral loads (European AIDS Clinical Society 2001; British HIV Association 2003; WHO 2003; Yeni, Hammer et al. 2004; Department of Health and Human Services 2005).

Analyses of the Cape Town AIDS Cohort have indicated that earlier initiation of ART is associated with reduced incidence of tuberculosis (Badri, Wilson et al. 2002), AIDS and death (Badri, Bekker et al. 2004) in comparison with deferring treatment to CD4<200 cells/μl. Extrapolation of available data with the use of Markov modelling concluded that undiscounted life expectancy would be 23, 21 and 19 if ART were started with CD4>350 cells/μl, CD4 200-350 cells/μl and CD4<200 cells/μl respectively. However, earlier initiation was inevitably associated with higher lifetime costs given that at least 50 per cent of ART costs are associated with recurrent medication and laboratory investigations. The study found that the ICER was lowest for starting ART at CD4<200 cells/μl (US$611 per QALY gained at a zero annual discount rate), followed by 200-350 cells/μl (US$915 per QALY gained) and finally >350 cells/μl (US$1,236 per QALY gained). Whether any of these therapy initiation criteria is considered to be technically efficient depends on the available budget. In general, if lower budgets are available, health could be maximised by starting treatment at CD4<200 cells/μl as this threshold has the lowest ICER. However, if higher budgets are available, it would become more cost-effective to start ART earlier.

On the other hand, data from the Khayelitsha HIV clinics has shown that within a group of patients starting ART with CD4<200 cells/μl, it could be less cost-effective to delay treatment to CD4<50 cells/μl in comparison to starting when the CD4 count is between 50 and 200 cells/μl. If one were to define need as illness, these sicker patients would be prioritised, but if need were defined as capacity to benefit, an opportunity for a better prognosis for those who have enrolled in the programme in a timely manner would be preserved. The conflict between these two principles is likely to be of ongoing concern for health professionals in settings such as Khayelitsha (Coetzee, Hildebrand et al. 2004).

While the widespread experience of Medecins sans Frontieres[34] in a number of countries suggests that putting those with the greatest risk of death first is most in keeping with the notion of fairness held by people living with HIV/AIDS (Calmy, Klement et al. 2004), this proposition needs to be verified and broader societal views need to be sought. A fair process could assist in deciding on the medical eligibility criteria for ART initiation, or in different language, to what extent capacity to benefit could be taken into account (claims 4 and 5 in Figure 29).

However, to do this using the framework proposed in this thesis, demographic models would need to be developed that provide estimates of patients in different CD4 categories instead of in WHO stages[35] so that it would be possible to assess the overall costs and outcomes of these alternatives. In the absence of these models, one might be tempted to assume that extending the criteria to higher CD4 levels would dramatically increase the number of eligible patients. The ASSA2003*lite* model indicates that in 2006 there were an estimated 542,000 adult AIDS cases and 1,771,000 Stage III cases. One might therefore be tempted to assume that extending eligibility criteria to include Stage III cases would more than triple the recurrent need. However this would fail to distinguish between total current cases at a particular date (prevalent cases) and the new cases (incident cases) that develop during a period of time. Only around 80 per cent of the prevalent AIDS cases by mid 2006 were incident cases during that year. While the ASSA model does not distinguish prevalent from incident Stage III cases, it might be the case that a lower proportion of the total Stage III cases by mid 2006 would be incident. This is because overall duration - the time that an individual can be expected to stay in Stage III before developing AIDS or dying – is higher than is the case for AIDS.

Imagine therefore that the medical eligibility criteria were changed from CD4<200 cells/μl to CD4≤350 cells/μl. As before, any patient that develops AIDS would also be eligible. Badri, Cleary et al. (2006) have calculated the one-month probability of movements between various No-ART health states for patients who enter "care" at CD4>300 cells/μl without AIDS. These can be used to estimate the monthly probability that a patient becomes eligible for ART. If the criterion remains CD4<200 cells/μl, the probability is 0.037 while if the criterion becomes CD4≤350 cells/μl, the probability increases to 0.049. Thus while clearly more new patients would

[34] Medecins sans Frontieres is an international NGO that has been critically involved in the provision of ART. By 2004, their experience included treating 21,000 patients in 27 different countries, including patients in Khayelitsha.
[35] The same is true of the demographic models used by UNAIDS as reported in their annual "AIDS Epidemic Update".

become eligible each year, the increase is likely to be lower than one might assume by looking at total patients in each group. However, because survival is superior when ART is initiated earlier, in the long-run a far larger number of patients would be in care.

It is therefore recommended that once a CD4-based demographic model has been created (where new cases are distinguished from total cases) these data should be fed into the social choice models so that a thorough assessment can be made of the total costs and outcomes associated with initiating ART at different CD4 levels.

5 Impact on health of society (claim 7)

An analysis of the impact of antiretroviral treatment on the health of society involves the balancing of a number of competing forces. On the one hand, the presence of effective treatment for HIV/AIDS is argued to have the potential to reduce HIV incidence. This is because there are additional incentives for patients to be diagnosed and it is hoped that people who know their HIV status will take steps to either remain negative or to avoid infecting others if they are positive (Moatti, N'Doye et al. 2003). In addition evidence from a longitudinal cohort dataset indicates that because ART suppresses the level of the virus in a patient's body, it can reduce per-partnership infectivity by as much as 60 per cent (Porco, Martin et al. 2004). On the other hand, the widespread availability of effective treatment might lead to what is known as disinhibition – where a reduction in the perceived severity of HIV infection leads to increased risky sexual behaviour. Evidence of increasing HIV incidence in recent years in the gay communities of San Francisco supports this hypothesis (Velasco-Hernandez, Gershengorn et al. 2002; Blower, Bodine et al. 2005). However, a meta-analysis of studies on sexual risk behaviour in antiretroviral treated patients in the developed world showed that these patients did not exhibit increased sexual risk behaviour (Crepaz, Hart et al. 2004). One study in Uganda at 6 months after starting ART showed reduced sexual risk behaviour. When coupled with viral suppression, this study concluded that integrated ART and prevention programmes had reduced the transmission of HIV by 98 per cent in these patients (Bunnell, Ekwaru et al. 2006).

The other key risk to the health of society associated with the introduction of ART is the development and transmission of drug resistant forms of the virus (Mugyeni 2004). The threat of drug resistance in Africa has led to calls for limiting ART owing to the public health risk.

However, others have argued that limiting ART in poor countries owing to fears of drug resistance amounts to double standards between the North and the South (Moatti, N'Doye et al. 2003; Moatti, Spire et al. 2004). A study in North America has shown that the proportion of new HIV infections that involve drug-resistant virus has increased from 3.4[36] per cent during the period 1995 to 1998 to 12.4 per cent between 1999 and 2000 (Little, Holte et al. 2002). Although some studies have shown no resistance in these primary infection cohorts, most studies show that resistance is increasing among newly infected patients (Blower, Bodine et al. 2005). It is important to distinguish between transmitted and acquired resistance. The former refers to the level of drug-resistance in the virus with which one is initially infected, while the latter refers to drug-resistance that develops over time as perhaps an inevitable consequence of long-term treatment. It is important to maintain good adherence to ART in order to maximise individual health gains and to minimise the chance of transmitting a drug resistant virus in the event of unprotected sexual intercourse. Given the importance of adherence, it is interesting to note that evidence from a large meta-analysis indicates that Africans receiving antiretrovirals achieved significantly higher levels of adherence than North Americans – the adjusted odds ratio of ART adherence in African studies in relationship to ART adherence in North America was 3.0 (Mills, Nachega et al. 2006).

Mathematical modelling studies have attempted to estimate the overall impact of these various factors on the health of society in San Francisco where prevalence in the gay community is around 30 per cent (Velasco-Hernandez, Gershengorn et al. 2002) and in sub-Saharan Africa (Blower, Bodine et al. 2005). Results from the San Francisco study indicate that a high usage of ART (between 50 and 90 per cent of all HIV-positive people receiving ART) would significantly reduce the severity of the HIV epidemic. If this was coupled with substantial reductions in risky sexual behaviour, there would be a high chance of eradicating the HIV epidemic entirely. However, the probability of epidemic eradication from the widespread use of ART is reduced to 50 per cent if risky sexual behaviour remains unchanged and if this behaviour increases, the probability is further decreased. However, high levels of ART coverage would still have an overall beneficial impact even with increased risky sexual behaviour and high levels of ARV resistance. This study therefore argues that the use of ART is recommended both as an effective therapeutic strategy and as an effective public health prevention intervention (Velasco-Hernandez, Gershengorn et al. 2002).

---

[36] Defined as high level resistance to one or more drugs.

Because the public health impact of ART is dependent on the percentage of the population that receives treatment, in sub-Saharan Africa, the impact of ART on overall societal health is likely to be small (Blower, Bodine et al. 2005). In South Africa in 2006, only about 3 per cent of the over 5 million HIV-positive people were receiving ART. However, if current treatment policy is implemented, this could increase to about 27 per cent by 2010 and 45 per cent by 2014. Modelling predicts that if 20 per cent are on treatment, prevalence could decrease by 5 to 10 per cent, while if 40 per cent are on treatment, it could decrease by 15 per cent 10 years after the rollout has commenced.

To conclude, empirical data and modelling studies suggest that the provision of ART provides an opportunity to reduce HIV incidence and prevalence and that the positive public health impact of ART is enhanced if a high percentage of HIV-positive people have access to ART. While acquired drug resistance is perhaps inevitable, transmitted drug resistance is predicted to remain low. Evidence also suggests that Africans are not more likely to have worse adherence than North Americans. In sum, there could be substantial positive externalities through the provision of ART on the health of society.

6 Impact on the social fabric (claim 6)

This section outlines the impact that HIV/AIDS and access to treatment is predicted to have on the social fabric. Social fabric, following Haacker (2004), includes social and economic institutions such as households, companies and the government, and less tangible concepts such as social cohesiveness and solidarity.

There is some debate about the impact that HIV/AIDS will have on economic growth and GDP per capita. A number of studies that consider the macroeconomic impact of HIV/AIDS in South Africa have been conducted, using a variety of different models including the Solow growth model (Young 2005), computable general equilibrium model (Arndt and Lewis 2000) and a macro-econometric demand-side model (Quattek 2000; Laubscher, Smit et al. 2001). While all concur that the level of GDP will be lower than it would have been in the absence of the HIV epidemic, there is debate about whether the decrease in the population from AIDS related mortality will on balance increase (Quattek 2000; Laubscher, Smit et al. 2001; Young 2005; Smit, Ellis et al. 2006), or decrease GDP per capita (Arndt and Lewis 2000). One study that included

the macroeconomic impact of the provision of ART to 50 per cent of new AIDS cases concluded that the cost of this programme[37] would be more than offset through enhanced GDP growth in comparison to a No-ART scenario (Smit, Ellis et al. 2006).

Thus while the impact of HIV/AIDS in terms of morbidity and mortality is clear, because this impact tends to be unevenly spread through the population it is less apparent at the aggregate level. A death in one household could be of benefit to another household if a new job becomes available (Haacker 2004). It is therefore argued that studying the impact of HIV/AIDS at the aggregate level provides little information about the impact on individual households. While poor households contribute only a limited amount to overall GDP, they are also less able to accommodate to adverse shocks to income or expenditure which increase their level of vulnerability (Crafts and Haacker 2004). HIV/AIDS could therefore lead to an increase in levels of poverty (Greener 2004). Those who take a human development approach would therefore emphasize that dramatic losses in life expectancy have an important welfare cost that is hidden within aggregate macroeconomic variables (Crafts and Haacker 2004).

In contrast to the more standard macroeconomic studies, research using an overlapping generations model to assess the impact of HIV/AIDS on the long-term formation of human capital pointed to a risk of economic collapse within a few generations unless policy is implemented to protect the development of human capital (Bell, Devarajan et al. 2004). In 1990, 36 per cent of 15 year old South African men and 21 per cent of 15 year old women were predicted to die before their sixtieth birthday. By 2010 it is predicted that these numbers will be 60 per cent for men and 56 per cent for women. The approximately 500,000 maternal orphans under the age of 18 in 1990[38] were estimated to be 1.5 million by 2006 (Dorrington, Johnson et al. 2006). According to Bell, Devarajan et al. (2004) these patterns of adult mortality and levels of orphaned children contain a real threat of economic collapse. Because HIV/AIDS makes it difficult for infected adults to provide for their children's education or to offer them the love and care they need to complement their formal schooling, the result is a generation of undereducated and hence underproductive youth who in turn might find it difficult to provide for their children. This study has however been criticized for its assumption of full employment. Given that

[37] As estimated using similar cost and utilization data as for this dissertation but with an alternative cost projection model – see Boulle and Cleary (2005).
[38] Maternal orphans are reported owing to the difficulty in calculating numbers of full orphans. However, given the importance of mothers in ensuring the health and well-being of children, this is nevertheless a key indicator.

HIV/AIDS is concentrated among the semi and unskilled sections of the labour force that can be replaced comparatively easily from the large pool of unemployed workers, the economic impact of HIV/AIDS could be less pronounced than suggested by studies that assume full employment (Smit, Ellis et al. 2006).

The nature of the care and treatment that HIV-positive people receive in South Africa also has the potential to have an impact on less tangible concepts such as social cohesiveness. Given the scale of the response that is required and the high proportion of the population[39] that is in need, there is a chance that mobilization around HIV/AIDS could assist South Africa to become a more solidaristic society with positive spin-offs for social justice in general. On the other hand if the response is perceived to be inadequate, levels of solidarity could decrease with a destabilizing effect on the security of the country (Haacker 2004).

An adequate response is however contingent on a commitment to equity, which will be influenced by the values of South African society and the level of compassion that South Africans have for the poor in general and for HIV-positive people in particular. South Africa's history of oppression of one group by another and more recent adoption of a neo-liberal[40] capitalist economic system (Terreblanche 2002) makes it uncertain that this commitment will be forthcoming. According to Coburn (2000), neo-liberalism produces higher income inequality and lowered social cohesion, partly through undermining the welfare state. If the economy, the state and civil society are inextricably linked, then a neo-liberal economic system will create a more individualistic society (Coburn 2000). According to Terreblanche (2002), in South Africa's neo-liberal system, members of the upper class[41] profit handsomely from mainstream economic activity while the mainly black lower classes[42] are increasingly pauperized. While the political landscape of South Africa in the twenty-first century is far more inclusive than in the past, the social landscape remains largely unchanged. A programme of ambitious budget deficit reductions combined with personal income tax reductions has also constrained the extent to which the government can increase social spending (McIntyre, Gilson et al. 2006).

---

[39] Approximately 11 per cent of South Africans are HIV-positive.
[40] Neo-liberalism is defined as the dominance of markets and market models in the economy, where it is assumed that markets are the best and most efficient allocators of resources in production and distribution, that societies are collections of autonomous individuals motivated chiefly by economic considerations and that competition is the major market vehicle for innovation (Coburn 2000).
[41] Approximately one third of the population consisting of 4 million blacks and 4 million whites.
[42] Approximately 50 per cent of the population.

If, as Mooney (2002) argues, a country's health care system is a reflection of the values of its citizens, then the South African health care system provides additional evidence of the individualistic nature of society. In 2003/04, the share of total health care financing captured by private and public intermediaries was an estimated 62 and 38 per cent respectively, while the private sector served less than 20 per cent of the population (McIntyre, Gilson et al. 2006). This tiered health care system has implications for inequities in health. A public health care system that serves the majority of the population will tend to be of higher quality than one that mainly serves the poor. If the political support of the middle and upper classes of the public health system is lost, differences in access and quality between income groups are further reinforced (Gilson 1998). While many societies might consider South Africa's inequalities in health status and access to health care as inequitable, it is not clear that South Africans (especially the elite) share this view. Instead South Africa's elite might view the public provision of health services to the poor in terms of "we are paying for them" as opposed to "we are paying for our health service" (Mooney 2002).

Gilson (1998) argues that active steps need to be taken to create the social capital required to promote equity through a broad process of social change in which all groups play an active role. Perhaps the most critical factor is the development of an enabling state which protects and promotes the interests of the poor. However, if political will is not forthcoming, civil society can play a key role, as has been seen through the work of the Treatment Action Campaign[43] (London 2003). In this way, HIV/AIDS could be considered to be a "resource for democracy" (Fassin and Schneider 2003 p. 497). Given the magnitude of the human tragedy of HIV/AIDS, it has the potential to raise awareness of health inequalities and to advance the battle for social rights through the mobilization of activists and lay people.

To conclude, HIV/AIDS has a devastating impact on society through morbidity and mortality, but this impact is less apparent at the macroeconomic level, especially given that HIV/AIDS primarily affects the low and semi skilled segments of the population that can be replaced relatively easily from the large pool of unemployed people. On the other hand, if one considers the impact of HIV/AIDS on long-term human capital formation, the affects could be far more

[43] According to London (2003) the Treatment Action Campaign was started in 1998 as a lobbying and advocacy group for HIV-positive people in South Africa, with the intention of campaigning for greater access to treatment for all South Africans by raising public awareness and understanding about issues surrounding the availability, affordability and use of HIV treatments.

devastating than predicted by the shorter-term macroeconomic forecasts. The extent to which South Africa responds to the massive need for treatment will be determined by the values of South African society and the level of commitment to equity. Given the high cost and the potential that increased taxation might be required to finance treatment, a commitment to treating HIV will require that citizens be motivated by social solidarity as opposed to self-interest.

7 Policy implications

The approach advocated in this dissertation could be used at the international level, by national governments and by provincial governments to increase the explicitness of HIV-treatment priority setting. Explicit priority setting is argued to be more equitable than implicit rationing, where rationing could be defined as the "controlled distribution of scarce goods and services" (Bennett and Chanfreau 2005 p. 542). This section will argue with the use of two examples that the process of priority setting that is currently used in many settings is more likely to lead to implicit rationing. This is because optimistic coverage targets tend to be set in one process without adequate consideration of the costs of meeting these targets. However, because targets appear equitable, policymakers are able to avoid the politically sensitive and divisive process of determining criteria for treatment (Bennett and Chanfreau 2005).

In contrast to the approach in this thesis, patient targets and treatment interventions are normally set within one forum, and the costs of meeting the resulting decisions are established separately. For example, the WHO's "3 by 5" strategy aimed to meet 50 per cent of new AIDS cases by 2005. While the costs of meeting this target were subsequently calculated (Gutierrez, Johns et al. 2004), the feasibility of meeting these resource needs seems to have been inadequately considered – there appears to be an inadequate feedback between the target setting and the costing processes. Over (2004) has compared the projected total HIV-treatment expenditure in 2004/05 that was required to meet the "3 by 5" targets against the non HIV-treatment health care expenditure in 34 key "3 by 5" countries in 2002. In 20 out of 34 countries, the ratio of HIV-treatment expenditure to other expenditure was one-third or less, in 14 countries the ratio was greater then one-third and in 3 countries the ratio was greater than one. What this means is that projected HIV-treatment costs in 2005 would exceed total public health care expenditure in 2002 in these 3 countries. When HIV-treatment costs are compared against the GDP per capita and the number of physicians, Over (2004) concludes that 21 per cent of the 3 million ART patients live in countries

facing a feasible challenge, 41 per cent live in countries facing a substantial challenge and 38 per cent live in countries facing the greatest challenge. During the period covered by the "3 by 5"[44], the number of patients on ART in these countries increased from 400,000 in December 2003 to approximately 1 million by June 2005. Although no final report on the "3 by 5" has been released, the June 2005 report concedes that it is unlikely that the target of 3 million would be reached by December of that year (UNAIDS and WHO 2005). Thus while the plan aimed to substantially increase coverage – an equity target - it was unsurprising that these goals were not met.

More recently, in September 2005, the United Nations General Assembly endorsed a target of as close as possible to universal[45] access by 2010 for all in need (UNGASS 2006). The practical interpretation of this target is to meet 80 per cent of new AIDS cases by this date. One example of an attempt to assist countries to meet this target is the "Five country study" that WHO Geneva, WHO AFRO and WHO country offices in Mozambique, Burkina Faso, Nigeria, Tanzania and Ghana initiated during 2006[46]. Consultants in each country were contracted to calculate costs using a generic costing model (Bollinger, Boulle et al. 2006) and to assess options for the financing of these costs. An earlier version of this generic model was used in the "3 by 5" costing mentioned above. However, as before, the process is one of costing the targets as opposed to using the opportunity cost of alternative approaches and the budget constraint to set feasible targets that are in keeping with the values of society. The costing of targets provides little guidance as to whether the chosen strategies are efficient.

A growing number of publications have also recently become available that assess the cost-effectiveness of alternative strategies (Yazdanpanah, Losina et al. 2005; Bachman 2006; Badri, Cleary et al. 2006; Cleary, McIntyre et al. 2006; Goldie, Yazdanpanah et al. 2006). While cost-effectiveness/utility analyses can theoretically assess efficiency, in practice the cost-effectiveness of an intervention is judged by comparing the incremental cost per QALY gained to a societal threshold. Besides the theoretical shortcomings of this approach and its assumption that health

---

[44] In addition, while the "3 by 5" clearly missed its patient targets, one cannot suggest that the 1 million patients who did start ART were initiated solely through "3 by 5" efforts as a variety of multilateral, private and individual country donors have also been involved in the efforts to increase access to ART during this period.

[45] The same caveats to the use of this term apply.

[46] The information in this section is based on personal experience in training country teams to use the costing model and providing technical assistance in this project.

maximisation is the only appropriate social choice rule (Birch and Gafni 1992; Donaldson, Currie et al. 2002; Gafni and Birch 2006) there are additional difficulties if a country has not defined the threshold. The result has been that the majority of the publications mentioned above have taken the advice of the Commission on Macroeconomics and Health (2001) and declared an intervention "very cost-effective" if it has an incremental cost per QALY gained less than per capita gross domestic product (GDP) and "cost-effective" if it is less than twice per capita GDP. In South Africa, where per capita GDP is US$ 3,089 (Statistics South Africa 2004), either ART option could be considered very cost-effective by this criterion. However, this provides no guidance on the total budget that would be required which is particularly relevant for HIV-treatment given the number of patients in need.

If the approach advocated in this dissertation were used, it is argued that the explicitness of HIV-treatment priority setting would be advanced, with positive spin-offs for both equity and efficiency. This is because implicit rationing, which is what happens when HIV-treatment targets are missed, is less likely to be consistent and fair and is less transparent and open to review than is the case with explicit priority setting (Bennett and Chanfreau 2005). To be equitable in HIV treatment does not mean that equal access to ART for all in need is the only possibility. Given resource scarcity, unequal access might be unavoidable. Instead, if the approach to priority setting includes a fair process where opportunity costs and the values of society are key considerations, the resultant HIV-treatment strategy would be equitable. Moreover, as Rosen, Sanne et al. (2004; 2005) have argued, governments that make explicit choices and then explain and defend these choices would be more likely to sustain social cohesion and extract a socially desirable return from the large investment that is being made.

The candidate has been involved in the development of two of the generic costing models that are currently used by international organisations (Boulle, Johnson et al. 2004; Bollinger, Boulle et al. 2006). While both use a similar (although simplified) approach to calculating costs as has been taken in this dissertation, neither of these models currently presents the life years or the QALYs that are associated with these costs. Both models currently calculate the costs associated with No-ART and first and second-line ART. If a first-line only strategy were to be included as a policy choice, this would require some additional programming. Once these changes to the models have been implemented, they could be used to calculate the total costs and total life years or QALYs associated with implementing each treatment strategy to all patients in need over a specified time

period. These results are fed into the mathematical programming algorithms, which are easy to provide as an additional feature of these models.

On a note of caution, however, experience to date in training participants in the use of generic costing models has shown that while these models attempt to make the process as simple as possible for users, many participants in training sessions appear to lack the computer skills required to enter data into the models and to navigate between different sheets to read the results. Thus while including a measure of benefits and mathematical programming would not add a great deal of complexity to the approach that is currently taken (the vast majority of work would still relate to collecting and entering unit cost and utilisation data in the models to calculate costs) one would still need to contend with the critical shortage of quantitative and computer skills in many poor countries.

8 Summary

This chapter has outlined the elements of a procedurally just approach to distributing the good to HIV-positive people grounded in the notion of claims. While no particular definition of a just distribution has been identified, this chapter argues that the use of a procedurally just approach grounded in the values of society, can increase the legitimacy of the resultant distribution.

Chapter 9: Conclusion

This final chapter outlines contributions, limitations, recommendations for further research and policy recommendations arising from the work presented in this dissertation.

1 Contributions and limitations

This thesis makes a number of new contributions. On the conceptual side, a new framework for distributing the good has been proposed, with application to HIV-treatment. This framework however has relevance in health economics more generally as it could be applied to other diseases and interventions. The framework is built through considering the nature of the good of health care, and debating the pros and cons of conceiving of the good as utility, health or capabilities. The next step in constructing the framework involved considering "to whom" the good should be distributed. This section assessed the potential claims that people might have on the good which included notions of personal responsibility versus morally arbitrary bad luck in HIV acquisition, the social context, need, the impact on the health of society and the impact on the social fabric. By spelling out in detail the various claims, the framework assists the decision-maker to consider issues of equity and efficiency more holistically than if one were only to consider health care costs and outcomes. The third step in developing the framework examined how the good should be distributed. This section advocated for the use of social choice rules to illustrate the equity/efficiency trade-off in HIV-treatment allocations and argued that a fair process could legitimize the choice of the rule.

This framework is tested against empirical data and substantially revisited in Chapter 8 where the usefulness of the notion of claims is explored in more detail. Here, it is ultimately argued that procedural justice can increase legitimacy in HIV-treatment decision-making - allocations that are determined within a fair process will be equitable, provided that this process is properly followed. This means that universal access, for example, is not necessarily the only ethical outcome; given resource constraints, universal access might be out of the reach of many poor countries. In addition, if universal access were attained to generalized first-line ART for example, the next step would be to determine whether additional health care benefits were affordable.

On the empirical side, the data that have been collected and presented are relatively good given the complexity of the disease and associated interventions, and the resource-poor setting in which the evaluations were undertaken where data tended to be scarce or of poor quality. The general absence of electronic records in these settings meant that data collection took a number of years. While a small number of cost-utility analyses of HIV-treatment from the developing world have been published to date, only one other has been based on the analysis of primary data (Badri, Cleary et al. 2006) while this dissertation continues to be the only analysis that has been based on the provision of services in routine settings as opposed to within a clinical trial. The lack of data around HIV-treatment costs has meant that policymakers have relied on normative costing exercises where a number of assumptions are made about the ingredients that would be required to provide HIV-treatment such as the quantities of different types of visits, laboratory investigations and medicines. This dissertation therefore makes a significant contribution to the existing empirical knowledge about the costs and effects of HIV-treatment. The wide number of facilities included in the assessment of unit costs, the size of the ART cohort and length of follow-up are additional strengths of the data in this thesis.

There are however shortcomings in the costing approach. In particular it has been assumed that the unit cost (long-run average total cost) is a proxy for the marginal cost, which is the correct cost statistic to use in economic evaluation. Using the long-run average total cost implies that one assumes that either one is producing at the lowest point on this schedule where the marginal and average costs coincide or there are constant returns to scale. Whether or not this is the case is an empirical question which this thesis has not been able to assess, although uncertainty in the results has been assessed in sensitivity analysis.

A key analytical contribution has been made in terms of the Markov modelling process. The approach has differed from the commoner strategy in the literature which uses changes in the CD4 count over time (sometimes stratified according to the viral load) as an intermediate outcome based on which the final outcome of life years or QALYs has been modelled. In this thesis the longer follow-up time has allowed survival to be estimated directly and then extrapolated. The CD4 has only been used at base case to differentiate between patients starting ART with very low CD4 counts because empirical data indicated that these patients had a higher probability of dying during the first 6 months on treatment.

The modelling approach in this thesis has also placed additional emphasis on getting the costs right. In most developing countries, there is a clear split between first and second-line ARV regimens with important cost implications while in the developed world a number of different drug combinations can be prescribed. In the latter situation it is not always necessary to assess the probability of moving from the first-line regimen to the second because the cost differences are less significant. In addition, many analysts ignore the period of treatment failure where patients develop opportunistic infections that require inpatient care. This dissertation has introduced the technique of modelling these as transition costs incurred when a "patient" in the model transitions to the "dead" Markov state which ensures that these costs are captured at the appropriate point in time.

Further analytical strengths in this thesis have included the validation of the models in terms of technical, predictive, face and modelling process validity. In technical and predictive validity, the objective is to assess whether the model is reproducing primary data accurately; in face validity one assesses whether the model produces the output that one might expect; and in modelling process validity one assesses whether the results are comparable to results in the literature. This process of validation has therefore thoroughly assessed any uncertainty relating to the modelling process and the extrapolation of primary data.

Other forms of uncertainty were also thoroughly assessed. These included uncertainty relating to the data requirements of the study, choice of analytic method and generalizability of results. Uncertainty relating to data requirements was assessed through probabilistic sensitivity analysis while uncertainty relating to the choice of analytic method was assessed in one-way sensitivity analyses where the discount rate was varied and results were presented as life years and QALYs. However, these sensitivity analyses have not been presented in the population level results where mean (expected value) results have been used, discounted at 3 per cent per annum with outcomes expressed as QALYs. Given that the results of these probabilistic and one-way sensitivity analyses did not significantly alter the conclusions of the study at the patient-level, and that additional tables of results at the population level would be potentially confusing, this was justified. Uncertainty relating to generalizability was assessed by comparing the results in the base case scenario to secondary data, HIV-treatment guidelines and to the cost structures that exist in other parts of South Africa where there is no access to tertiary hospitals for example. This led to the construction of a generalized scenario which took these elements into account.

While patient-level models have been constructed, validated and used to assess uncertainty, the standard approaches to decision-making in cost-effectiveness/utility analyses at the patient-level have not been used in this thesis. Ordinarily, if one has a positive ICER (where a new intervention is both more effective and more costly than a comparator) then the decision of whether to implement this intervention is a value judgment that involves weighing up the extra costs and benefits under the assumption that health maximization is the appropriate social choice rule. Instead, a mathematical programming approach has been developed that is capable of integrating concerns of efficiency, equity and the costs of scaling up. It is argued that this approach simplifies HIV-treatment decision-making. Instead of facing a number of choices, decision-makers only need to be explicit about the health care budget constraint and the value base. These two elements determine the proportion of need that is met, the health gain that is achieved and the proportion of patients receiving alternative treatment strategies. While a similar mathematical programming approach has previously been recommended in the health economics literature (Birch and Gafni 1992; Stinnett and Paltiel 1996), and the utility of this type of approach has been tested with "convenience"[47] data (Anand 2003) to the best of the candidate's knowledge this has yet to be implemented in practice. However, in the implementation of this approach one limitation has been that constant returns to scale has been assumed - the costs and effects that were estimated in Khayelitsha were applied uniformly to other parts of South Africa when it is likely that costs and effects would change as the service is scaled up. This limitation has been addressed to a certain extent through the generalized scenario.

Outcomes have been presented as QALYs. Some would fault the use of the QALY approach to measuring and valuing health. While one could agree that the use of QALYs in resource allocation across diseases and interventions requires interpersonal comparability assumptions that are unrealistic, the application of the QALY to one disease, and hence one group of people, is less questionable. However, it is important that QALYs are used with care, in particular because of the assumption of constant marginal value. What this means is that decision-makers should consider the individual costs and benefits of the health care intervention as well as the population level costs and gains. Finally, although the assessment of health care costs and health care benefits has been narrow, the concept of claims has been used to include other considerations that might be relevant to participants in a fair process. This information could be presented and discussed in order to inform the choice of the health care budget and the value base.

---

[47] Data that are invented to illustrate the usefulness of the approach.

2 Recommendations for further work

While there are many aspects of this dissertation that could be taken forward, three key areas for future work are:

- Eliciting societal preferences regarding the constituents of claims for guiding allocations to HIV-treatment
- Determining the health care budget that should be made available for HIV-treatment
- Clarifying the steps that are required to implement procedurally just decision-making

This thesis has outlined a number of claims that HIV-positive people might have on the good including personal characteristics and disadvantage, need as illness or as capacity to benefit, the impact on the health of society and the impact on the social fabric. However, it is not known whether South African society would view these claims as relevant to decision-making. A potential for future research therefore lies in eliciting societal preferences regarding claims on the good. A number of different techniques could be used to elicit these preferences including discrete choice experiments (conjoint analysis) or contingent valuation (willingness to pay) surveys. Discrete choice experiments would be helpful in unpacking the views of society regarding the strengths of claims relative to each other while contingent valuation would help to determine whether society views HIV-treatment as an intervention that has positive net-benefits. Including societal preferences regarding the importance of the claims of HIV-positive people on the good is an instrumentally valuable input into the procedurally just decision-making process because it ensures that the value base of the participants at this level is as far as possible reflective of the value base of society. Eliciting societal preferences also has intrinsic value because this allows the voice of the community or society to be heard.

A crucial input into the process advocated in this thesis has been the level of the health care budget allocated to HIV-treatment. This, together with the value base determines the percentage of need that can be met within each treatment strategy and the quantity of QALYs that can be gained. However, decision-makers are not always well-equipped to make the types of resource allocation decisions required to set budgets and as a result often rely on historical budgeting approaches. Historical budgeting is particularly ill-suited to the context of HIV where required

increases in health care budgets would exceed the amounts that would be allocated through historical budgeting approaches. In addition, historical budgeting is likely to lead towards suboptimal allocations because it does not adequately consider the concepts of opportunity cost and the margin. In an ideal situation, it would be possible to apply the mathematical programming approach that has been outlined in this dissertation to the entire spectrum of independent programmes within the public health care sector. Given current data scarcity, a more pragmatic alternative is programme budgeting and marginal analysis (PBMA). This approach relies upon an advisory panel that is tasked with identifying areas where services are expanding and areas where resources could be released in order to fund the areas of expansion (Mitton and Donaldson 2004).

Finally, additional research is required regarding procedurally just decision-making before this approach can be implemented. This would include clarifying institutional levels of decision-making, developing structures to address decisions at each level, training to develop competence in fair process, learning from experience, improving the process through training and research, and developing mechanisms for enforcement (Daniels 2004). In particular, the choice of stakeholders and mechanisms to include the preferences of the community requires careful thought.

3 Policy recommendations

This dissertation has not aimed to provide conclusive recommendations regarding just allocations of the good to HIV-positive people, but has rather outlined a framework for decision-making. This section provides more concrete recommendations about how this framework could be implemented, where the aim would be to increase the explicitness of priority setting to HIV-treatment as one independent programme out of a number that are funded within the public health care system. As far as possible, each element in this process needs to respect the criteria of publicity, relevance, appeals, and enforcement in order to ensure as far as possible that the resulting choices are viewed as legitimate.

The first step involves identifying the budget that is available for HIV-treatment from all sources including donors over a ten-year period. If a ten-year period is not feasible, it might be possible to extrapolate current budgets. Based on the evidence in this thesis, it is not recommended that the decision-making time frame be reduced as this has the potential to mask the full cost implications

of HIV-treatment which could lead to the implicit rationing of treatment. If current budgets or budget plans are not readily available, it might be necessary to implement a PBMA-style exercise which means that this step is far from insignificant. However, unless the budget has been set there is a danger that the priority setting exercise will result in the setting of highly optimistic targets, once again resulting in implicit rationing.

Once the budget has been set, the next step is to choose the set of stakeholders. The aim is to be representative of all key interested parties while bearing in mind that reaching consensus is likely to become more difficult if the group is particularly large. Stakeholders also need to be able to commit to attending a number of meetings. Key stakeholders include civil society groups such as people living with HIV/AIDS and other advocacy groups, clinical experts in HIV medicine, representatives from key government departments such as the National Department of Health and National Treasury, representatives from parliament and representatives from academic groups that are familiar with priority setting and equity issues. Many countries have a National AIDS Council – stakeholders can also be drawn from this group. Once the stakeholders have been identified, the aim of the first meeting would be to clearly define the aim of the priority setting exercise and the terms of reference of the stakeholders. If possible, an introduction to the priority setting approach would be provided, including the concept of opportunity cost, the good and the implications of the different social choice rules for health gain or the proportion of need that can be met. There should also be a discussion about the value base of the group and the influence that claims on the good have on this value base. If societal preferences regarding claims have been elicited, these would be a key input at this stage – as far as possible, the values of society should determine the values adopted by the stakeholder group.

The next step would be to present the group with patient-level results regarding each of the interventions under consideration. All members need to understand the implications of each intervention in terms of average life years or QALYs gained as well as the uncertainty and range associated with these estimates. A full discussion of uncertainty would be particularly relevant given that modeling has been used and it is likely that at least some stakeholders would be skeptical about this "black box" approach. Stakeholders might also request that certain modeling assumptions are changed, that certain interventions are excluded or that other interventions are added. For example, it might be of interest to further unpack a less resource-intensive model of care that has greater reliance on nurses. There should be sufficient time between this meeting and subsequent meetings to allow for any reanalysis of data that might be required.

After this, population-level results for each social choice rule given the health care budget constraint would be presented. These results include the proportion of patients receiving each intervention and the QALYs that can be gained. It might also be necessary to present these results for a range of budgets because it is possible that the stakeholder group might recommend that the budget is reassessed. Results should also be presented for all social choice rules instead of only according to the value base of the group because the opportunity cost of one value base over another in terms of health gains foregone or unmet need would not have been apparent during earlier discussions. The ultimate aim at this point is to reach consensus on the value base – this choice will determine which interventions are prioritized. If the group recommends that the budget be expanded, it would be necessary to repeat the budget setting process. This would imply that a decision would be deferred and the group would be expected to meet again to reach consensus on the social choice rule once additional clarity on the budget had been gained.

4 Summary

Responding to the HIV-epidemic is among one of the many challenges currently facing the new South African democracy. Whether this response becomes a "resource for democracy" (Fassin and Schneider 2003 p. 497) or whether it undermines social cohesiveness within poor communities and between rich and poor communities will be partially determined by the steps that are taken during the next ten years. Implicit rationing of treatment, where optimistic treatment targets are set and subsequently missed, is less equitable and less likely to generate a socially reproducible return than explicit priority setting. To aid in the latter, a framework for decision-making that is capable of assessing equity, efficiency and the costs of scaling up in a way that makes these trade-offs transparent has been proposed and applied. The final step in the framework involves the use of procedural justice to increase the legitimacy of the chosen strategies. If decision-making around HIV-treatment is undertaken within this approach, the resulting consequences will be equitable.

References

(2003). Perspectives and Practice in Antiretroviral Treatment: Antiretroviral Therapy in Primary Health Care: Experience of the Khayelitsha Programme in South Africa, Medecins sans Frontieres South Africa, Department of Public Health University of Cape Town, Provincial Administration of the Western Cape South Africa: 1-10.

(2005). Federal Reserve Statistical Release: Foreign Exchange Rates (Annual), USA Federal Reserve Board. **2005**.

(2005). Provincial budgets and expenditure review: 2001/02-2007/08, National Treasury, Pretoria.

(2005). Uniform Fee Schedule for Patients Attending Public Hospitals. Pretoria, National Department of Health.

Abdullah, F. (2006). Lessons from the field: South Africa. XVI International AIDS Conference. Toronto, Canada.

Anand, P. (2003). "The integration of claims to health-care: a programming approach." Journal of Health Economics **22**: 731-745.

Anis, A., D. Guh, R. Hogg, X.-H. Wang, B. Yip, K. Craib, M. O'Shaughnessy, M. Schechter and J. Montaner (2000). "The Cost Effectiveness of Antiretroviral Regimens for the Treatment of HIV/AIDS." Pharmacoeconomics **4**: 393-404.

Arndt, C. and J. D. Lewis (2000). "The Macro Implications of HIV/AIDS in South Africa: A Preliminary Assessment." South African Journal of Economics **68**(5): 857-887.

Arrow, K. J. (1973). Values and Collective Decision-Making. Economic Justice. E. S. Phelps. Harmondsworth, England, Penguin Education: 117-136.

ART-LINC and ART-CC (2006). "Mortality of HIV-1-infected patients in the first year of antiretroviral therapy: comparison between low-income and high-income countries." Lancet **367**: 817-824.

Bachman, M. (2006). "Effectiveness and cost effectiveness of early and late prevention of HIV/AIDS progression with antiretrovirals or antibiotics in Southern African adults." AIDS Care **18**(2): 109-120.

Badri, M., D. Wilson and R. Wood (2002). "Effect of highly active antiretroviral therapy on incidence of tuberculosis in South Africa: a cohort study." Lancet **359**(9323): 2059-2064.

Badri, M., L. G. Bekker, C. Orrell, J. Pitt, F. Cilliers and R. Wood (2004). "Initiating highly active antiretroviral therapy in sub-Saharan Africa: an assessment of the revised World Health Organization scaling-up guidelines." AIDS **18**(8): 1159-1168.

Badri, M., S. Cleary, G. Maartens, J. Pitt, L.-G. Bekker, C. Orrell and R. Wood (2006). "When to initiate HAART in sub-Saharan Africa? A South African cost-effectiveness study." Antiviral Therapy **11**: 63-72.

Barnum, H. and J. Kutzin (1993). Public hospitals in developing countries: resource use, cost, financing. Baltimore, MD, Johns Hopkins University Press for the World Bank.

Bayoumi, A. and D. Redelmeier (1998). "Preventing Mycobacterium avium complex in patients who are using protease inhibitors: a cost-effectiveness analysis." AIDS **12**(12): 1503-1512.

Bell, C., S. Devarajan and H. Gersbach (2004). Thinking About the Long-Run Economic Costs of AIDS. The Macroeconomics of HIV/AIDS. M. Haacker. Washington, D.C., International Monetary Fund: 96-133.

Benatar, S. R. (2004). "Health Reform and the crisis of HIV/AIDS in South Africa." New England Journal of Medicine **351**(1): 81-92.

Bennett, S. and C. Chanfreau (2005). "Approaches to rationing antiretroviral treatment: ethical and equity implications." Bulletin of the World Health Organisation **83**(7): 541-547.

Bertozzi, S., J.-P. Gutierrez, M. Opuni, N. Walker and B. Schwartlander (2004). "Estimating resource needs for HIV/AIDS health care services in low-income and middle-income countries." Health Policy **69**: 189-200.

Birch, S. and C. Donaldson (1987). "Applications of Cost-Benefit Analysis to Health Care" Journal of Health Economics **6**: 211-225.

Birch, S. and A. Gafni (1992). "Cost effectiveness/utility analyses. Do current decision rules lead us to where we want to be?" Journal of Health Economics **11**: 279-296.

Black, M. and G. Mooney (2002). "Equity in Health Care from a Communitarian Standpoint." Health Care Analysis **10**: 193-208.

Blower, S., E. Bodine, J. Kahn and W. McFarland (2005). "The antiretroviral rollout and drug-resistant HIV in Africa: insights from empirical data and theoretical models." AIDS **19**: 1-14.

Bollinger, L., A. Boulle, S. Cleary and J. Stover (2006). Resource Needs for HIV/AIDS: Model for Estimating Resource Needs for Prevention, Care, Mitigation. Glastonbury, USA; Cuernavaca, Mexico; Cape Town, South Africa, WHO/UNAIDS.

Booysen, F. (2001). The Measurement of Poverty. Poverty and Chronic Diseases in South Africa. D. Bradshaw and K. Steyn. Cape Town, Medicines Research Council: 15-52.

Boulle, A. and S. Cleary (2005). Costing of four antiretroviral treatment scenarios. Cape Town, School of Public Health and Family Medicine, University of Cape Town.

Boulle, A., L. Johnson, S. Cleary and F. Abdullah (2004). The Cape Town (CT) Antiretroviral Costing Model. Version 2. Cape Town, University of Cape Town.

Brazier, J. and R. Fitzpatrick (2002). "Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions"." Health Economics **11**: 17-19.

Brazier, J., J. Roberts and M. Deverill (2002). "The estimation of a preference-based measure for health from the SF-36." Journal of Health Economics **21**: 271-292.

Briggs, A. (1995). "Handling uncertainty in the results of economic evaluation." OHE Briefing **32**.

Briggs, A. (1999). "A Bayesian Approach to Stochastic Cost-Effectiveness Analysis." Health Economics **8**: 257-261.

Briggs, A. and M. Sculpher (1998). "An Introduction to Markov Modelling for Economic Evaluation." Pharmacoeconomics **13**(4): 397-409.

Briggs, A. H. (2001). Handling uncertainty in economic evaluation and presenting the results. Economic evaluation in health care: merging theory with practice. M. Drummond and A. McGuire. Oxford, Oxford University Press.

Briggs, A., M. Sculpher and M. Buxton (1994). "Uncertainty in the economic evaluation of health care technologies: the role of sensitivity analysis." Health Economics **3**: 95-104.

British HIV Association (2003). BHIVA guidelines for the treatment of HIV-infected adults with antiretroviral therapy.

Broome, J. (1991). Weighing Goods: Equality, Uncertainty and Time. Oxford, Basil Blackwell.

Broome, J. (1993). "Qalys." Journal of Public Economics **50**: 149-167.

Brouwer, W. and M. Koopmanschap (2000). "On the economic foundations of CEA. Ladies and gentlemen, take your positions!" Journal of Health Economics **19**: 439-459.

Brouwer, W., F. Rutten and M. Koopmanschap (2001). Costing in economic evaluation. Economic Evaluation in Health Care: Merging Theory with Practice. M. Drummond and A. McGuire. Oxford, Oxford University Press.

Buchanan, A. (1989). "Assessing the Communitarian Critique of Liberalism." Ethics **99**(4): 852-882.

Bunnell, R., J. P. Ekwaru, P. Solberg, N. Wamai, W. Bikaako-Kajura, W. Were, A. Coutinho, C. Liechty, E. Madraa, G. Rutherford and J. Mermin (2006). "Changes in sexual behavior and risk of HIV transmission after antiretroviral therapy and prevention interventions in rural Uganda." AIDS **20**: 85-92.

Calmy, A., E. Klement, R. Teck, D. Berman, B. Pecoul, L. Ferradini and N. Ford (2004). "Simplifying and adapting antiretroviral treatment in resource-poor settings: a necessary step to scaling-up." AIDS **18**: 2353-2360.

Campbell, C. (1997). "Migrancy, masculine identities and AIDS: the psychosocial context of HIV transmission on the South African gold mines." Social Science and Medicine **45**(2): 273-281.

Campbell, C. (2003). Letting them Die: How HIV/AIDS prevention programmes often fail. Bloomington, Indiana University Press.

Campbell, C. and C. MacPhail (2002). "Peer education, gender and the development of critical consciousness: participatory HIV prevention by South African youth." Social Science and Medicine **55**: 331-345.

Cleary, S. and D. Ross (2002). "The 1998-2001 legal struggle between the South African government and the international pharmaceutical industry: a game-theoretic analysis." Journal of Social, Political and Economic Studies **27**: 445-494.

Cleary, S., A. Boulle, D. McIntyre and D. Coetzee (2004). Cost-effectiveness of Antiretroviral Treatment for HIV-positive adults in a South African township. Durban, Health Systems Trust: 1-67.

Cleary, S., D. McIntyre and A. Boulle (2006). "The cost-effectiveness of Antiretroviral Treatment in Khayelitsha, South Africa: a primary data analysis." Cost-effectiveness and resource allocation.

Cleary, S., O. Okorafor, W. Chitha, A. Boulle and S. Jikwana (2005). Financing Antiretroviral Treatment and Primary Health Care Services. South African Health Review 2005. P. Ijumba and P. Barron. Durban, Health Systems Trust.

Cleary, S., W. Chitha, M. Castillo, A. Boulle and D. McIntyre (2005). Health system burden of HIV/AIDS in the Western Cape. Pretoria, Joint Economics AIDS and Poverty Programme.

Clewer, A. and D. Perkins (1998). Economics for health care management. Harlow, Pearson Education Limited.

Coburn, D. (2000). "Income inequality, social cohesion and the health status of populations: the role of neo-liberalism." Social Science and Medicine **51**: 135-146.

Coetzee, D., A. Boulle, K. Hildebrand, V. Asselman, G. Van Cutsem and E. Goemaere (2004). "Promoting adherence to antiretroviral therapy: the experience from a primary care setting in Khayelitsha, South Africa." AIDS **18**(Supplement 3): S27-S31.

Coetzee, D., K. Hildebrand, A. Boulle, G. Maartens, F. Louis, V. Labatala, H. Reuter, N. Ntwana and E. Goemaere (2004). "Outcomes after two years of providing antiretroviral treatment in Khayelitsha, South Africa." AIDS **18**(6): 887-896.

Commission on Macroeconomics and Health (2001). Macroeconomics and health: investing in health for economic development. Geneva, World Health Organisation.

Conteh, L. and D. Walker (2004). "Cost and unit-cost calculations using step-down accounting." Health Policy and Planning **19**(2): 127-135.

Cook, J., E. Dasbach, P. Coplan, L. Markson, Y. Dongpin, A. Meibohm, B.-Y. Nguyen, J. Chodakewitz and J. Mellors (1999). "Modeling the Long-Term Outcomes and Costs of HIV Antiretroviral Therapy Using HIV RNA Levels: Application to a Clinical Trial." AIDS Research and Human Retroviruses **15**(6): 499-508.

Crafts, N. and M. Haacker (2004). Welfare Implications of HIV/AIDS. The Macroeconomics of HIV/AIDS. M. Haacker. Washington, D.C., International Monetary Fund: 182-197.

Creese, A. and D. Parker, Eds. (1994). Cost Analysis in Primary Health Care: A training manual for programme managers. Geneva, World Health Organization.

Crepaz, N., T. A. Hart and G. Marks (2004). "Highly Active Antiretroviral Therapy and Sexual Risk Behavior." JAMA **292**(2): 224-236.

Culyer, A. and A. Wagstaff (1993). "Equity and equality in health and health care." Journal of Health Economics **12**: 431-457.

Culyer, A. J. (2001). "Equity - some theory and its policy implications." Journal of Medical Ethics **27**: 275-283.

Currie, G., C. Donaldson and E. McIntosh (1999). Cost effectiveness analysis and the "societal approach": should the twain ever meet? Alberta, Canada, Institute of Health Economics: 1-18.

Daniels, N. (2004). How to achieve fair distribution of ARTs in 3 by 5: Fair process and legitimacy in patient selection. Geneva, World Health Organisation.

Day, C. and A. Gray (2005). Health and Related Indicators. South African Health Review 2005. P. Ijumba and P. Barron. Durban, Health Systems Trust: 248-367.

De V Graaff, J. (1973). Some Elements of Welfare Economics. Economic Justice. E. S. Phelps. Harmondsworth, England, Penguin Education: 92-113.

Department of Health (2000a). HIV/AIDS/STD Strategic Plan for South Africa 2000-2005. Pretoria, National Department of Health.

Department of Health (2000b). Recommendations for the Prevention and Treatment of Opportunistic and HIV Related Diseases in Adults. Pretoria, National Department of Health.

Department of Health (2000c). The South African Tuberculosis Control Programme: Practical Guidelines. Pretoria, National Department of Health.

Department of Health (2003), Operational Plan for Comprehensive HIV and AIDS Care, Management and Treatment for South Africa. Pretoria, National Department of Health.

Department of Health (2004). National Antiretroviral Treatment Guidelines. Pretoria, National Department of Health.

Department of Health (2007), National Strategic Plan for HIV and AIDS & STIs 2007-2011. Pretoria, National Department of Health.

Department of Health and Human Services (2005). Guidelines for the use of antiretroviral agents in HIV-1 infected adults and adolescents.

Doctor, J. N., H. Bleichrodt, J. Miyamoto, N. R. Temkin and S. Dikmen (2004). "A new and more robust test of QALYs." Journal of Health Economics **23**: 353-367.

Dolan, P., C. Gudex, P. Kind and A. Williams (1995). A social tariff for EuroQol: Results from a UK general population survey. York, Centre for Health Economics, York Health Economics Consortium, NHS Centre for Reviews & Dissemination.

Donaldson, C. (1990). "The state of the art of costing health care for economic evaluation." Community Health Studies **14**(4): 341-356.

Donaldson, C. (1998). "The (Near) Equivalence of Cost-Effectiveness and Cost-Benefit Analyses." Pharmacoeconomics **13**(4): 386-396.

Donaldson, C., G. Currie and C. Mitton (2002). "Cost effectiveness analysis in health care: contraindications." British Medical Journal **325**: 891-894.

Dorrington, R., L. Johnson and D. Budlender (2004). ASSA2002 AIDS and Demographic Models: Users Guide. Cape Town, Centre for Actuarial Research, University of Cape Town and AIDS Committee of the Actuarial Society of South Africa.

Dorrington, R., L. Johnson, D. Bradshaw and T.-J. Daniel (2006). The Demographic Impact of HIV/AIDS in South Africa. National and Provincial Indicators for 2006. Cape Town, Centre for Actuarial Research, South African Medical Research Council and Actuarial Society of South Africa.

Dowie, J. (2002). "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decision." Health Economics **11**: 1-8.

Drummond, M. and T. Jefferson (1996). "Guidelines for authors and peer reveiwers of economic submissions to the BMJ." British Medical Journal **313**: 275-83.

Dunkle, K., R. Jewkes, H. Brown, G. Gray, J. McIntyre and S. Harlow (2004). "Gender-based violence, relationship power, and risk of HIV infection in women attending antenatal clinics in South Africa." Lancet **363**: 1415-1421.

Edgar, A., S. Salek, D. Shickle and D. Cohen (1998). The Ethical QALY: Ethical Issues in Healthcare Resource Allocation. Haslemere, United Kingdom, Euromed Communications.

Egger, M., M. May, G. Chene, A. N. Phillips, B. Ledergerber, F. Dabis, D. Costagliola, A. D. Monforte, F. de Wolf, P. Reiss, J. D. Lundgren, A. C. Justice, S. Staszewski, C. Leport, R. S. Hogg, C. A. Sabin, M. J. Gill, B. Salzberger and J. A. Sterne (2002). "Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies." Lancet **360**(9327): 119-129.

Elster, J. (1992). Local Justice: How Institutions Allocate Scarce Goods and Necessary Burdens. Cambridge, Cambridge University Press.

Etard, J.-F., I. Ndiaye, M. Thierry-Mieg, N. F. N. Gueye, P. M. Gueye, I. Laniece, A. B. Dieng, A. Diouf, C. Laurent, S. Mboup, P. S. Sow and E. Delaporte (2006). "Mortality and causes of death in adults receiving highly active antiretroviral therapy in Senegal: a 7-year cohort study." AIDS **20**(8): 1181-1189.

European AIDS Clinical Society (2001). European guidelines for the clinical management and treatment of HIV infected adults in Europe.

Evans, R. and G. Stoddart (1990). "Producing health, consuming health care." Social Science & Medicine **31**: 1347-1363.

Evans, R. and G. Stoddart (1994). Producing health, consuming health care. Why are some people healthy and others not? R. Evans, M. Barer and T. Marmor. New York, de Gruyter.

Farmer, P., F. Leandre, J. S. Mukherjee, M. Claude, P. Nevil, M. C. Smith-Fawzi, S. P. Koenig, A. Castro, M. C. Becerra, J. Sachs, A. Attaran and J. Y. Kim (2001). "Community-based approaches to HIV treatment in resource-poor settings." Lancet **358**(9279): 404-9.

Fassin, D. and H. Schneider (2003). "The politics of AIDS in South Africa: beyond the controversies." British Medical Journal **326**: 495-497.

Feeny, D. (2002). "Commentary on Jack Dowie, "Decision validity should determine whether a generic or condition-specific HRQOL measure is used in health care decisions"." Health Economics **11**: 13-16.

Ferradini, L., A. Jeannin, L. Pinoges, J. Izopet, D. Odhiambo, L. Mankhambo, G. Karungi, E. Szumilin, S. Balandine, G. Fedida, M. P. Carrieri, B. Spire, N. Ford, J.-M. Tassie, P. J. Guerin and C. Brashner (2006). "Scaling up of highly active antiretroviral therapy in a rural district of Malawi: an effectiveness assessment." Lancet **367**: 1335-1342.

Filmer, D. and L. H. Pritchett (2001). "Estimating Wealth Effects without Expenditure Data - or Tears: An Application to Educational Enrollments in States of India." Demography **38**(1): 115-132.

Freedberg, K. A., E. Losina, M. C. Weinstein, A. D. Paltiel, C. J. Cohen, G. R. Seage, D. E. Craven, H. Zhang, A. D. Kimmel and S. J. Goldie (2001). "The cost effectiveness of combination antiretroviral therapy for HIV disease." N Engl J Med **344**(11): 824-31.

Freedberg, K. A., J. Scharfstein, G. R. Seage, E. Losina, M. C. Weinstein, D. E. Craven and A. D. Paltiel (1998). "The Cost-effectiveness of Preventing AIDS-Related Opportunistic Infections." JAMA **279**(2).

Furber, A. S., I. J. Hodgson, A. Desclaux and D. S. Mukasa (2004). "Barriers to better care for people with AIDS in developing countries." British Medical Journal **329**: 1281-1283.

Gafni, A. and G. W. Torrance (1984). "Risk attitude and time preference in health." Management Science **30**(440-51).

Gafni, A. and S. Birch (2006). "Incremental cost-effectiveness ratios (ICERs): The silence of the lambda." Social Science and Medicine **62**: 2091-2100.

Gerard, K. and G. Mooney (1993). "QALY League Tables: Handle with Care." Health Economics **2**: 59-64.

Gilks, C., S. Crowley, R. Ekpini, S. Gove, J. Perriens, Y. Souteyrand, D. Sutherland, M. Vitoria and T. Guerma (2006). "The WHO public-health approach to antiretroviral treatment against HIV in resource-limited settings." Lancet **368**(9534): 505-511.

Gilson, L. (1998). Re-addressing equity: the search for the holy grail? The importance of ethical processes. Eighth Annual Public Health Forum, "Reforming Health Sectors", London School of Hygiene and Tropical Medicine.

Goemaere, E., F. Louis, H. Reuter, M. Darder, V. Labatala, N. Ntwana, G. Van Cutsem, K. Hildebrand, A. Boulle and D. Coetzee (2004). Evolving experience after three years of ART in Khayelitsha. International AIDS Conference, Bangkok.

Goldie, S. J., Y. Yazdanpanah, E. Losina, M. C. Weinstein, X. Anglaret, R. P. Walensky, H. E. Hsu, A. Kimmel, C. Holmes, J. E. Kaplan and K. A. Freedberg (2006). "Cost-Effectiveness of HIV Treatment in Resource-Poor Settings - The Case of Cote d'Ivoire." New England Journal of Medicine **355**(11): 1141-1153.

Goldie, S., J. E. Kaplan and E. Losina (2002). "Prophylaxis for human immunodeficiency virus-related Pneumocystis carinii pneumonia: using simulation modeling to inform clinical guidelines." Arch Intern Med **162**: 921-928.

Goldie, S., M. Weinstein, K. M. Kuntz and K. Freedberg (1999). "The Costs, Clinical Benefits, and Cost-Effectiveness of Screening for Cervical Cancer in HIV-Infected Women." Ann Intern Med **130**(2): 97-107.

Goodman, C., P. Coleman and A. Mills (1999). "Cost-effectiveness of malaria control in sub-Saharan Africa." Lancet **354**: 378-385.

Govender, V., D. McIntyre, A. Grimwood and G. Maartens (2000). The Costs and Perceived Quality of Care for People Living with HIV/AIDS in the Western Cape Province in South Africa, Partnerships for Health Reform.

Gray, A. (2005). NDOH ARV Tenders Awarded, Drug Information & Policy Network. **2005**.

Greener, R. (2004). The Impact of HIV/AIDS on Poverty and Inequality. The Macroeconomics of HIV/AIDS. M. Haacker. Washington, D.C., International Monetary Fund: 167-181.

Guinness, L., G. Arthur, S. Bhatt, G. Achiya, S. Kariuki and C. Gilks (2002). "Costs of hospital care for HIV-positive and HIV-negative patients at Kenyatta National Hospital, Nairobi, Kenya." AIDS **16**: 901-908.

Gutierrez, J.-P., B. Johns, T. Adam, S. Bertozzi, T. Tan-Torres Edejer, R. Greener, C. Hankins and D. Evans (2004). "Achieving the WHO/UNAIDS antiretroviral treatment 3 by 5 goal: what will it cost?" Lancet **364**: 63-64.

Haacker, M. (2004). HIV/AIDS: The Impact on the Social Fabric and the Economy. The Macroeconomics of HIV/AIDS. M. Haacker. Washington, D.C., International Monetary Fund: 41-95.

Haile, B. (2000). The Costs of Adult Inpatient Care for HIV Disease at GF Jooste Hospital [dissertation]. Health Economics Unit. Cape Town, University of Cape Town.

Hall, E. and J. Erasmus (2005). Medical Practitioners and Nurses. HRD Review 2003, Human Sciences Research Council.

Hansen, K., G. Chapman, I. Chitsike, O. Kasilo and G. Mwaluko (2000). "The costs of HIV/AIDS care at government hospitals in Zimbabwe." Health Policy and Planning **15**(4): 432-440.

Harling, G., R. Wood and E. J. Beck (2005). "Efficiency of Interventions in HIV Infection, 1994-2004." Dis Manage Health Outcomes **13**(6): 371-394.

Harries, A., E. Schouten and E. Libamba (2006). "Scaling up antiretroviral treatment in resource-poor settings." Lancet **367**(9525): 1870-1872.

Harris, J. and S. Holm (1995). "Is there a moral obligation not to infect others?" British Medical Journal **311**: 1215-1217.

Hendriks, J. C., G. A. Satten, I. M. Longini, H. A. van Druten, P. T. A. Shellekens, R. A. Coutinho and G. J. van Griensven (1996). "Use of immunological markers and continuous-time Markov models to estimate progression of HIV infection in homosexual men." AIDS **10**: 649-656.

Hickey, A. (2004). New allocations for ARV treatment: An analysis of 2004/5 national budget from an HIV/AIDS perspective. IDASA Occasional Papers. Cape Town, IDASA - Budget Information Service: 1-43.

Ho, D. (1995). "Time to hit HIV, early and hard." New England Journal of Medicine **333**: 450-451.

Hoffmann, T. and H. Brunner (2004). "Model for simulation of HIV/AIDS and cost-effectiveness of preventing non-tuberculous mycobacterial (MAC) disease." Eur J Health Econ **5**(2): 129-135.

Hogg, R. S., B. Yip, K. J. Chan, E. Wood, K. J. Craib, M. V. O'Shaughnessy and J. S. Montaner (2001). "Rates of disease progression by baseline CD4 cell count and viral load after initiating triple-drug therapy." JAMA **286**(20): 2568-77.

Holmberg, S., F. Palella, K. Lichtenstein and D. Havlier (2004). "The case for earlier treatment of HIV infection." Clin Infect Dis **39**: 1699-1705.

Jacobs, P. and J.-F. Baladi (1996). "Biases in cost measurement for economic evaluation studies in health care." Health Economics **5**: 525-529.

Jaffar, S., T. Govender, A. Garrib, T. Welz, H. Grosskurth, P. G. Smith, H. Whittle and M. L. Bennish (2005). "Antiretroviral treatment in resource-poor settings: public health research priorities." Tropical Medicine and International Health **10**(4): 295-299.

Jelsma, J., E. MacLean, J. Hughes, X. Tinise and M. Darder (2005). "An investigation into the Health Related Quality of Life of individuals living with HIV who are receiving HAART." AIDS Care **17**(5): 579-588.

Jelsma, J., K. Hansen, W. De Weerdt and P. De Cock (2002). How do Zimbabweans value health states? The burden of disease due to disability in a high density area of Harare, Zimbabwe. J. Jelsma, Katholieke Universiteit Leuven: 64-79.

Johnson, L. and D. Budlender (2002). HIV risk factors: a review of the demographic, socio-economic, biomedical and behavioural determinants of HIV prevalence in South Africa. Cape Town, Centre for Actuarial Research, University of Cape Town.

Karstaedt, A. S., T. C. M. Lee, A. W. Kinghorn and H. Schneider (1996). "Care of HIV-infected adults at Baragwanath Hospital, Soweto: Part II. management and costs of inpatients." S Afr Med J **86**(11): 1490-1493.

Kinghorn, A. W., T. C. M. Lee, Karstaedt.A.S., B. Khounane and H. Schneider (1996). "Care of HIV-infected adults at Baragwanath Hospital, Soweto: Part I. Clinical management and costs of outpatient care." S Afr Med J **86** (11): 1484-1489.

Kober, K. and W. Van Damme (2004). "Scaling up access to antiretroviral treatment in southern Africa: who will do the job?" Lancet **364**: 103-107.

Koch, R. J. (2003). The Quantum Year Book. Port Elizabeth, Van Zyl, Rudd & Associates.

Koenig, S. P., F. Leandre and P. E. Farmer (2004). "Scaling-up HIV treatment programmes in resource-limited settings: the rural Haiti experience." AIDS **18 (suppl 3)**: s21-s25.

Kopelman, L. M. (2002). "If HIV/AIDS is Punishment, Who is Bad?" Journal of Medicine and Philosophy **27**(2): 231-243.

Kuntz, K. M. and M. C. Weinstein (1997). Modelling in economic evaluation. Methods for the Economic Evaluation of Health Care Programs. M. Drummond, B. O'Brien, G. Stoddart and G. W. Torrance. Oxford, Oxford University Press: 141-171.

Lamont, J., Ed. (2003). Distributive Justice. The Stanford Encyclopedia of Philosophy (Fall 2003 Edition).

Lane, H. and J. Neaton (2003). "When to start therapy for HIV infection: a swinging pendulum in search of data." Ann Intern Med **138**: 680-681.

Laubscher, P., B. Smit and L. Visagie (2001). The Macroeconomic Impact of HIV/AIDS in South Africa. Research Note No. 10. Stellenbosch, Bureau for Economic Research.

Laurent, C., C. Kouanfack and S. Koulla-Shiro (2004). "Effectiveness and safety of a generic fixed-dose combination of nevirapine, stavudine, and lamivudine in HIV-1 infected adults in Cameroon: open-label multicentre trial." Lancet **364**: 29-34.

Lipsey, R. G., P. N. Courant, D. D. Purvis and P. O. Steiner (1993). Economics. New York, HarperCollins College Publishers.

Little, S. J., S. Holte, J.-P. Routy, E. S. Daar, M. Markowitz, A. C. Collier, R. A. Koup, J. W. Mellors, E. Connick, B. Conway, M. Kilby, L. Wang, J. M. Whitcomb, N. W. Hellman and D. D. Richman (2002). "Antiretroviral-drug resistance among patients recently infected with HIV." New England Journal of Medicine **347**(6): 385-394.

Litva, A., J. Coast, J. Donovan, J. Eyles, M. Shepherd, J. Tacchi, J. Abelson and K. Morgan (2002). "'The public is too subjective': public involvement at different levels of health-care decision making." Social Science and Medicine **54**: 1825-1837.

Loewenson, R. and D. McCoy (2004). "Access to antiretroviral treatment in Africa." British Medical Journal **328**: 241-242.

London, L. (2003). Can human rights serve as a tool for equity? Equinet Policy Series. R. Loewenson. Harare, EQUINET. **14:** 1-39.

Luce, B., W. Manning, J. Siegel and J. Lipscomb (1996). Estimating Costs in Cost-Effectiveness Analysis. Cost-effectiveness in health and medicine. M. Gold, J. Siegel, L. Russell and M. Weinstein. New York, Oxford University Press.

Luchini, S., B. Cisse, S. Duran, M. De Cenival, C. Comiti, M. Gaudry and J.-P. Moatti (2003). Decrease in prices of antiretroviral drugs for developing countries: from political "philanthropy" to regulated markets? Economics of AIDS and access to HIV/AIDS care in developing countries. Issues and challenges. J.-P. Moatti, B. Coriat, Y. Souteyrand, T. Barnett, J. Dumoulin and Y.-A. Flori. Paris, ANRS: 169-213.

Lurie, M. (2000). "Migration and AIDS in southern Africa: a review." South African Journal of Science **96**: 343-347.

Macklin, R. (2004). Ethics and equity in access to HIV treatment - 3 by 5 initiative. Geneva, World Health Organisation.

Maman, S., J. Campbell, M. D. Sweat and A. C. Gielen (2000). "The intersections of HIV and violence: directions for future research and interventions." Social Science and Medicine **50**: 459-478.

Marseille, E., J. Saba, S. Muyingo and J. G. Kahn (2006). "The costs and benefits of private sector provision of treatment to HIV-infected employees in Kampala, Uganda." AIDS **20**(6): 907-914.

Martin, A. J., P. Glasziou, R. Simes and T. Lumley (2000). "A comparison of standard gamble, time trade-off, and adjusted time trade-off scores." International Journal of Technology Assessment in Health Care **16**(1): 137-147.

Mbeki, T. (2004). Address of the President of South Africa, Thabo Mbeki, to the first joint sitting of the third democratic Parliament, Cape Town, South African Government Information.

McIntyre, D. and L. Gilson (2002). "Putting equity in health back onto the social policy agenda: Experience from South Africa." Social Science and Medicine **54**: 1637-1656.

McIntyre, D., L. Gilson, H. Wadee, M. Thiede and O. Okorafor (2006). "Commercialisation and extreme inequality in health: the policy challenges in South Africa." Journal of International Development **18**: 435-446.

Miller, D. (1992). "Distributive Justice: What the People Think." Ethics **102**: 555-593.

Mills, E. J., J. B. Nachega, I. Buchan, J. Orbinski, A. Attaran, S. Singh, B. Rachlis, P. Wu, C. Cooper, L. Thabane, K. Wilson, G. H. Guyutt and D. R. Bangsberg (2006). "Adherence to Antiretroviral Therapy in Sub-Saharan Africa and North America: A Meta-analysis." JAMA **296**(6): 679-690.

Miners, A. H., C. A. Sabin, P. Trueman, M. Youle, A. Mocroft, M. Johnson and E. J. Beck (2001). "Assessing the cost-effectiveness of HAART for adults with HIV in England." HIV Med **2**(1): 52-8.

Mitton, C. and C. Donaldson (2004). "Health care priority setting: principles, practice and challenges." Cost Effectiveness and Resource Allocation **2**(3).

Moatti, J. P., I. N'Doye, S. M. Hammer, P. Hale and M. D. Kazatchkine (2003). "Antiretroviral treatment for HIV infection in developing countries: an attainable new paradigm." Nature Medicine **9**(12): 1449-1452.

Moatti, J., B. Spire and M. Kazatchkine (2004). "Drug resistance and adherence to HIV/AIDS antiretroviral treatment: against a double standard between the north and the south." AIDS **18 (suppl)**: s55-s61.

Mooney, G. (1996). "And now for vertical equity? Some concerns arising from Aboriginal health in Australia." Health Economics **5**: 99-103.

Mooney, G. (1998). ""Communitarian claims" as an ethical basis for allocating health care resources." Social Science and Medicine **47**(9): 1171-1180.

Mooney, G. (2002). Access and service delivery issues. Health Policy Round Table, Canberra, Australia.

Mooney, G. (2003). Economics, Medicine and Health Care. Harlow, Pearson Education Limited.

Mooney, G. (2005). "Communitarian claims and community capabilities: furthering priority setting?" Social Choice and Welfare **60**(247-255).

Mooney, G. and S. Jan (1997). "A Second Opinion: Cost-utility analysis and varying preferences for health." Health Policy **41**: 201-205.

Mooney, G., S. Jan and V. Wiseman (2002). "Staking a claim for claims: a case study of resource allocation in Australian Aboriginal health care." Social Science & Medicine **54**: 1657-1667.

Mooney, G., S. Jan and V. Wiseman (2002). "Staking a claim for claims: a case study of resource allocation in Australian Aboriginal health care." Social Science & Medicine **54**: 1657-1667.

Moore, R. and J. Bartlett (1996). "Combination antiretroviral therapy in HIV infection: an economic perspective." Pharmacoeconomics **10**(2): 109-113.

Moore, R. and R. Chaisson (1997). "Cost-utility analysis of prophylactic treatment with oral ganciclovir for cytomegalovirus retinitis." J Acquir Immune Defic Syndr Hum Retrovirol **16**(1): 15-21.

Moore, R., J. Hidalgo, J. Bareta and R. Chaisson (1994). "Zidovudine therapy and resource utilization in AIDS." J Acquir Immune Defic Syndr **7**(4): 349-354.

MSF (2002). Untangling the Web of Price Reductions, Medecins Sans Frontieres.

Mugyeni, P. (2004). "Highly active antiretroviral therapy: we need to scale up its use and reach with existing facilities in poor countries." British Medical Journal **329**: 1118-1119.

Nachega, J., M. Hislop, D. Dowdy, M. Lo, S. Omer, L. Regensberg, R. Chaisson and G. Maartens (2006). "Adherence to highly active antiretroviral therapy assessed by pharmacy claims predicts survival in HIV-infected South African adults." J Acquir Immune Defic Syndr **43**(1): 78-84.

Nattrass, N. (2002). Unemployment, Employment and Labour Force Participation in Khayelitsha/Mitchell's Plain. Cape Town, Centre for Social Science Research.

Nattrass, N. (2004). The Moral Economy of AIDS in South Africa. Cambridge, Cambridge University Press.

Nord, E., J.-L. Pinto, J. Richardson, P. Menzel and P. Ubel (1999). "Incorportating societal concerns for fairness in numerical valuations of health programmes." Health Economics **8**(1): 25-39.

O'Hagan, A. and B. R. Luce (2003). A Primer on Bayesian Statistics in Health Economics and Outcomes Research, MEDTAP International.

O'Hagan, A., C. McCabe, R. Akehurst, A. Brennan, A. Briggs, K. Claxton, E. Fenwick, D. Fryback, M. Sculpher, D. Spiegelhalter and A. Willan (2005). "Incorporation of Uncertainty in Health Economic Modelling Studies." Pharmacoeconomics **23**(6): 529-536.

O'Keefe, E. and R. Wood (1996). "The impact of human immunodeficiency virus (HIV) infection on quality of life in a multiracial South African population." Quality of Life Research **5**: 275-280.

Oliveira-Cruz, V., C. Kurowski and A. Mills (2003). "Delivery of priority health services: searching for synergies within the vertical versus horizontal debate." Journal of International Development **15**: 67-86.

Olsen, J. A., J. Richardson, P. Dolan and P. Menzel (2003). "The moral relevance of personal characteristics in setting health care priorities." Social Science & Medicine **57**: 1163-1172.

Over, M. (2004). Impact of the HIV/AIDS Epidemic on the Health Sectors of Developing Countries. The Macroeconomics of HIV/AIDS. M. Haacker. Washington, D.C., International Monetary Fund: 311-344.

Paltiel, A. and K. Freedberg (1998). "The Cost-Effectiveness of Preventing Cytomegalovirus Disease in AIDS Patients." Interfaces **28**(3): 34-51.

Paltiel, A., S. Goldie, E. Losina, M. Weinstein, G. Seage, A. Kimmel, H. Zhang and K. Freedberg (2001). "Preevaluation of Clinical Trial Data: The Case of Preemptive Cytomegalovirus Therapy in Patients with Human Immunodeficiency Virus." Clin Infect Dis **32**(5): 783-793.

Phillips, A., A. Lepri, F. Lampe, M. Johnson and C. Sabin (2003). "When should antiretroviral therapy be started for HIV infection? Interpreting the evidence from observational studies." AIDS **17**: 1863-1869.

Pienaar, D., L. Myer, S. Cleary, D. Coetzee, D. Michaels, K. Cloete, H. Schneider and A. Boulle (2006). Models of Care for Antiretroviral Service Delivery. Cape Town, University of Cape Town: 1-101.

Pitt, J., M. Badri and R. Wood (2005). Changes in quality of life in a South African antiretroviral programme. International AIDS Conference, Bangkok, Thailand.

Porco, T., J. Martin, K. Page-Shafer, A. Cheng, E. Charlebois, R. Grant and D. Osmond (2004). "Decline in HIV infectivity following the introduction of highly active antiretroviral therapy." AIDS **18**: 81-88.

Posnett, J. and S. Jan (1996). "Indirect cost in economic evaluation: the opportunity cost of unpaid inputs." Health Economics **5**: 13-23.

Post, F. A., R. Wood and G. Maartens (1996). "CD4 and total lymphocyte counts as predictors of HIV disease progression." QJM **89**(7): 505-8.

Provincial Government of the Western Cape (2004). Antiretroviral Treatment Protocol - Western Cape (based on National Treatment Guidelines): Version 2.

Quattek, K. (2000). Economic Impact of AIDS on the South African Economy. Johannesburg, A study by Wefa SA commissioned by INB Barings.

Rankin, W. W., S. Brennan, E. Schell, J. Laviwa and S. H. Rankin (2005). "The Stigma of Being HIV-Positive in Africa." PLoS Medicine **2**(8).

Rawls, J. (1971). A Theory of Justice. Oxford, Oxford University Press.

Rice, T. (2003). The economics of health reconsidered. Chicago, Health Administration Press.

Richardson, J. and J. McKie (2005). "Empiricism, ethics and orthodox economic theory: what is the appropriate basis for decision-making in the health sector?" Social Science & Medicine **60**: 265-275.

Richter, A., B. Hauber, K. Simpson, A. Mauskopf and D. Yin (2002). "A Monte Carlo Simulation for Modelling Outcomes of AIDS Treatment Regimens." Pharmoeconomics **20**(4): 214-224.

Roemer, J. (1996). Theories of Distributive Justice. Cambridge, Harvard University Press.

Roemer, J. E. (1993). "A Pragmatic Theory of Responsibility for the Egalitarian Planner." Philosophy and Public Affairs **22**(2): 146-166.

Rosen, S., I. Sanne, A. Collier and J. L. Simon (2004). Hard choices: rationing antiretroviral therapy for HIV/AIDS in Africa. Lancet.

Rosen, S., I. Sanne, A. Collier and J. L. Simon (2005). "Rationing Antiretroviral Therapy for HIV/AIDS in Africa: Choices and Consequences." PLoS Medicine **2**(11): 1098-1104.

Ryan, M. (1999). "Using conjoint analysis to take account of patient preferences and go beyond health outcomes: an application to in vitro fertilisation." Social Science and Medicine **48**: 535-546.

Schackman, B. R., K. A. Freedberg, M. C. Weinstein, P. E. Sax, E. Losina, S. M. Hong Zang and S. J. Goldie (2002). "Cost-effectiveness Implications of the Timing of Antiretroviral Therapy in HIV-Infected Adults." Arch Intern Med **162**: 2478-2486.

Schackman, B. R., S. J. Goldie, M. C. Weinstein, E. Losina, H. Zhang and K. A. Freedberg (2001). "Cost-effectiveness of earlier initiation of antiretroviral therapy for uninsured HIV-infected adults." Am J Public Health **91**(9): 1456-63.

Schneider, H. and J. Stein (2001). "Implementing AIDS policy in post-apartheid South Africa." Social Science and Medicine **52**: 723-731.

Schneider, H., D. Blaauw, L. Gilson, N. Chabikuli and J. Goudge (2006). "Health Systems and Access to Antiretroviral Drugs for HIV in Southern Africa: Service Delivery and Human Resources Challenges." Reproductive Health Matters **14**(27): 12-23.

Schneider, M., M. Zwahlen and M. Egger (2004). Natural history and mortality in HIV-positive individuals living in resource-poor settings: A literature review. UNAIDS Obligation HQ/03/463871. Geneva, UNAIDS.

Sen, A. (1982). Choice, Welfare and Measurement. Oxford, Basil Blackwell.

Sen, A. (1992). Inequality Reexamined. Cambridge, MA, Harvard University Press.

Sen, A. (1999). Development as Freedom. New York, Anchor Books.

Sendi, P. P., B. A. Craig, D. Pfluger, A. Gafni and H. C. Bucher (1999). "Systematic validation of disease models for pharmoeconomic evaluations." Journal of Evaluation in Clinical Practice 5(3): 283-295.

Sendi, P. P., B. A. Craig, G. Meier, D. Pfluger, A. Gafni, M. Opravil, M. Battegay and H. C. Bucher (1999). "Cost-effectiveness of azithromycin for preventing *Mycobacterium avium* complex infection in HIV-positive patients in the era of highly active antiretroviral therapy." Journal of Antimicrobial Chemotherapy **44**: 811-817.

Sendi, P. P., H. C. Bucher, T. Harr, B. A. Craig, M. Schwietert, D. Pluger, A. Gafni and M. Battegay (1999). "Cost effectiveness of highly active antiretroviral therapy in HIV-infected patients." AIDS **13**(1115-1122).

Sendi, P., A. Gafni and S. Birth (2002). "Opportunity costs and uncertainty in the economic evaluation of health care interventions." Health Economics **11**: 23-31.

Shaw, J. W., J. A. Johnson and S. J. Coons (2005). "US Valuation of the EQ-5D Health States: Development and Testing of the D1 Valuation Model." Medical Care **43**(3): 203-220.

Shisana, O. and L. Simbayi (2002). Nelson Mandela/HSRC Study of HIV/AIDS. Cape Town, Human Sciences Research Council: 1-121.

Shisana, O., E. Hall and K. Maluleke (2002). The Impact of HIV/AIDS on the Health Sector: National Survey of Health Personnel, Ambulatory and Hospitalised Patients and Health Facilities. Pretoria, Human Sciences Research Council, Medical University of South Africa, Medical Research Council.

Shisana, O., T. M. Rehle, L. C. Simbayi, W. Parker, K. Zuma, C. Connolly, S. Jooste and V. Pillay (2005). South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey, 2005. Cape Town, HSRC Press.

Siegel, J., M. Weinstein and G. Torrance (1996). Reporting Cost-Effectiveness Studies and Results. Cost-effectiveness in health and medicine. M. Gold, J. Siegel, L. Russell and M. Weinstein. New York, Oxford University Press.

Sinanovic, E., K. Floyd, L. Dudley and V. Azevedo (2000). Cost and cost-effectiveness of community-based and clinic-based supervision of tuberculosis treatment in Cape Town, South Africa. Cape Town, Health Economics Unit, University of Cape Town.

Sinanovic, E., K. Floyd, L. Dudley, V. Azevedo, R. Grant and D. Maher (2003). "Cost and cost-effectiveness of community-based care for tuberculosis in Cape Town, South Africa." Int J Tuberc Lung Dis 7(9): S56-S62.

Sinnott-Armstrong, W. (2003). "Consequentialism." Stanford Encyclopedia of Philosophy http://plato.stanford.edu/entries/consequentialism/. June 2005.

Smit, B. W., L. Ellis and P. Laubscher (2006). The macroeconomic impact of HIV/AIDS under alternative intervention scenarios (with specific reference to ART) on the South African economy. Stellenbosch, Bureau for Economic Research.

Smith De Scherif, T., J. H. Schoeman, S. Cleary, G. Meintjies, K. Rebe and G. Maartens (2005). The costs of inpatient care at GF Jooste Hospital.

Sonnenberg, F. A. and J. R. Beck (1993). "Markov models in medical decision making: a practical guide." Med.Decis.Making **13**(4): 322-338.

Spiegelhalter, D., J. P. Myles, D. R. Jones and K. R. Abrams (1999). "Methods in health service research: An introduction to bayesian methods in health technology assessment." British Medical Journal **319**: 508-512.

Statistics South Africa (2004). Gross Domestic Product - First Quarter 2004: 1-42.

Sterling, T., R. Chaisson and M. RD (2001). "CD4 T-lymphocytes, and clinical response to highly active antiretroviral therapy." AIDS **23**: 2251-2257.

Sterling, T., R. Chaisson, J. Keruly and R. Moore (2003). "Improved outcomes with earlier initiation of highly active antiretroviral therapy among Human Immunodeficiency Virus-infected patients who achieve durable virologic suppression: longer follow-up of an observational cohort study." J Infect Dis **188**: 1659-1665.

Stewart, R. and M. Loveday (2005). The operational plan: implementation of the antiretroviral therapy component. South African Health Review 2005. P. Ijumba and P. Barron. Durban, Health Systems Trust: 224-246.

Stillwaggon, E. (2002). "HIV/AIDS in Africa: Fertile Terrain." Journal of Development Studies **38**(6): 1-22.

Stinnett, A. A. and A. D. Paltiel (1996). "Mathematical programming for the efficient allocation of health care resources." Journal of Health Economics **15**: 641-653.

Sugden, R. (1993). "Welfare, Resources, and Capabilities: A Review of *Inequality Reexamined* by Amartya Sen." Journal of Economic Literature **XXXI**: 1947-1962.

Tebas, P., K. Henry, R. Nease, R. Murphy, J. Phair and W. Powderly (2001). "Timing of antiretroviral therapy. Use of Markov modeling and decision analysis to evaluate the long-term implications of therapy." AIDS **15**(5): 591-599.

Terreblanche, S. (2002). A History of Inequality in South Africa. 1652-2002. Pietermaritzburg, University of Natal Press.

Thiede, M., N. Palmer and S. Mbatsha (2005). South Africa: Who goes to the public sector for HIV/AIDS counselling and testing? Reaching the Poor with Health, Nutrition, and

Population Services. D. Gwatkin, A. Wagstaff and A. Yazbeck. Washington DC, World Bank.

Torrance, G. and D. Feeny (1989). "Utilities and Quality-Adjusted Life Years." International Journal of Technology Assessment in Health Care **5**: 559-575.

Torrance, G. W., D. Feeny and W. Furlong (2001). "Visual Analog Scales: Do They Have a Role in the Measurement of Preferences for Health States?" Medical Decision Making **21**(4): 329-334.

Ubel, P. A., M. L. DeKay, J. Baron and D. A. Asch (1996). "Cost-effectiveness Analysis in Setting of Budget Constraints. Is It Equitable?" New England Journal of Medicine **334**(18): 1174-1177.

UNAIDS and WHO (2005). AIDS epidemic update. Geneva, UNAIDS/WHO.

UNAIDS and WHO (2005). Progress on Global Access to HIV Antiretroviral Therapy: An update on "3 by 5". Geneva, UNAIDS and WHO: 1-34.

UNGASS (2006). Towards universal access: assessment by the Joint United Nations Programme on HIV/AIDS on scaling up HIV prevention, treatment, care and support. Geneva, United Nations General Assembly: 1-21.

Van Niekerk, A. A. (2001). "Moral and Social Complexities of AIDS in Africa." Journal of Medicine and Philosophy **27**(2): 143-162.

van Sighem, A., M. van de Wiel, A. Ghani, M. Jambroes, P. Reiss and I. Gyssens (2003). "Mortality and progression to AIDS after starting highly active antiretroviral therapy." AIDS **17**: 2227-2236.

Velasco-Hernandez, J. X., H. B. Gershengorn and S. M. Blower (2002). "Could widespread use of combination antiretroviral therapy eradicate HIV epidemics?" The Lancet Infectious Diseases **2**: 487-93.

Viner, J. (1973). Bentham and J.S. Mill: The Utilitarian Background. Economic Justice. E. S. Phelps. Harmondsworth, England, Penguin Education.

von Neumann, J. and O. Morgenstern (1944). Theory of Games and Economic Behavior. Princeton, Princeton University Press.

Wagstaff, A. (1991). "QALYs and the equity-efficiency trade-off." Journal of Health Economics **10**: 21-41.

Wailoo, A. and P. Anand (2005). "The nature of procedural preferences for health-care rationing decisions." Social Science & Medicine **60**: 223-236.

Walker, D. and L. Kumaranayake (2002). "How to do (or not to do). Allowing for differential timing in cost analyses: discounting and annualization." Health Policy and Planning 17(1): 112-118.

Walzer, M. (1983). Spheres of Justice: A Defence of Pluralism & Equality. Oxford, England, Basil Blackwell.

Whitehead, M. (1992). "The concepts and principles of equity and health." International Journal of Health Services **22**(3): 429-445.

WHO (1993). "Proposed World Health Organization Staging System for HIV Infection and Disease: preliminary testing by an international collaborative cross-sectional study." AIDS 7: 711-718.

WHO (2002). Scaling up antiretroviral therapy in resource limited settings: guidelines for a public health approach. Geneva, World Health Organisation.

WHO (2003). Scaling up antiretroviral therapy in resource-limited settings: treatment guidelines for a public health approach. Geneva, World Health Organisation.

WHO and UNAIDS (2004). Consultation on ethics and equitable access to treatment and care for HIV/AIDS. Geneva, World Health Organisation/Joint United Nations Programme on HIV/AIDS.

WHO (2006). Antiretroviral therapy of HIV infection in adults and adolescents in resource-limited settings: recommendations for a public health approach (2006 revision). Geneva, WHO.

Williams, A. (1996). "QALYs and ethics: a health economist's perspective." Social Science and Medicine **43**(12): 1795-1804.

Williams, A. and R. Cookson (2000). Equity in Health. Handbook of Health Economics. A. Culyer and J. Newhouse. Amsterdam, Elsevier: 1863-1910.

Williams, B., D. Gilgen, C. Campbell, D. Taljaard and C. MacPhail (2000). The natural history of HIV/AIDS in South Africa: a biomedical and social survey in Carletonville. Pretoria, Council for Scientific and Industrial Research.

Wiseman, V., G. Mooney, G. Berry and K. Tang (2003). "Involving the general public in priority setting: experiences from Australia." Social Science and Medicine **56**: 1001-1012.

Wood, K., K. Maforah and R. Jewkes (1998). ""He forced me to love him": putting violence on the adolescent sexual health agenda." Social Science and Medicine **47**: 233-242.

Yazdanpanah, Y., E. Losina, X. Anglaret, S. J. Goldie, R. P. Walensky, M. C. Weinstein, S. Toure, H. E. Smith, J. E. Kaplan and K. A. Freedberg (2005). "Clinical impact and cost-

effectiveness of co-trimoxazole prophylaxis in patients with HIV/AIDS in Cote d'Ivoire: a trial-based analysis." AIDS **19**(12): 1299-1308.

Yazdanpanah, Y., S. Goldie, A. Paltiel, E. Losina, L. Coudeville, M. Weinstein, Y. Gerard, A. Kimmel, H. Zhang, R. Salamon, Y. Mouton and K. Freedberg (2003). "Prevention of Human Immunodeficiency Virus-Related Opportunistic Infections in France: A Cost-Effectiveness Analysis." Clin Infect Dis **36**(1): 86-96.

Yeni, P., S. Hammer, M. Hirsch and e. al (2004). "Treatment for adult HIV infection: 2004 recommendations of the International AIDS Society-USA Panel." JAMA **292**(251-265).

Young, A. (2005). "The gift of the dying: the tragedy of AIDS and the welfare of future African generations." The Quarterly Journal of Economics **CXX**(2): 423-466.

Appendix A – WHO staging system

Clinical stage I:

1. asymptomatic infection
2. persistent generalized lymphadenophathy
3. acute retroviral infection

Clinical stage II:

1. unintentional weight loss < 10 per cent of body weight
2. minor mucocutaneous manifestations
3. herpes zoster within the previous 5 years
4. recurrent upper respiratory tract infections

Clinical stage III:

1. unintentional weight loss > 10 per cent of body weight
2. chronic diarrhoea > 1 month
3. prolonged fever > 1 month
4. oral candidiasis
5. oral hairy leukoplakia
6. pulmonary TB
7. severe bacterial infections
8. vulvovaginal candidiasis, chronic

Clinical stage IV (AIDS):

1. HIV wasting syndrome
2. Pneumocystis carnii pneumonia (PCP)
3. toxoplasmosis
4. cryptosporidiosis with diarrhoea
5. isosporiasis with diarrhoea
6. cryptococcosis extrapulmonary
7. cytomegalovirus disease of an organ other than liver, spleen or lymph node
8. herpes simplex virus infection, mucocutaneous, (> 1 month) or visceral (any duration)
9. progressive multifocal leukoencephalopathy

10. any disseminated endemic mycosis
11. candidiasis of the oesophagus, trachea, bronchi or lungs
12. atypical mycobacteriosis, disseminated (MAI)
13. non-typhoid salmonella septicaemia (SAL)
14. extrapulmonary TB
15. lymphoma
16. Kaposi's sarcoma
17. HIV encephalopathy

Source: WHO (1993)

Appendix B – summary of cost-effectiveness/utility analyses

Table 48: Summary of cost-effectiveness/utility analyses - undiscounted

Reference	Intervention and characteristics of patients	Setting	Currency, price year and discount rate	Lifetime costs in US$	Lifetime outcome
Undiscounted - No ART					
Moore, Hidalgo et al 1994	No-ART for patients with AIDS	United States of America	US$, 1990, ?	31,300	0.64 life years
Sendi, Bucher et al 1999	No-ART for patients representative of the Swiss HIV cohort	Switzerland	Swiss francs, 1997, 4%	140,065	6.64 life years
Yazdanpanah, Losina et al 2005	Patients enter care with CD4=331 and receive no prophylaxis and No-ART	Cote d'Ivoire	US$, 2000, 3%	1,260	3.2 life years
Yazdanpanah, Losina et al 2005	Patients enter care with CD4=331 and receive cotrimoxazole prophylaxis and No-ART	Cote d'Ivoire	US$, 2000, 3%	1,320	3.5 life years
Badri, Cleary et al 2006	Patients enter care with CD4>350 and No-AIDS and receive No-ART	South Africa	US$, 2004, 8%	7,877	6.2 life years or 4.3 QALYs
Bachman, 2005	Patients enter care with CD4>350 and No-AIDS and receive No-ART	South Africa	US$, 2005, 3%	2,952	8.2 life years
Bachman, 2005	Patients enter care with CD4>350 and No-AIDS and receive No-ART with cotrimoxazole prophylaxis from CD4<200	South Africa	US$, 2005, 3%	3,140	8.4 life years
Undiscounted - ART					
Sendi, Bucher et al 1999	ART for patients representative of the Swiss HIV cohort	Switzerland	Swiss francs, 1997, 4%	338,544	14.85 life years
Miners, Sabin et al 2001	ART for patients with CD4<200 and No AIDS	United Kingdom	English Pounds, 1999/2000, 6% on costs and 0% on outcomes	191,624	14.5 life years
Bayoumi and Redelmeier 1998	No MAC prophylaxis and ART for patients with CD4<100	North America	US$, 1997, 3%	233,000	6.48 life years or 4.16 QALYs
Hoffmann and Brunner 2004	No MAC prophylaxis and ART for patients with CD4<100	Berlin, Germany	Euro, 1998, 0%	114,828	5.8 life years
Badri, Cleary et al 2006	Patients enter care with CD4>350 and No-AIDS and start ART with CD4<200	South Africa	US$, 2004, 8%	14,230	18.8 life years or 14.6 QALYs
Bachman, 2005	Patients enter care with CD4>350 and No-AIDS and receive ART from CD4<200	South Africa	US$, 2005, 3%	19,300	14.9 life years

Table 49: Summary of cost-effectiveness/utility analyses – discounted

Reference	Intervention and characteristics of patients	Setting	Currency, price year and discount rate	Lifetime costs in US$	Lifetime outcome
Discounted No-ART					
Freedberg, Losina et al 2001	Patients enter care with CD4<200 and receive No-ART	United States of America	US$, 1998, 3%	55,400	3.33 life years or 2.7 QALYs
Schackman, Goldie et al 2001	Patients enter care with CD4 of 500 and receive No-ART	United States of America	US$, 1998, 3%	69,900	7.02 life years or 6.23 QALYs
Schackman, Freedberg et al 2002	Patients enter care with CD4 of 350 and receive No-ART	United States of America	US$, 1999, 3%	61,600	4.3 QALYs
Goldie, Weinstein et al 1999	Women representative of US HIV-positive population enter care and receive no papanicolaou smears and No-ART	United States of America	US$, 1996, 3%	75,410	3.16 life years or 2.66 QALYs
Goldie, Weinstein et al 1999	Women representative of US HIV-positive population enter care and receive an annual papanicolaou smear and No-ART	United States of America	US$, 1996, 3%	76,700	3.24 life years or 2.73 QALYs
Freedberg, Scharfstein et al 1998	Patients enter care with CD4 200-300 and receive No-ART	United States of America	US$, 1995, 3%	40,288	3.8 life years
Freedberg, Scharfstein et al 1998	Patients enter care with CD4 200-300 and receive No-ART and cotrimoxazole prophylaxis is initiated at CD4<200	United States of America	US$, 1995, 3%	44,786	4.2 life years
Paltiel and Freedberg, 1998	Patients enter care with CD4<100 and receive No-ART	United States of America	US$, 1991, 3%	32,100	2.2 life years
Discounted ART					
Freedberg, Losina et al 2001	Patients enter care and initiate ART with CD4<200	United States of America	US$, 1998, 3%	88,250	5.32 life years or 4.6 QALYs
Schackman, Goldie et al 2001	Patients enter care with CD4 of 500 and start ART with CD4<200	United States of America	US$, 1998, 3%	98,000	8.51 life years or 7.64 QALYs
Schackman, Freedberg et al 2002	Patients enter care with CD4 of 350 and initiate ART with CD4<200	US	US$, 1999, 3%	139,700	9.31 QALYs
Goldie, Kaplan et al - 2002	Patients enter care and initiate ART with a CD4 of 350	US	US$, 1999, 3%	144,260	10.9 life years or 9.8 QALYs
Sendi, Craig et al 1999	Patients on ART enter care with CD4<50 and no AIDS	Switzerland	Swiss francs, 1997, 4%	140,804	6.15 life years
Yazdanpanah, Goldie et al 2003	Patients enter care with CD4 of 370 and start HAART with CD4<350	France	Euro, 2000, 3%	171,060	11.2 life years or 9.4 QALYs
Yazdanpanah, Goldie et al 2003	Patients enter care with CD4 of 370, start HAART with CD4<350 and start cotrimoxazole prophylaxis with CD4<200	France	Euro, 2000, 3%	173,180	11.4 life years or 9.5 QALYs