Measuring Technical Efficiency in Public Spending on Antiretroviral Treatment in Primary Healthcare Facilities in Umgungundlovu District, South Africa

Mziwoxolo Israel Christopher Maxon

Research assignment presented in partial fulfilment
of the requirements for the degree of
Master of Philosophy (MPhil): Development Finance
at Stellenbosch University

Supervisor: Dr Master Mushonga

ABSTRACT

With health needs and expenditure in South Africa on the rise, international support declining and the economic situation deteriorating; efficiency in health spending has become an important matter for decision and policy makers. This study seeks to measure the technical efficiency in public spending on antiretroviral treatment in primary healthcare facilities in uMgungundlovu district, South Africa.

The study examines technical efficiency among 24 primary healthcare facilities in uMgungundlovu district in the financial year 2017/18. District representative data was extracted from KwaZulu Natal Department of Health's reports. We applied input-oriented Data Envelopment Analysis (DEA) models to estimate efficiency scores. To elaborate further on performance, we used bootstrapped truncated regression to identify the determinants of primary healthcare facilities efficiency.

The results indicate that the mean technical efficiency score of 86% was reported for primary public facilities, and the mean scale efficiency was 89%. Moreover, out of 24 primary healthcare facilities only 14 DMUs were fully efficient (TE = 1), followed by the group of 4 that scored mean score between 70 – 99%, while the remaining 6 facilities recorded mean technical efficiency scores ranging between 50 – 70%.

The results suggest that the Department of Health needs to put more effort into reducing (even eliminating) inefficiency in ART service provision for the sustainability of the programme, given the rise in medicine costs and the increasing number of people in need of ART. Furthermore, there is a need to constantly measure technical efficiency, as efficiency information is key to the upgrading of service quality.

DEDICATION

This study is dedicated to my children, Nombulelo, Melokuhle, Nikita and Clarissa and my friends Sandile Buthelezi, Siphokazi Zuma and Vuyokazi Luthuli.

ACKNOWLEDGEMENTS

I would like to thank the people who have always provided support, motivation and guidance to me at various stages of my life and for this thesis. Foremost, I am greatly thankful to my supervisor Dr Master Mushonga, and to Dr J P Opperman and Dr. Tita Tita for their invaluable guidance and suggestions. I acknowledge the support and encouragement I received from my colleagues in my class, especially the support when I fell ill. I further acknowledge help with data analysis from Dr. Mushoni Bulagi.

Special appreciation goes to the KwaZulu Natal Department of Health for providing data used in the thesis. Specifically, I thank Ms Siphokazi Zuma for her tireless assistance and for facilitating the acquisition of the data. On a personal level, I am indebted to my children, for their support during the period when I was pursuing this research study. Additionally, I appreciate the encouragement I received from my friends, Dr. Sandile Sabelo Siyabonga Buthelezi, Dr. Gavin George, Ms Vuyisile Luthuli and Mr. Sandile Miya.

DECLARATION

By submitting this thesis electronically, I, **Mziwoxolo Israel Christopher Maxon**, declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

M.I.C. Maxon

22224025

Copyright © 2019 Stellenbosch University
All rights reserved

TABLE OF CONTENTS

ABSTRACT ii
DEDICATION iii
ACKNOWLEDGEMENTS iv
DECLARATION v
LIST OF FIGURES vii
LIST OF TABLES viii
ACRONYMS AND ABBREVIATIONS ix
CHAPTER 1: INTRODUCTION 1
1.1 Introduction 1
1.2. Problem Statement and Research Questions 3
1.3 Objectives of the Study 4
1.4 Significance of the Study 4
1.5 Chapter Outline 5
CHAPTER 2 6
BACKGROUND OF THE STUDY 6
2.1 Introduction 6
2.2 KwaZulu Natal Levels of Care 7
2.3 HIV Estimates in Umgungundlovu KwaZulu Natal 7
2.4 Key Gaps in the Response 8
2.5 Funding for HIV and AIDS 11
2.6 Chapter Summary and Conclusion 16
CHAPTER 3 18
THEORETICAL AND LITERATURE REVIEW 18
3.1 Introduction 18
3.2 Theoretical Review 18
3.2.1 Why is health sector efficiency important? 18
3.2.2 The health production function theory 19
3.2.3 Basic concept of efficiency 22
3.3 Literature Review 25
3.4 Chapter Summary and Conclusion 30
CHAPTER 4: METHODOLOGY 31
4.1 Introduction 31

4.2 Measuring efficiency in healthcare 31
4.3 Data Envelopment Analysis 35
4.3.1 First stage: Estimation of DEA efficiency score 36
4.3.2 Second stage: Bootstrap truncated regression 39
4.4 Data and variables 40
4.5 Chapter Summary and Conclusion 43
CHAPTER 5: RESULTS AND DISCUSSION 44
5.1 Introduction 44
5.2 Descriptive Statistics 44
5.2.1 Statistics of DEA variables 44
5.3 Results of the DEA Approaches 46
5.3.1 Input-oriented technical and scale efficiency 46
5.4 Results of the Bootstrap Truncated Regression Model 48
CHAPTER 6: 50
6.1 Introduction 50
6.2 Summary of Main Findings 50
6.3 Policy Recommendations 51
6.4 Limitations of the Study 52
6.5 Further Areas of Study 53
6.6 Conclusion 53
REFERENCES 55

LIST OF FIGURES

Figure 2.1: Districts focussed for impact 10
Figure 2.2: Conditional grant allocation between 2012 and 2018 in KZN 11
Figure 2.3: Annual ART spend and TROA: District and Province 16

Figure 3.1: Health production function 21
Figure 3.2: Input-oriented measure of efficiency 22
Figure 3.3: Output-oriented measure of efficiency 24

Figure 4.1: The production process in ART delivery 41
Figure 6.1: Graphic representation of technical efficient DMUs .**Error! Bookmark not defined.**

LIST OF TABLES

Table 2.1: Levels of care in the KwaZulu Natal Department 7
Table 2.2: 2018/2019 Conditional Grant allocations by sub-programme 12
Table 2.3: Annual investment by programme area 15
Table 4.1: Facility (DMU) data showing annual spending per unit 42
Table 5.1: Descriptive statistics for output and input variables 45
Table 5.2: Descriptive statistics for operational variables 45
Table 5.3: Frequency distribution of input-oriented technical and scale efficiency 46
Table 5.4: Input-oriented technical efficiency returns to scale 47
Table 5.5: Truncated regression results for VRS determinants of TE 48
Table 5.6: Truncated regression results for CRS determinants of TE 49

ACRONYMS AND ABBREVIATIONS

AIDS	Acquired Immune Deficiency Syndrome
AP	Average Productivity
ART	Antiretroviral Treatment
BAS	Basic Accounting System
CHC	Community Health Centre
CI	Confidence Interval
CMMT	Comprehensive Care Management and Treatment
CRS	Constant Return to Scale
DALY	Disability-Adjusted Life Years
DEA	Data Envelopment Analysis
DHIS	Department of Health Information System
DMU	Decision Making Unit
DoH	Department of Health of the Republic of South Africa
DRS	Decreasing Return to Scale
FY	Financial Year
GFAN	Global Fund Advocate Network
HIV	Human Immunodeficiency Virus
HSRC	Human Sciences Research Council
IEC	Information, Education and Communication
IRS	Increasing Return to Scale
KAM	Key Account Management
KZN	KwaZulu Natal
MP	Marginal Productivity
NAC	National AIDS Council
NHLS	National Health Laboratory Services
NSP	National Strategic Plan
PCR	Polymerase Chain Reaction
PEP	Post Exposure Prophylaxis
PERSAL	Personnel Salary System
PHC	Primary Healthcare
PMTCT	Prevention of Mother to Child Transmission
PSP	Provincial Strategic Plan

QALY	Quality-Adjusted Life Year
SADC	South African Development Community
SA DoH	South African National Department of Health
SFA	Stochastic Frontier Approach
SSA	Sub-Saharan Africa
STI	Sexually Transmitted Infection
TB	Tuberculosis
TE	Technical Efficiency
TFP	Total Factor Productivity
TROA	Total Remaining in Care
VMMC	Voluntary Medical Male Circumcision
VRS	Variable Returns to Scale
WHO	World Health Organisation

CHAPTER 1: INTRODUCTION

1.1 Introduction

The battle against the Human Immunodeficiency Virus (HIV) epidemic has reached scales of unprecedented proportions in Southern Africa, with South Africa having the highest number of infected persons worldwide. The South African National HIV Prevalence, Incidence, Behaviour and Communication Survey (2012) estimated that among those in the 15 – 49-year age group, the overall prevalence was 18.8% (95% Confidence Interval (CI) 17.5 – 20.3).

The distribution of HIV across provinces shows KwaZulu Natal being amongst the worst affected with a prevalence of 27.9% (CI 25.2 – 30.80) compared to the Western Cape with a prevalence of 7.8% (95% CI 5.5 – 10.9%) (HSRC, 2012). Young women acquire HIV at least five to seven years earlier than their male peers and seem to bear the disproportionate burden of HIV (Abdool Karim, Singh, Short & Ngxongo, 1992; Abdool Karim, Churchyard, Abdool Karim & Lawn, 2009). In 2012, the HIV prevalence in young women in the 15 – 24 age group was 11.4% (95% CI 9.8 – 13.2) compared to 2.9% (95 CI 2.1 – 3.9%) in young men in the same age group (HSRC, 2012).

In addition to household surveillance, the South African National Department of Health (SA DoH) conducts annual anonymous seroprevalence surveys among pregnant women attending ante-natal clinics in public health facilities. These surveys have showed increases in HIV prevalence over a period from 0.8% in 1990 reaching 30.2% in 2010. Some stability was observed in 2011 and 2012 at 29.5% (South African Department of Health, 2014). Similar to the household data, the data from the pregnant women mask geographical variations, and in 2012 the HIV prevalence in KZN was 37.4% (95% CI 36.0 – 38.7) in contrast to 16.9% (95% CI 13.8 – 20.5 in the Western Cape. In addition, seven of the nine districts in KZN showed HIV prevalence above 38.0% with uMgungundlovu District having a prevalence of 40.7% (South African Department of Health, 2010).

In order to counter these higher HIV trends, the government launched a multi-stakeholder programme in 2010 (SANAC, 2012) with an aim of increasing HIV counselling and testing and antiretroviral treatment (ART) uptake, rigorous implementation of prevention of mother to child transmission (PMTCT) of HIV, roll out of voluntary medical male circumcision (VMMC) and the provision of post exposure prophylaxis (PEP) (SANAC, 2012).

Some degree of success has been realised with these programmes (Johnson, 2012), however, providing optimal coverage to prevent further spread of HIV requires strengthening the effectiveness of the ART rollout, including monitoring of outcomes that may be influenced by other factors such as the rate of uptake, adherence and retention in follow-up care.

Over the past sixteen years unprecedented resources and global commitment have added value in the fight against HIV and AIDS. To sustain this level of progress, government will require innovative ways to do more with less. Resources must be adequately allocated to make the biggest impact in the most efficient and effective ways possible. For proper resource allocation, decisions require data on the cost and impact of investments on outputs/outcomes.

There are several ways in which decision makers could use information obtained from HIV programme implementation for innovative decision making (Forsythe, 2006). Key approaches are listed below.

- *Planning and budgeting:* Allow planners to determine, plan and budget for the resources required for current and future HIV programmes and services. This allows for the preparation of budgets and strategic plans, and to compare actual expenditure with budgets.
- *Sustainability:* Costing exercises assist programme planners to ensure that interventions are sustainable by estimating and planning for the costs over time (Forsythe, 2006).
- *Economic evaluation:* Costs analysts are a common component of all types of economic evaluations in HIV/AIDS evaluations that determine the efficiency of the programmes and guides the allocation of resources (Forsythe, 2006). By analysing the inputs for service delivery, planners are able to identify cost drivers, thereby helping identify areas for cost savings by improving efficiencies in the future. This helps in deciding whether to continue with a programme and/or determine cost-effective ways to expand the programme or service.

Having made significant strides thus far the focus is now aimed at ensuring the achievement of the UNAIDS 90-90-90 targets. These include expanded HIV testing services, and ensuring that those tested HIV positive are enrolled in care (since the adoption of a Universal Test and Treat policy). Another is ensuring that HIV services are integrated with other health (chronic) services so that there is seamless integration delivery to promote access to healthcare.

The focus on youth, especially on young women and adolescent girls, and key populations will be further enhanced by access to health information, services and support for youth. The *She Conquers* campaign seeks to protect and promote the rights of both male and female adolescents and young people, and promotes scaling-up the *Comprehensive Package of Interventions* to include biomedical, socio-behavioural and structural interventions.

The main objective of this research study was to evaluate public expenditure efficiency in clinics under Umgungundlovu District, South Africa, during the 2017/18 financial year. Specifically, the research was anchored on the proposition that the cost of ART remains high, despite the massive reductions in drug and operational costs. Since 2009/10 HIV budgets have grown in accordance with the need, and this has had marginal cost increases as prevalence and incidence drive total ART costs. As the need to put more people on ART grows, in order to avoid future HIV care costs, the next frontier should not be about adding more money, but should focus on improving efficiencies in the health system.

1.2. Problem Statement and Research Questions

In response to the impact of HIV/AIDS, substantial amounts of resources have been invested in prevention, treatment and care, and research. The growth in funding for health has increased by 2.2% in real terms. Of this R43.5 billion has been budgeted for HIV which includes HIV/AIDS conditional grant and provincial equitable share. This may be translated to a 6.1% increase from R41 billion in 2009/10 to 2013/14.

HIV constitutes 9% of the total health budget and 1.1% of total government expenditure (Meyer-Rath, 2018). HIV/AIDS prevention and care has outpaced the budget for most other public health programmes, and yet information on the level and nature of the costs of HIV/AIDS programmes remains generally inadequate. There are two important uses for information on costs. The first is to address financing issues. Programme managers and others supporting them need to know what is being spent on HIV/AIDS programmes and how much more will be required in the future as the demand for services changes.

The second use of cost information is to estimate how efficiently or cost-effectively resources are being used and where adjustments might be made to reduce costs and optimise outcomes. Given the substantial amounts of resources being devoted to tackling HIV/AIDS, all managers should be actively working to ensure that they are being used cost effectively.

The process of conducting cost studies in itself has additional benefits: it helps develop an in-house capacity to make better use of existing data; it can inspire improvements in the existing accounting systems; and it can provoke more complex thinking about the relationship between costs and outcomes.

The following research question was addressed for this study:

- Is the comprehensive HIV and AIDS conditional grant spending on ART efficient in primary healthcare facilities in uMgungundlovu District, South Africa?

1.3 Objectives of the Study

The main objective of this study was to measure the technical efficiency of primary healthcare facilities in uMgungundlovu District in delivering antiretroviral treatment.

From this main objective, the following sub-objectives followed:

- to determine the cost efficiency of the ART programme in uMgungundlovu District; and
- based on the research findings, to suggest possible recommendations for the improvement of efficiency in primary healthcare facilities of uMgungundlovu District, South Africa.

1.4 Significance of the Study

Biologically plausible (Gilliam, Fiscus, Marcus & Zhou, 1997) observational studies (Lima, Barrios, Yip & Wood, 2010) and a trial (Cohen, Chen, McCauley, Gamble & Hosseinipour, 2011) show that antiretroviral treatment reduces the transmission of HIV between partners. Modelling papers (Granich, Gilks, Dye, De Cock & Williams, 2009; Hontelez, de Vlas, Tanser, Bakker & Barnighausen, 2009); Wagner & Blower, 2010) have estimated the reduction in HIV incidence and the impact on healthcare costs that would follow from achieving close-to-universal coverage with HIV testing and ART. These reports have suggested that sufficient ART coverage would subsequently pay for itself by suppressing new HIV cases and therefore preventing future need for care, including ART.

On the other hand, the Global Fund has claimed that increasing value for money is now an ethical obligation (Williams, 1992). This is why technical efficiencies are key considerations when deciding on health programmes. Costing a programme is not simply a matter of adding up the bills, as HIV interventions tend to be implemented in complex community-based or healthcare environments, with many different types of expenditures and sources of funding. Also, HIV

responses can (and usually should) differ from one population to another. This objective reality necessitates that comprehensive data be collected from all sources and analysed in order to obtain an accurate estimate of total HIV programme costs in a given context.

This study is intended to estimate technical efficiency in ART spending in a primary healthcare setting in uMgungundlovu District. The research is also intended to make a contribution to establishing how funds are spent, which might help predict how costs may change over time. In this way, cost information can be used to estimate the efficiency of resources being used and other measures of efficiency (Beck, Santas & DeLay, 2008). The study aims to identify if there is technical efficiency in ART spending at primary healthcare facilities. It is expected that the study will: contribute to the debate on the value for money agenda; improve resource allocation; optimise service delivery; and mobilise new resources.

1.5 Chapter Outline

Chapter 1 introduces the research topic, the research problem statement, the research questions and objectives, and the significance of the study.

Chapter 2 gives the background to the study, providing information about the HIV programme in South Africa and more in particular in KwaZulu Natal, the HIV/AIDS conditional grant and how the department is structured.

Chapter 3 presents the theoretical and literature review. The chapter also addresses the concept of technical efficiency measurement in the health sector, and explores in some detail the fundamental concept of technical efficiency

Chapter 4 focuses on research design and methodology, paying attention to how data were collected, and gives details of the technique used to analyse data.

Chapter 5 presents a report on the results of the data collected and some discussion on the results.

Chapter 6 looks at recommendations and conclusions. This chapter also recommends areas for future studies.

CHAPTER 2

BACKGROUND OF THE STUDY

2.1 Introduction

Although the major focus of this study was to estimate the technical efficiency in ART spending, a brief background of KZN Health as a service delivery authority (levels of care), funding for and management of HIV and AIDS Conditional grant.

South Africa spent 30.15 billion rands on health services in 1992/93, equivalent to 8.5 percent of GDP and currently stood at 13, 1 % of GDP in 2016 (Blecher, Chiu, Pillay, & al, 2016). Health expenditure per person was 740 rands or US$247. The established market economies spent a similar proportion of their GDP on health to South Africa, but countries with GDPs per capita comparable to South Africa's spent a lower proportion. For example, the percentage of GDP spent on health was 6.0 in Hungary, 3.0 in Malaysia, 3.6 in Venezuela and 4.7 in Chile.

South African health expenditure disparity is that the data represents more fully the private sector and thus it is believed the data could be overstated relative to many other countries which are unlikely to have undertaken such extensive health expenditure reviews (McIntyre, Bloom, Doherty, & Brijlal, 2007). In the past four years the department of health's budget allocation has increased from R36, 468 billion (2015/16) to R47, 142 billion (2018/19) and is estimated to grow to R61, 939 billion in 2021/22 (National Treasury, 2019).

The HIV/AIDS epidemic remains government's priority and it had (Doyle et al,1991) already been estimated that it would account for between 34 and 75 percent of total health expenditure. The HIV allocations made by the government have grown from R1.2 billion in 2004/5 to R17.5 billion in 2016/17, representing a more than 14-fold growth in nominal public spending (National Treasury, 2019).

2.2 KwaZulu Natal Levels of Care

The KZN province is divided into eleven administrative districts where the health facilities are located. The health facilities are divided into four levels of care.

Table 2.1: Levels of care in the KwaZulu Natal Department

Level of care	Type of facility	Core function/description
Level 1	Primary Healthcare Clinic	▪ The first point of call when citizens need medical attention.
	Community Health Centre	▪ The second sub-level of care offering similar services as the clinic but operating 24-hour maternity, emergency care and wards for short period stays.
	District Hospital	▪ Support to lower level facilities in disease prevention, management and control, health promotion, training and supervision of all sub-district facilities.
Level 2	Regional Hospitals ('Combos')	▪ Offer specialists services, research and training of clinicians and also at times some district hospitals have the capacity to offer regional level services and these are colloquially called "combos".
Level 3	Tertiary Hospitals	▪ Upper level of care offering highly specialised services and delivers for populations beyond the province and country borders.
Level 4	Specialised Hospital	▪ The apex level of healthcare with much specialised referral units providing an environment for multi-speciality clinical services, innovation and research. People are referred by tertiary hospitals. ▪ These facilities offer care only for a certain category of patients, including psychiatric and TB hospitals; but may also include spinal injury and acute infectious disease hospitals.

The backbone of HIV, sexually transmitted infection (STI) and tuberculosis (TB) service delivery is at the first level (Level 1) of care. In 2016/17, the provincial total number of fixed health facilities offering ART services was 825 (580 fixed Primary Healthcare (PHC) clinics, 21 Community Health Centres (CHCs), 81 hospitals and 143 mobiles). In Umgungundlovu District there are 51 clinics: 46 public health clinics; 2 state aided; and 3 community healthcare centres).

2.3 HIV Estimates in Umgungundlovu KwaZulu Natal

Of the 11,074,784 KwaZulu Natal population (StatsSA, 2017), there is an 18.1% HIV prevalence (i.e. 2,004,535 people living with HIV) (HSRC, 2017). The incidence (new infections per year) of

HIV in KZN is 2.3% compared to the national incidence of 1.8%. As at end of September 2016 there were 3,494,260 people on ART in South Africa of which 1,186,932 were in KwaZulu Natal, making it the largest ART programme in the world (KZN Department of Health, 2018). This is a clear sign of a growing demand for healthcare and for more funding allocation in South Africa.

According to estimates from the Thembisa model (mathematical model used to analyse the South African HIV epidemic), new HIV infections for KwaZulu Natal (2016) stood at above 64,600 (KZN Government, 2017). Total HIV prevalence in 2016 was the highest of all provinces (KZN Government, 2017) at 18.0%, having increased by 0.9% from 17.4% in 2012. District specific total HIV prevalence for 2016 ranged from 16.1% in ILembe to 20.6% in uMgungundlovu. Other districts with relatively high prevalence rates were Zululand (19.2%), Ugu (18.5%) and King Cetshwayo (17.8%) (KZN Government, 2017).

The total number of people remaining on care (treatment) in uMgungundlovu stands at 143,320 in the financial year 2017/18 (Provincially, 1,271,116) according to the District Health Information System.

2.4 Key Gaps in the Response

The province remains the leader in most HIV, TB and STIs data; i.e. new HIV infections based on all categories, such as total new infections and by age groups and sex, are the highest in the country. Also, the TB incidence is roughly five times higher than the World Health Organisation (WHO) threshold of 200,000/100,000 and TB multi-drug resistance cases have increased close to double between 2012 and 2016.

The TB/HIV co-infection has increased. The incidence of sexually transmitted infections is high and data showed that a cumulative 31% of the sexually active population had been treated for new STI episodes between 2012 and 2016 (KZN Government, 2017). While the province has put slightly above one million on treatment, the ART coverage still stands at about 61% and retention remains a challenge.

The 2012-2016 Provincial Strategic Plan (PSP) review report noted that the province did not meet its targets in many of its HIV, TB and STIs prevention programmes, and that coverage efforts for these prevention programmes were generally low. While the province achieved its condoms distribution targets, these were not adequate when calculated at number of condoms per eligible

male. Sexually transmitted infections were also high, while reduction of pregnancies for females under 18 years has generally been slow.

The Human Sciences Research Council (HSRC) Behaviour Survey (2012) findings showed poor sexual behaviour outcomes for the province by stating that (a) 11% of women and men aged 15 - 49 years reported having had more than one sexual partner in the past 12 months; (b) only 21% of the women and men aged 15 – 49 years who had had more than one sexual partner in previous 12 months reported the use of condoms during their last sexual intercourse and; (c) 24% of people in the province had misconceptions about the transmission of HIV, TB and STIs.

The 2016 South Africa demographic and health survey stated that 50% of women and 58% of men reported having had sex with a person who was neither their spouse nor had been living with them in the past 12 months. This suggests that reach of information, education and communication (IEC) programmes to the population to effect positive behaviour change has not been adequate, despite an abundance of platforms to convey sustained messages. Both the National Strategic Plan (NSP) review and PSP review noted an absence of a communication strategy to be used in implementing adequate and sustained behaviour change messages.

The NSP 2017-2022 emphasizes 'Focus for Impact' (SANAC, 2017) as a fundamentally new way of doing business in order to achieve a decisive transition from disease control to eliminating HIV, TB and STIs as public health threats. *Focus for Impact* methodology promotes localized responses and moves away from the assumption that the epidemic is uniformly distributed and all areas require 'a one fits all' set of interventions. The essence of the methodology is to deliver services/interventions that can provide the greatest impact with limited resources and highest value for money.

The province has adopted this approach and is using it in prioritizing interventions based on the burden of the epidemics per specific district. Five of the province's eleven districts have the highest proportion of HIV, TB and STIs. These are: (1) eThekwini; (2) iLembe; (3) uThukela; (4) uMgungundlovu; and (5) Ugu. In regard to STIs five districts namely: (1) eThekwini; (2) uMgungundlovu; (3) King Cetshwayo; (4) uMkhanyakude; and (5) Zululand accounted for 72% of those diagnosed with Male Urethritis Syndrome in 2015. Sixteen local municipalities accounted for 42% of the STIs burden, namely: (1) Newcastle; (2) Greater Kokstad; (3) Mandeni; (4) uMhlabuyalingana; (5) Jozini; (6) Mtubatuba; (7) Alfred Duma (8) uPhongolo; (9) Nongoma; (10)

Abaqulusi; (11) Msunduzi; (12) KwaDukuza; (13) uMhlathuze; (14) uMlalazi; (15) KwaDukuza; and (16) Hibiscus Coast.

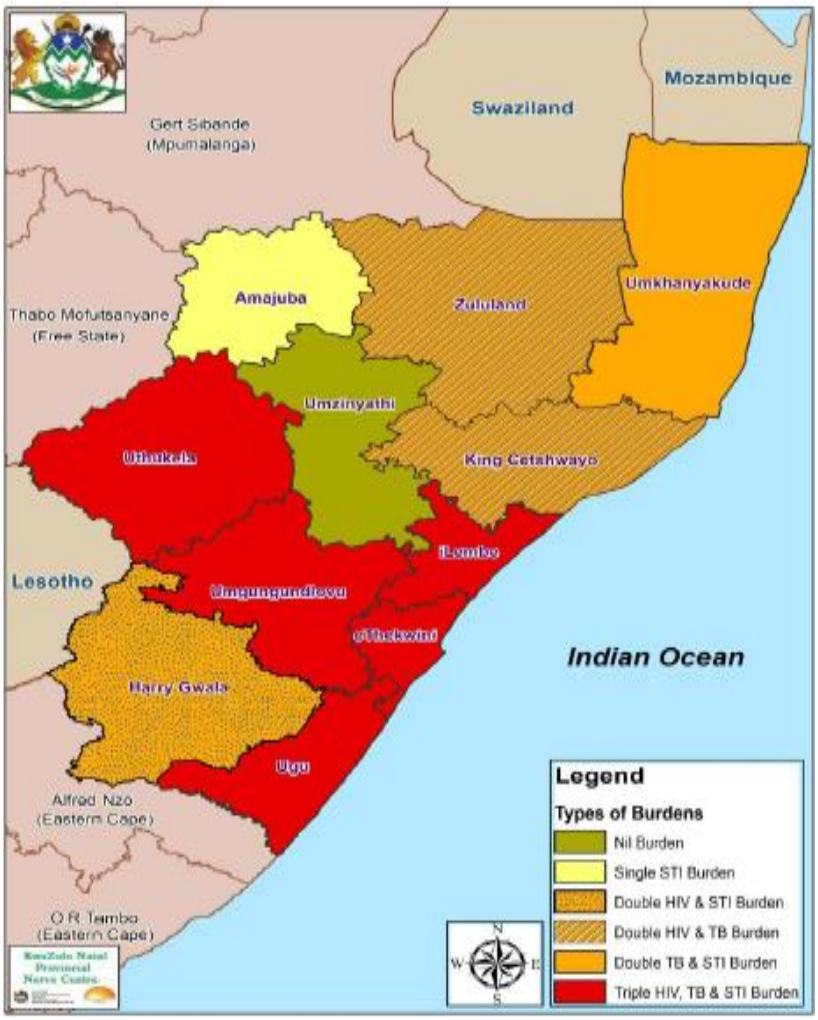


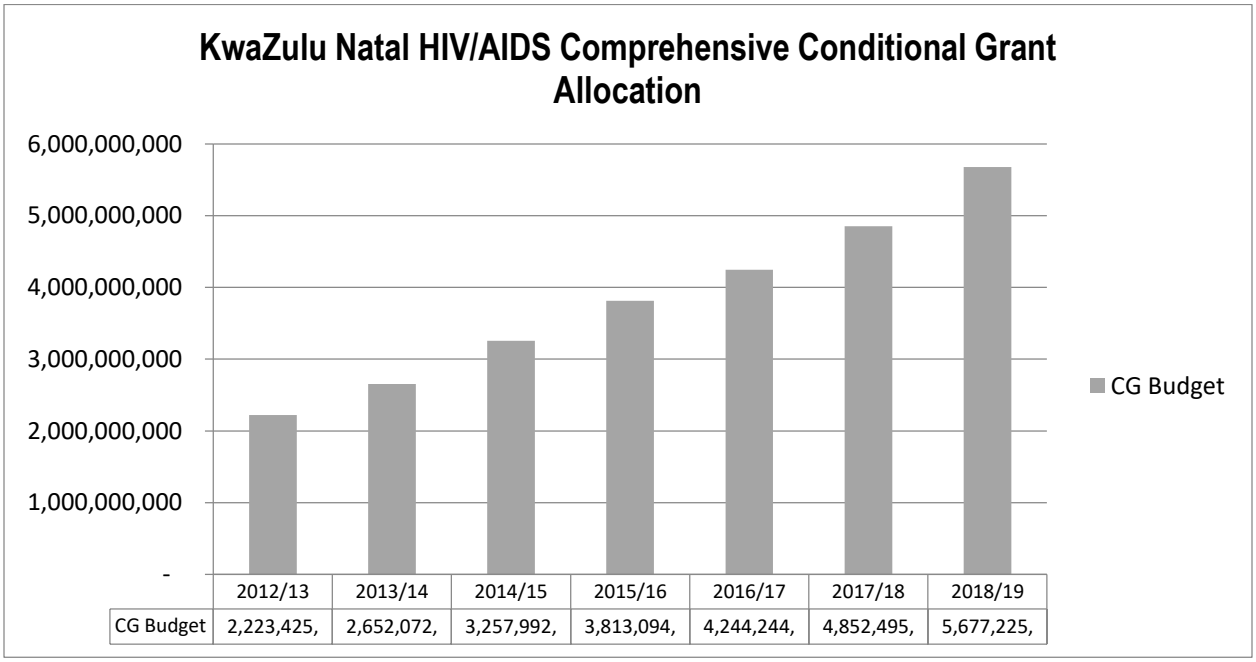
Figure 2.1: Districts 10picent for impact
(Source: KZN Government, 2017)

2.5 Funding for HIV and AIDS

The Comprehensive HIV and AIDS Conditional Grant mainly provides financial resources in order to accelerate the effective implementation of the Comprehensive Care Management and Treatment (CCMT) programme. The CCMT grant is implemented in line with the National Operational Plan for HIV and AIDS Care, Management and Treatment in South Africa and the National HIV/AIDS/STI Strategic Plan 2017-2022, as well as meeting global 90-90-90 targets for achievement by 2020, working towards ending HIV/AIDS epidemic in 2030.

The allocation to the province is based on the HIV and AIDS disease burden and number of patients on the ART programme. KwaZulu-Natal has always been regarded as the epicenter of HIV and, in some instances, had to request additional funds from the National Department of Health for antiretroviral treatment because of the high number of patients on ART. The final allocation for 2018/2019 stood at R 5,677,225,000 (R5.6 billion).

Between 2012/13 and 2018/19 the conditional grant has grown from R2,223,425,000 to R5,677,225,000 with an annual increase of above inflation, i.e. 10% and 19% (KZN, 2018).



	2012/13	2013/14	2014/15	2015/16	2016/17	2017/18	2018/19
CG Budget	2,223,425,	2,652,072,	3,257,992,	3,813,094,	4,244,244,	4,852,495,	5,677,225,

Figure 2.2: Conditional grant allocation between 2012 and 2018 in KZN
(Source: KZN DoH, 2018)

The ART programme receives the largest allocation of the grant. During the 2016-17 financial year the ART programme received 72.4% of the total conditional grant budget. The budget that is allocated to the ART programme is distributed for the following inputs:

- Procurement of the ARV drugs;
- Payment of National Health Laboratory Services (NHLS);
- Compensation of employees providing ART services;
- Procurement of Micronutrients for ART patients;
- Capacity Building of Programme Staff working in the HIV and AIDS programme.

Table 2.2: 2018/2019 Conditional grant allocations by sub-programme
(Source: KZN Department of Health)

Sub-Programme	**Allocation**
Antiretroviral Treatment	R4,319,938
Community Outreach Services	R261,434
Condoms	R124,235
High Transmission Areas	R22,118
HIV Testing Services	R276,288
Male Medical Circumcision	R121,608
Post Exposure Prophylaxis	R4,000
Prevention of Mother to Child Transmission	R189,118
Programme Management	R100,285
Regional Training Centre	R25,200
TB/HIV	R132,000
TB Control	R101,000
Total	**R5,677,224**

Over the years the budget allocations for the HIV and AIDS response in South Africa has grown tremendously, due to the demand for the HIV services. Between 2015/16 and 2017/18 the national comprehensive HIV and AIDS conditional grant has grown exponentially from R13,670,730,000 to R17,557,903,000 (22.1% increase). In KZN the allocation (for the period 2017/18) stands at R4,852,495,000 (National Treasury, 2019) which is the largest share of the conditional grant, followed by Gauteng, though the KZN population is smaller than Gauteng province.

HIV spending between 2005 and 2013 increased by 262% (U.S. Agency for International Dvelopment **(**USAID, 2016). The consolidated national and provincial health HIV and AIDS allocations constituted 7.9% of health expenditure in the financial year (FY) 2012/13. The

government currently provides over three quarters of the funding for HIV. For the FY 2016/17 the government was projected to contribute more to its HIV response; an estimated 82.9% of funding comes from the government, 3.8% from the Global Fund to Fight AIDS, Tuberculosis and Malaria, and 13.3% from PEPFAR (Department of Health of the Republic of South Africa (DoH) and South African National AIDS Council (SANAC), 2019).

Combating HIV and AIDS remains a priority of government (National Department of Treasury, 2019). The DoH is committed to the 90-90-90 targets of the Joint United Nations Programme on HIV/AIDS: ensuring that 90 per cent of people living with HIV are diagnosed; 90 per cent of those diagnosed are initiated on antiretroviral treatment; and 90 per cent of those initiated on treatment have suppressed viral loads.

The implementation of the HIV programme is funded through the Comprehensive HIV, AIDS and TB Grant (hereafter the conditional grant). Its purpose is to enable the health sector to develop and implement an effective responsive to HIV, AIDS and TB (National Department of Treasury, 2019).

The HIV and AIDS component will receive the bulk of the total grant allocation to continue the implementation of the universal test-and-treat policy for antiretroviral treatment, and to provide services intended to prevent HIV. An additional R1 billion is allocated to this component in 2021/22, resulting in an increase from R20 billion in 2019/20 to R24.5 billion in 2021/22.

HIV treatment accounts for a substantial share of the national budget. At the same time, funding from donor sources, such as the United States President's Emergency Plan for AIDs Relief (PEPFAR), is declining. The Global Fund Advocates Network (GFAN) (2018) reported at the International AIDS Conference, in Amsterdam, that if current levels of investments against the three epidemics (malaria, HIV and TB) are maintained, global targets for 2025 and 2030 will be unattainable.

Flat or declining funding from international donors, with an assumption of progress through greater efficiencies and increased domestic investment, will drive the world off-course in its attempts to control and end the epidemics.

The GFAN have estimated the total funding need for AIDS, TB and malaria at US$46 billion annually, of which GFAN estimates that at least US$16.8 to US$18 billion should be invested

through the Global Fund for the Sixth Replenishment (2020-2022) (Global Fund Advocates Network (GFAN), 2018).

They also assert that flat funding pushes the world off-track. The world risks losing control of all three epidemics. Uncontrolled and potentially worsened epidemics will cost countless lives, undermine economic and human development, and threaten the health security for all people on the planet.

The reality of the KwaZulu Natal province is that 23% of the population who are HIV positive do not know that they are HIV positive. Furthermore, there is a gap of 13% of people who are HIV positive but not started on ART. Of those on ART, only 54% are tested for viral load to see the effects of the treatment in fighting the virus, and 90% of those with viral loads done are suppressed. This indicates that there are programmatic deficiencies that have created a gap that must be filled in order to meet the targets set, i.e. eliminating HIV by 2030 (KZN DoH, 2018). It also means resources could be wasted as the desired outcomes are not realized.

In order to sustain the ART rollout and realise the goals of the HIV and AIDS conditional grant, an additional one to one-and-a-half billion rands (above inflation) would need to be provided yearly, over the coming years (Health System Trust, 2016). Again, to expand the programme faster and achieve the international targets, a more efficient management and allocation of resources is needed.

A prior study on the subject (South African Health Review, 2016) indicates significant ineficiencies in the spending at levels below national. The report states that spending performance has been erratic with regular overspending against ART therapy budgets as well as underspending in other areas. Also, the management of the fund and performance against targets varies, with clear indications of poor budgeting in some areas.

Seeing that there is growth in the total cost of the programme (Meyer-Rath, 2018) that is dictated by HIV prevalence and not eligibility, and the ART costs absorb not less than 60 percent of the HIV budget, the question arises whether the programme is affordable in the long run. Also, since HIV budgets have grown over the years, in accordance to need, without bankrupting the country, would putting more people drive up ART costs?

Table 2.3: Annual investment by programme area
(Source: PEPFAR, 2018)

Area	Total Expenditure in R	% GoSA (FY 2016/17)	% PEPFAR (FY 2017)	% GFATM (FY 2016/17)
Clinical care, treatment and support	13,692,640,152	84%	14%	2%
Community-based care, treatment and support	2,783,314,368	81%	17%	2%
Prevention of Mother-to-Child Transmission	587,304,916	45%	55%	0%
HIV Testing Services	1,773,294,168	45%	55%	0%
Voluntary Medical Male Circumcision	1,347,420,440	22%	78%	0%
Priority Population Prevention	1,148,428,584	48%	45%	7%
Key Population Prevention	446,053,778	52%	23%	25%
Orphaned and Vulnerable Children	1,744,238,890	63%	37%	0%
Laboratory	225,170,094	0%	100%	0%
SI, surveys and surveillance	507,752,702	0%	89%	11%
Health Systems Strengthening (HSS)	162,937,616	0%	81%	19%
Other HIV spending	2,014,456,838	93%	0%	7%
Total	**26,433,012,546**	**71%**	**26%**	**3%**

General Notes:
GoSA figures are based on Basic Accounting System (BAS) actual expenditures for FY 2016/17 (exchange rate: ZAR14.58:USD1). GFATM figures are actual expenditures from FY 2016/17 (exchange rate: ZAR14.58:USD1). PEPFAR figures are based on FY 2017 expenditures (exchange rate: ZAR13.5:USD1).

The table provides a broad profile of expenditures and budgets for HIV spending in SA, and is not comprehensive of all HIV expenditures in SA. BAS data do not provide specific information on spending and budget allocation for several of the programme areas or component areas listed, including laboratory, OVC, Communications, Monitoring and Evaluation, Other Prevention, Policy and Systems Development. This lack of information does not reflect a lack of GoSA expenditure in these programme areas.

The current procurement of ARTs is mainly by the government (97%) and minimally by the Global Fund for AIDS, Malaria and TB (3%), estimated (in 2016/17) at R5.89 billion and R1.26 million, respectively (PEPFAR, 2018).

uMgungundlovu District is home to a population of 1,095,861, where 225,747 (20.6%) people are already living with HIV (StatsSA, 2016). There are 7,543 new HIV infections annually (Johnson, Dorrington & Moolla, 2017). In order close the 18% gap of those already on treatment and to reach the 90-90-90 goal, there will be more than a need for money to close the gaps.

In the 2017/2018 financial year uMgungundlovu District had a budget allocation of R160,009,000 for the procurement of ARV drugs. The unit costs of ARVs for the fixed dose therapy is estimated, according to the KwaZulu Natal Department of Health, at R920.95 per annum. This means that to put the 7,543 newly infected persons (annually) on treatment will require an addition of R6,946,725.85 per annum or R34,733,629.25 over the next five years. Sustaining the programme moving forward will require a constant allocation of no less than the current allocation, plus the addition for new infections, and improvements in efficiencies in the health system.

A review of data from uMgungundlovu District shows that while there has been an increase in the budget allocation (see Figure 2.3) which is accompanied by a slight increase in outputs as demonstrated by an uptake rate of between 5% and 15%.

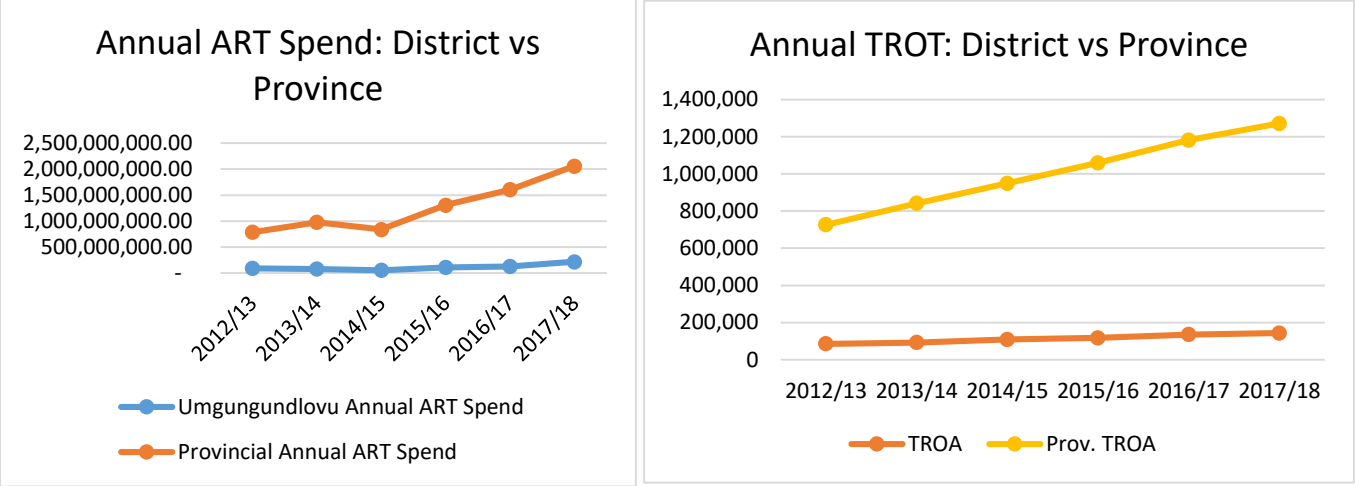


Figure 2.3: Annual ART spend and TROA: District and Province
(Source: KZN DoH, 2018)

This brings to the fore the question of value for money in public health (and public service, in general) which underscores the need to focus on efficiencies. The current service gaps suggest that there is going to be a greater need for expansion (both service and financial) to ensure that all those in need of ARTs get them in order to live a longer life without bankrupting the country.

2.6 Chapter Summary and Conclusion

Increasing access to antiretroviral treatment, especially for people who would not have qualified in previous years, has become a major goal in the response to HIV and AIDS. In 2016 the government adopted the 'test and treat' policy, meaning that more people will be started on ART

sooner, and stay on them longer. In the face of these demands for scale up, healthcare systems and HIV programmes need to clearly understand the resources that ART demands. The next chapter of this research study presents the theoretical framework and literature review.

CHAPTER 3

THEORETICAL AND LITERATURE REVIEW

3.1 Introduction

The pursuit of efficiency must be the central preoccupation of health programme managers and policy makers, and it is justifiably the cause for much concern in the HIV and AIDS programme sustainability. Inefficient delivery of healthcare may lead to unnecessarly poor outcomes for the patients directly affected, measured either in terms of their health improvement, or in their broader satisfaction with the health system.

This chapter offers a framework for thinking about the conceptualisation and the measurement of efficiency in healthcare. The chapter explains why the understanding of health sector efficiency is important. The chapter goes on to clarify what is meant by efficiency, and explores in some detail the fundamental concept of technical efficiency.

3.2 Theoretical Review

3.2.1 Why is health sector efficiency important?

The notion of health efficiency, including related issues such as cost effectiveness and value for money, are some of the important dimensions of healthcare performance. These concepts seek to capture the extent to which the inputs to the health system, in the form of expenditure and other resources, are used to secure valued health system goals. In other sectors of the economy, consumer preferences help to ensure that the most valued outputs are produced at market prices. But, due to market failures in the health sector it means that traditional market mechanisms cannot work, thereby allowing poor quality or inappropriate care to persist at high prices, if no policy action is taken.

In the context of HIV and AIDS programmes, recognizing that the HIV epidemic is a long-term challenge that requires a sustainable response, the UNAIDS (2011) decided to ‘rethink the approach to investments in HIV, UNAIDS’.

The HIV investment case framework provides an important vehicle for countries to deliver strategic, rights-based, sustainable responses to HIV. The process of developing investment cases provides countries with new opportunities to explore options for innovative funding and service delivery, to identify specific steps to enhance equity and inclusiveness for key populations, to use

available evidence to understand better the health and economic benefits of timely, rights-based, smart HIV investments and to eliminate inefficiency in HIV programmes.

Inefficient use of healthcare resources is a serious concern because:

- It may deny health gain to patients who have received treatment because they may not receive the best possible care available within available resources;
- When consuming excess resources, inefficient treatment may deny others who could have benefitted from treatment had the resources been used better;
- Inefficient use of resources in healthcare may sacrifice loss of consumption opportunities elsewhere in the economy, such as education; and
- Wastage due to inefficient care may reduce society's willingness to contribute to the fulfilment to the funding of health services, thereby harming social solidarity, health system performance and social welfare (Cylus, Papanicolas & Smith, 2017).

Efficiency becomes more important in light of the financial pressures and concerns over long-term financial sustainability.

In working to remove inefficiency from HIV programmes and services, costs may potentially be saved in three key areas: lowering commodity costs, including the costs of antiretroviral medicines; identifying alternative service delivery models, including community-based services; and eliminating parallel structures and integrating HIV services. Value for money involves a balance that maximizes high-quality output while optimizing resources. At all times, efforts to save money should ensure that the quality of services is not compromised. Efforts to ensure value for money must be rights-based and gender-sensitive, ensuring that the HIV-related needs of the most severely affected populations are effectively addressed (Vogel, 2008).

3.2.2 The health production function theory

Health system efficiency seeks to capture the extent to which the inputs to the health system, such as expenditure and other resources, are used to achieve valued health goals. In other areas of the economy consumer preferences help to ensure that the most valued outputs are produced at market prices. But there are a number of market failures in the health sector that mean that the traditional market mechanisms cannot work, allowing poor quality or inappropriate care to persist at higher prices, if no policy action is taken.

The measurement of health efficiency is simple, represented at its most straightforward as a ratio of resources consumed (inputs) to some measure of the valued output they create. (Cylus, Papanicolas & Smith, 2017). However, putting this somewhat simple concept into practice can be complex.

Economists regard the transformation of inputs into outputs as a productive function, which indicates the maximum feasible level of output for a given set of inputs. Any level of production below that maximum is an indication of inefficiency (Atun, Bennett & Duran, 2008). The concept of the production function can be applied to the functioning of a very small unit (such as a doctor's practice) through to a huge macro unit (such as the entire health system).

Grossman (1972) developed a theoretical health production function, which can be expressed as:

H=F(X)

Where **H** is a measure of individual health output and **X** is a vector of individual inputs to the health production function **F**. The elements of the vector include: nutrient intake; income; consumption of public good; education; time devoted to health related procedures; initial individual endowments like the environmental determinants and personal choices. This theoretical model was designed for analysis of health production at micro level.

In economics there is talk of firms "producing" outputs through a combination of "factor inputs", chiefly these being labour and machines. The relationship between inputs and the final outputs is called "production function" (Wagstaff & Yol, 2014). The demand for healthcare uses this theory and conceives of the individual producing his/her health by combining health inputs, such as medical care, is just one example.

The health production function combines inputs to realize the desired output/s. Figure 3.1 illustrates the health production function. The output-health is measured along the vertical axis. The figure also indicates how inputs (on the horizontal axis) are used to produce a certain level of outputs (on the vertical axis. As more units of health inputs are utilised, more output is produced.

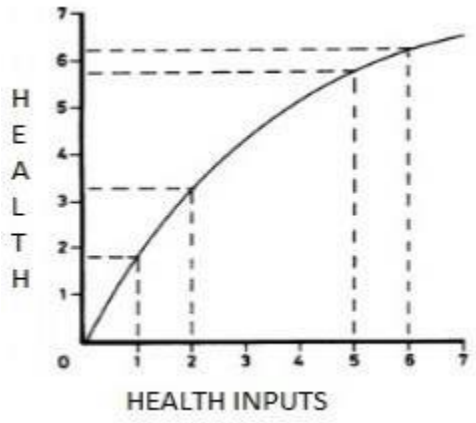


Figure 3.1: Health Production Function Curve
(Source: Adapted and Modified from Wagstaff (1986)

The curve further indicates that successive additions to the quantity of health inputs result in successively smaller increments in health. For example, increasing the amount of health inputs from 1 to 2 units results in an increase in health of about 1.5 units. Increasing the input from 5 to 6 units, however, results in an increase in health of only 0.5 units – the 'law of diminishing marginal products'.

The importance of specifying health production functions becomes apparent when one attempts to determine the amount of resources to be allocated among alternative health inputs to produce the largest possible increase in health levels. Without specifying a production function, equi-marginal analysis, an important tool for economic optimization, cannot be carried out (Wibowo & Tisdell, 1992).

Feldstein (1983) argued that the key information requirement for economic optimisation, among others, is empirical information on the marginal effect on health of each of the health programmes. Therefore, where information is not available, allocation decisions are made on the basis of the average benefits, instead of the marginal benefits of the programmes (Wibowo & Tisdell, 1992). This approach assumes a linear production function without a constant-term, which is not practically possible.

The health production function is important when one seeks to analyse the cost-benefit or cost-effectiveness analysis of alternative programmes. Warner and Luce (1982) suggested that,

irrespective of the method chosen, identifying inputs and outputs and their specific linkage provides the basis for estimating costs and benefits (or effectiveness).

3.2.3 Basic concept of efficiency

To economists, efficiency is a relationship between ends and means. When we call a situation inefficient, we are claiming that we could achieve the desired ends with less means, or that the means employed could produce more of the desired ends. “Less” and “more” in this context necessarily refer to less and more value. Thus, economic efficiency is measured not by the relationship between the physical quantities of ends and means, but by the relationship between the value of the ends and the value of the means.

The desirable goals of any production unit are the attainment of high levels of economic efficiency and productivity. The commonly used ratios for productivity measurement are average productivity (AP), marginal productivity (MP) and total factor productivity (TFP). On the other hand, efficiency of a production unit consists of a comparison between observed and optimal values of its output and input (Lovell, 1993).

The idea of measuring the efficiency of a production unit, using frontier analysis, was developed initially by Farrel (1957). Farrel’s idea of efficiency estimation can be explained graphically using Figure 3.2.

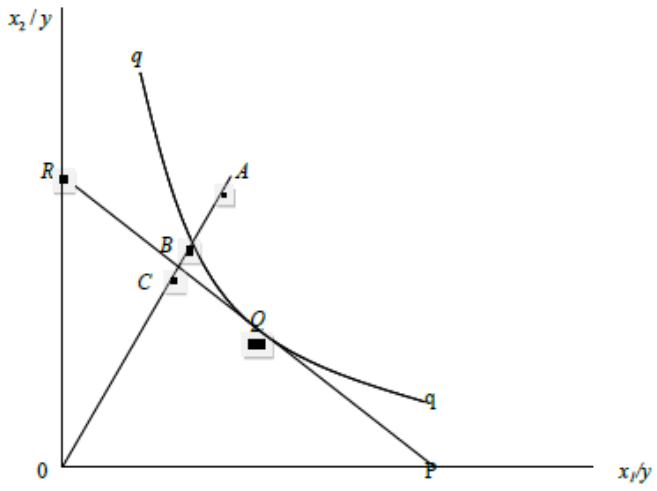


Figure 3.2: Input-oriented measure of efficiency (Source: Green, 2007 and Coelli, 1996)

Taking a case in which a firm uses two inputs, x1 and x2, the production function is generally specified as follows: $y = f(x1, x2)$, where y is output. Assuming constant return to scale, so that $1 = f(x1/y, x2/y)$, the frontier is characterised by the efficient unit isoquant denoted by the curve qq (Lovell, 1993; Schmidt and Lovell, 1979).

Figure 3.2 shows the unit efficient isoquant *qq* and a budget line RP. Assume that the firm uses *(x1*, x2*)* to produce output *y**. In Figure 3.2, point A is given by *(x1*/ y*, x2*/ y*)*. The technical efficiency is measured by *0B/0A*. This ratio measures the proportion of *(x1*, x2*)* that is actually necessary to produce *y**. The technical inefficiency, given by *1-(0B/0A)*, measures the proportion by which *(x1*, x2*)* could be reduced without reducing output, the input ratio being constant. It also measures the possible reduction in cost of producing *y**, holding input ratio constant.

In Figure 3.2, Q represents the equilibrium point at which cost is minimized. The cost of production is equal at points C and Q, but level of output at C is less than at point Q. However, output level is the same at point B and Q. Therefore, the ratio OC/OB gives the allocative efficiency and 1-(0C/0B) gives allocative inefficiency. The allocative inefficiency measures the possible reduction in cost from using the correct input proportion. 0C/0A gives the total efficiency, while 1-(0C/0A) is total inefficiency. The latter measures the possible reduction in cost by moving from the observed point A to the cost minimising point Q (Forsund et al., 1980).

The above graphical analysis gives input-oriented measures of efficiency. However, it is also possible to address the same measures using output-oriented approach. Assuming a multi-product case of two outputs y1 and y2, the graphical analysis of efficiency is illustrated in Figure 3.3.

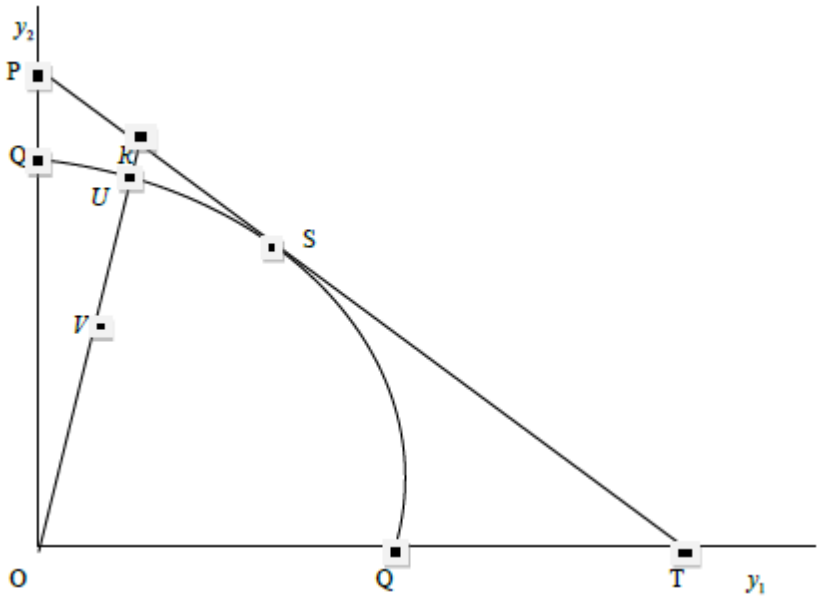


Figure 3.3: Output-oriented measure of efficiency

The figure shows the production possibility curve QQ and the price line PT. The output y1 lies on the x-axis and output y2 lies on the y-axis. Let V be a point on the line OR and U be point on the intersection of line OR and curve QQ. If production is taking place at point V, technical efficiency is given by OV/OU and technical inefficiency is 1-(OV/OU). Likewise, allocative efficiency is measured as OU/OR, and allocative inefficiency is 1-(OU/OR). Lastly, total efficiency is given as a product of technical and allocative efficiency, i.e. (OV/OU) x (OU/OR) = OV/OR (Coelli et al., 1998).

The determination of the above efficiency measures requires the production function to be known. It is very difficult, however, to know exactly the underlying function. This problem led to the use of the efficient isoquant estimated from the observed data. Farrell (1957) suggested the use of either of the following two approaches in the estimation. First, using a non-parametric piece-wise-linear convex isoquant constructed so that no observed point lies above or below it. Second, using parametric function, such as the Cobb-Douglas form, fitted so that no observed point lies to the left or below it (Coelli et al., 1998).

The first approach led to the development of data envelopment analysis (DEA), and the second one led to use of parametric techniques such as deterministic and probabilistic frontier models.

3.3 Literature Review

Empirical literature on efficiency is very broad and includes all economic sectors. With specific reference to health economics, studies started in the 1970s (Lave & Lave, 1970) and have been continuing at an increasing rate.

In South African HIV/AIDS context, technical efficiency was mooted in the South African HIV and TB Investment Case (2016) wherein the partners reported that over the previous decade, there had been large decreases in new HIV infections and deaths from AIDS. Tuberculosis prevalence and deaths had levelled off, in large part due to the massive rollout of ART, and the country had recovered from the significant decrease in life expectancy in the early years of the epidemic. South Africa is on the way to reaching the target of an average life expectancy of 70 years for all by 2030, as envisaged in the National Development Plan (2030).

The Investment Case helped the government and partner organisations in four ways: by reviewing the evidence base for all known interventions against HIV and TB, including those that were part of the HIV and TB programmes and a number of those that could be added; comparing the impact and cost of each of these interventions and suggesting an optimal package of services to reach important targets in controlling the two diseases; calculating the total budget needed to implement the optimal package over twenty years so that the return on the initial investment becomes clear; and pointing out in greater clarity where the gaps are in the collective knowledge on what works against the two diseases.

The Investment Case acknowledged that the projected spending on HIV and TB would continue to rise in future years and that the share of spending covered by the South Africa government would also continue to rise. However, the projected increases are unlikely to meet resource needs, with a substantial resource gap projected for each of the following five years (National Department of Health and SANAC, 2016).

The Investment Case (2016) realises that real growth of healthcare budgets was strong until 2011/12 (exceeding 5% per annum), but slowed substantially from 2012/13, and especially in the subsequent years, when it started to level off in real terms (Blecher et al., 2016). The Investment Case further noted that achieving high levels of technical and allocative (an examination of whether limited resources are directed towards producing the correct mix of health outputs)

efficiencies is crucial to sustainability, especially in a constrained fiscal environment. It further recommended that further gains may be realised by improving technical efficiency.

Technical efficiency has since become a critical feature of subsequent Investment Case reports where it is suggested that improved technical efficiency could yield savings and rationalisation for the HIV programme and value for money in health service delivery platforms. It should be noted, however, that while efficiency enablers are identified, these do not take into account the constraints as identified hereinunder by Mikkelsen et al. (2017).

There has been an increase in literature (Kirigia et al., 2010, Tlotlego et al., 2010) analysing the efficiency of public health expenditure in Africa. Other studies have analysed efficiency at a disease level within and across countries (Baily et al., 1997, Garber and Skinner, 2008). The ORPHEA study design and methods was key breaking ground in looking at the cost of scale up and technical efficiency required (Bautista-Arredondo et al., 2014). The study reviewed unit cost and information from different delivery contexts, both within and across countries, which it proposes is critical in identifying and addressing inefficiencies. The study is regarded as one of the most comprehensive studies on the cost and technical efficiency of HIV programmes.

The study collected data to estimate cost efficiency in Kenya, Rwanda, South Africa and Zambia. It was found that scaling up HIV services (i.e. HIV counseling and testing and provision of treatment) to achieve national and global targets will require that countries make concerted efforts to optimise the use of available funding. The study also recommended that information on the unit cost of services is critical in identifying and addressing the inefficiencies impeding the delivery of the highest possible quality HIV services at the lowest feasible cost. The researchers did not publish results of their study, but recommended that further work needs to build on their efforts in order to expand knowledge on the topic.

Meyer-Rath and Over (2012) produced a policy discussion document on the feasibility of massively scaling up antiretroviral treatment to reduce HIV transmission and incidence, which they argue depends more on accurately projecting the cost of such a scale up in comparison to the benefits from reduced HIV incidence and mortality. The researchers reviewed existing modelled estimates of the cost of providing ART to different populations around the world and suggested modelling methods that use cost accounting versus flexible cost function in the future.

The researchers suggest that projections of alternative ART scale up scenarios need to be plausible, supported by observational studies and alternative delivery methods. They suggest that the envisaged cost-effectiveness analyses, together with planned large-scale trials rollout, provide a historic opportunity to collect data and allow more precise cost functions. Data collection on large samples of facilities should go beyond measuring the quantity and quality of ART services, to capturing the actual cost of services delivered in facilities at different levels of care. The availability of such data may, the researchers suggest, also help improve the cost-effectiveness (technical efficiency) of the programmes, which we hope to achieve through this study.

Hontelez et al. (2011) also conducted a study on the impact of fully adopting the WHO guidelines on HIV epidemic treatment dynamics and associated costs where the start of HIV treatment is recommended for those infected with HIV and their CD4 counts are equal or less than 350 cells per ul. The study strengthened the WHO recommendation of starting ART at ≤350 cells/μl for all HIV-infected patients. Apart from the benefits associated with many life-years saved, a modest frontloading appears to lead to net savings within a limited time-horizon. This finding is robust to alternative assumptions and foreseeable changes in ART prices and effectiveness. It was recommended that South Africa should aim to rapidly expand its healthcare infrastructure to fully embrace the new WHO guidelines. Conducted in Hlabisa (KwaZulu Natal, South Africa) where it predicted the HIV dynamics in a 30-year time horizon, number on ART (instead of an annual 7% increase, it predicted 28% more people on ART) and programme costs. The study further suggested that there is a need to improve the delivery infrastructure so as to realise the desired impacts.

Sweeny et al. (2012) reviewed literature on the potential efficiency gains of integrating HIV services with other health services. This was done in the context of increasing political commitment and previous reviews suggesting a strong public health argument for the integration of HIV services. The researchers found that integration of services improves efficiency. However, evidence gaps remain and there was an urgent need for further research in this area.

Mikkelsen et al. (2017) also looked at the impact of scaling up HIV treatment in sub-Saharan Africa (SSA) and argued that current evidence for scale up may be theoretically optimal but practically infeasible to implement. The researchers suggested that the cost-effectiveness analyses of scaling up ART in SSA should be made more responsive to the context by taking into account the local health system. They argued that there are constraints at both the supply-side (service

delivery side) and demand-side (community side) which impact on the quality and quantity of health services being offered.

The supply-side is based on the WHO building blocks which are:

(i) adequate service delivery – availability of goods, timeliness and sustainable care to all patients along the full cascade; (ii) sufficient number of qualified health workers to provide treatment and care to all eligible patients; (iii) a health information system that monitors health system performance and systematically records essential information on all patients; (iv) a reliable supply chain management of medical products and technologies to ensure sufficient supply of drugs and medical equipment, sustainable financing to maintain care and treatment for those eligible; and (vi) good governance that allows for balanced, contextual decisions in allocating and managing limited resources.

The demand-side constraints may include: stigma; limited knowledge about HIV; fatalism, out-of-pocket payments; and waiting times which can affect health seeking behaviour such as treatment uptake and adherence. These constraints tend to be related to the disease stage, as people who are relatively stable or at an early stage of the disease experience fewer symptoms and thus may be less motivated to seek care.

Mikkelsen et al. (2017) note the work done by Cleary, McIntyre and Boulle (2008) in South Africa, Fraser et al. (2015) in Nigeria and Anderson et al. (2014) in Kenya, and posit that in order to move towards a more realistic and methodological innovation cost-effective approach there is a need to incorporate health system constraints into modelling models. Financial, infrastructural and human resources needed for various alternatives need to be explicitly incorporated in a dynamic way. Furthermore, researchers point out that supply-side constraints cannot be treated statistically, e.g. by simply assuming a maximum number of people who can receive ART (Mikkelsen, et al., 2017). Rather, it is proposed, the effects of insufficient resources should dynamically translate into reduced performance along the healthcare continuum to adequately reflect the patient and population-level effects of policy decisions.

Mbonigaba and Oumar (2014) found that the ranking of municipalities is not the same across both types of healthcare when efficiency scores and efficiency score growth are contemplated. The results implied that municipalities in South Africa are generally inefficient, but with the possibility of learning from each other's practice in order to increase their technical efficiency. The health

system authority should monitor service-specific best practices among municipalities so that they can use them as practice guidelines for other municipalities.

Kirigia, Emrouznejad, Sambo, Munguti and Liambila (2004) measured the technical efficiency of public health centres in Kenya. Their finding suggested that 44% of the public health centres are inefficient. They recommended that this type of study be conducted in other countries so as to empower countries to play their stewardship role more effectively.

Van der Westhuizen and Dollery (2009) conducted efficiency measures in local government. The researchers sought to measure the productive efficiency with which municipal councils had delivered electricity, domestic waste removal, sanitation and water in line with their new responsibilities applied to cross-sectional data covering the period 2006/2007 for 231 local municipalities and 46 district municipalities.

The results suggested that eight out of the nine South African provinces had one or more local municipality that was fully technically efficient. Only the Western Cape had no local municipality that was fully technically efficient. The Gauteng Province had the highest average efficiency estimate of 79.4 percent under the output-orientated approach and 76.7 percent under the input-orientated approach. The Western Cape was the worst-performing province with an average efficiency estimate of 55.25 percent under the output-orientated approach and 50.8 percent under the input-orientated approach.

Jayasuriya and Wodon (2003) measured health and education efficiencies in 76 countries from 1990 to 1998. The researchers found that urbanisation and, to some extent, the quality of the leadership are strong determinants of the efficiency of countries in improving education and health outcomes, while the impact of corruption is not statistically significant.

A study by Kirigia, Mensah et al. (2010) in South Africa and the Southern African Development Community revealed that 87%, 87%, 61%, 52% and 35% of the hospitals were run inefficiently in 2003, 2004, 2005, 2006 and 2007, respectively, and needed to either increase their outputs or reduce their inputs in order to become efficient. The average variable returns to scale (VRS) technical efficiency scores were 63%, 64%, 78%, 78% and 86%, respectively during the years under consideration.

Facility based efficiency studies in Ghana (Osei et al., 2005; Nuti et al., 2011; Akazili et al., 2008; Zere, 2000) were used to look at the efficacy of policy reforms and advise on further reforms to be made. Similarly, in China (Chu et al., 2015) and in India (Jat & Sebastian, 2013) efficiency studies concerned themselves with the policy reforms that were needed in order to improve efficiency and, importantly, to answer the question of sustainability for universal health coverage. These provided a broader context of analysing efficiency which may depend on the health architecture and may not be applicable in a country where health is a dual responsibility or a federal arrangement. Their importance in measuring sustainability for a country like South Africa that seeks to introduce universal health coverage cannot be overstated.

In the context of HIV and AIDS a number of studies (Martin, 2003; Blecher et al., 2016; World Bank, 2015; 2015; Guthrie et al., 2016) have looked at the efficiency and its impact on sustainability. These studies have brought greater focus to allocative efficiency with important questions being whether the problem is allocative, and/or whether it poses sustainability challenges.

The World Bank (Glassman, 2013) and the International Monetary Fund (Grigoli & Kasoli, 2013) have warned against the inefficient use of resources for HIV and AIDS and its sustainability. The WHO Report 2000 brought attention to efficiency in healthcare delivery by providing estimates of health system performance for countries.

3.4 Chapter Summary and Conclusion

Given the existing evidence that largely supports HIV treatment scale up from a public health and clinical perspective, there is also evidence that suggests efforts to ensure efficiency of health systems. Investments in context-based operational levels needs to be supported by reliable efficiency reviews and system building blocks to support the scale up efforts and sustain the HIV treatment programme. The following chapter provides a framework on how data gathering and analysis were done.

CHAPTER 4: METHODOLOGY

4.1 Introduction

The main objective for measuring the technical efficiency of health facilities is the raping increase in the cost of healthcare services. In the context of HIV and AIDS, the concern is mostly the sustainability of the treatment programme in the long term given that it's chronic disease. Irrespective of the healthcare system, being publicly or privately funded, efficiency is a critical indicator for ensuring the quality of care and to minimize wastage in the system.

Compared with oter industries, measuring efficiency in the health sector is complicated by charecteristics specific to health and health services (Ozcan, 2008). The measurement of efficiency in the health organisation takes into account the process and outcome. The former (process) concerns itself with intermediate services such as the number of tests performed or patients served and the outcome approach conerns itself with the health status as an ultimate outcome (Ozcan, 2008).

This research aims to compare various primary healthcare facilities using data from the public health sector. In the context of measuring efficiency of a set of decion making units (DMUs), DEA has been increasingly used in most of the studies carried out in the last decades where there are multiple inputs and outputs (Chilingerian & Sherman, 2011) .

4.2 Measuring efficiency in healthcare

Healthcare costs have grown dramatically in the recent past and it is widely believed that the inefficiencies of healthcare delivery have contributed to this rise. There is a growing body of literature that seeks to address the measurement of efficiency in healthcare. And while hospitals have been a subject of efficiency studies, the efficiency of other levels of care has also been addressed.

Efficiency is composed of two main components (Coelli T. , Rao, O'Donnell, & Battese, 1998); technical efficiency and allocative efficiency. Technical efficiency occurs when a firm obtains maximum output from a set of inputs, whereas allocative efficiency occurs when a firm choices optimal combination of inputs given the level of prices and the production technology (Coelli, Rao et al 1998 and Rogers 1998).

The subject of methodology distinguishes between two main types efficiency analysis used to in the health service. These are categorized into: economic evaluation and benchmarking analysis

(Peacock, Chan, Mangolini, & Johansen, 2001). The economic evaluation matches progammes or healthcare interventions in terms of their cost and benefit. Whereas, benchmarking analysis matches service providers against best practice standards of the production process.

The economic evaluation analysis is about choosing the most efficient way of delivering care. Because of the scarcity of resources, we also need to choose an intervention that has value for money (Goodacre & McCabe, 2002). The key determinants of economic evaluation are costs and outcomes of healthcare interventions. For this analysis to work, it needs to have two features; (i) both cost and outcome must be analyzed, and (ii) more than one intervention must be compared.

The techniques used for economic evaluation include: cost-effectiveness analysis (CEA), cost-minimization analysis CMA), cost-utility analysis (CUA), and cost-benefit analysis (CBA). These techniques have been used in various studies (Lacey et al 1999; Babigumira et al 2009; Babigumira et al 2010; Babigumira et al 2009; Kuznik et al 2012; and Brent, 2010) healthcare, and HIV and AIDS in particular, where they were seen to be advantageous because of their simplicity, easy to interpret and also ease of communication and specificity (Babigumira J. B., 2017).

The noted disadvantages include (Babigumira J. B., 2017): weighting of different attributes is left to individual decision makers, increases welfare of decision maker, and decision by individual might not be in the best interest of patients or society.

Also, the cost-mineralization analysis is no longer considered by many to be "valid" (Briggs, 2001) and omitted from seminal text as a type of economic evaluation (Drummond, Weatherly, & Ferguson, 2008). The argument is that scientists should do CEA and estimate joint density of costs and effects and examine uncertainty regardless of whether there is a statistically significant difference in effectiveness.

Another category is called benchmarking where a collection of service providers that delivery a service or package of services. Benchmarking is defines as continuous measurement and improvement of an organization's performance against the best (Kozak, 2002). In healthcare, benchmarking is about implementing best practices at best cost (Ettorchi-Tardy, Levif, & Phillipe, 2012). The techniques used for benchmarking are: simple ratio analysis, average cost analysis, data envelopment analysis, and stochastic frontier analysis.

The well-known benchmarking tool developed for healthcare was developed by the World Health Organization (WHO) is the Performance Assessment Tool for Quality Improvement in Hospitals

(PATH). It was designed by WHO to support hospitals in defining quality improvement strategies, questioning their own results and translating them into actions for improvement. PATH is a toll that measures and manages performance, and foster and support a culture of measurement and continuous improvement (Buttigieg, 2012).

The major (two) major features that differentiate between the alternative empirical approaches for forming the frontier and measuring efficiency and productivity methods are whether: (i) they are parametric or not, and (ii) whether they are deterministic or stochastic (Hollingsworth, Dawson & Mniadakis, 1999).

Parametric methods assume a specific functional form for the frontier whereas non-parametric methods do not; and deterministic methods assume that the distance of a unit from its frontier is a result of inefficiency whereas stochastic methods assume that some of it is due to random error.

A number of healthcare studies have analysed efficiency of health facilities using the non-parametric method called the Data Envelopment Analysis (DEA) (Nedelea, Fannin, & Barnes, 2010). Over 400 published studies have used both the parametric and non-parametric methods within health care settings over the past 30 years (Hollingsworth, 2003, 2008, 2012).

The DEA uses linear programming by creating virtual units that serve as benchmarks for measuring Decision Making Unit (DMUs) comparative efficiency. The Stochastic Frontier Analysis (SFA) uses a hypothesized function to calculate estimates of the efficiencies of individual DMUs. Only SFA can separate random noise from efficiency; DEA incorporates noise as part of the efficiency score (ÖNDER, DELIKTAŞ, & KARADAĞ, 2013).

Each of these techniques has its associated strengths and weaknesses, so that a trade-off exists in selecting the 'correct' approach. Compared with two ratio methods (simple ratio and average cost analyses) frontier techniques deliver benchmarks that reflect industry best-practice and efficiency scores that reflect deviations between observed and potential performance. In contrast, simple ratios benchmark hospitals against average industry behaviour. Industry-average benchmarks are quite useful for setting reimbursement rates but problems arise if we want to say anything about the gap between observed and potential performance, or about how hospital decision-makers might go about bridging this gap (Mortimer & Peacock, 2002).

Whereas DEA and SFA evaluate hospital performance over multiple dimensions, simple ratios take account of only two summary dimensions: typically input (or cost) and output. Because it is

more difficult to make multi-attribute comparisons using simple ratios, there is no guarantee that inefficient (efficient) organizations lie above (or below) the industry-average level of cost per output.

DEA seems to produce different rankings from the traditional production function, and thus the deterministic production frontier. Studies (Burrow, 1988; Schimdt, 1986 and Wagstaff, 1988) suggest that estimates of inefficiency can change dramatically when allowance for statistical noise and random shocks are made. Also, the advantage of being relatively flexible in modelling the underlying production technology but makes no adjustment for random error.

The SFA has the disadvantages of both parametric and deterministic approaches, that is the latter has the advantage that it accounts for stochastic error, which is not included as part of the measure of (in)efficiency but its disadvantage is the imposition of a parametric structure on the production function and on the distribution of efficiency (Banker, Conrad & Strauss, 1986; Fare,Grosskopf & Lovell, 1985; Ganley & Cubbin, 1992; Silkman, 1986 and Wagstaff, 1992).

The dominance of DEA in healthcare efficiency measurement studies is its ability to quantify multiple inputs and outputs in different units of measurement is one consideration. For example, many healthcare studies (Grosskopf and Valdmanis, 1987; Valdmanis, 1992 & Parkin and Hollingsworth, 1997) have defined inputs as the number of doctors, nurses and ancillary staff along with non-labor inputs in dollar terms, especially buildings and equipment assets. Also, outputs are often defined as the number of patient days, surgeries or discharges, (Wagstaff, 1989 and Gonzalez Lopez-Valcarcel & Barber Perez, 1996).

Even though the DEA has a number of advantages, it also has disadvantages. DEA scores are biased and serially correlated (Simar & Wilson, 2007). Therefore the inferences in two stage approaches are invalid (Simar & Wilson, 2007). Simar and Wilson (2007) propose a double bootstrap method that enables researchers to correct for serial correlation, measurement error and to stimulate a true sampling distribution by mimicking the data generating process (Latruffe, Davidora, & Balcombe, 2008). The procedure in this study follows the double bootstrap.

The input-oriented DEA model is chosen because of its common use, ability to correct for bias and correlation. In this study it focuses on minimizing the use of inputs for producing the given amount of outputs. This model fits with the context of primary healthcare since these facilities can control the inputs such as staffing and operating expenses rather than the number people who visit the facility seeking healthcare services (Nayar, Ozcan, & Nguyen, 2013).

4.3 Data Envelopment Analysis

DEA is a non-parametric linear programming technique developed by Charnes et al (1978) to estimate efficiency and capacity utilization and effectively identifying the production frontier. This was progressing from Farrell's (1957) single input-output productive efficiency concept into an efficiency measure of DMUs involving multiple inputs-outputs to calculate a best practice efficient production frontier.

DEA is based on the principle that the performance of each DMU must be compared relative to the 'best-practice' frontier, that is a benchmark continuum of highly-efficient, virtual DMUs. The 'best-practice' virtual frontier is essentially the convex combination of all efficient points of operation. In this technique, any deviation from the 'best-practice' frontier must be an indication of technical inefficiency. DEA is the most extensively employed method in production frontier analysis, does not require a *priori* assumptions on the functional form and can handle multiple inputs and outputs (Charnes, Cooper & Levin, 1997).

Literature reviews conducted for the measurement of efficiency in health care found that there is a lot to be learned from empirical studies, particularly regarding the interpretation of results derived from frontier analysis. These studies exploring the technical efficiency in health care used their findings to inform policy decisions, such as to identify ways of achieving resource savings and possible improvement of efficiency scores.

The DEA measures relative efficiencies (i.e., relative to a best practice frontier) and allows for specialization in one or another input or output variable, and the technique also does not handle measurement errors (Charnes, Cooper & Seiford, 1978), hence the need to bootstrap. DEA has certain inherent inefficiency created by noise, formed by the distance from the efficient boundary. Bootstrapping helps to overcome these efficiencies for bias and to develop the correct confidence intervals, whilst accepting that the data has random noise.

The technique assumes homogeneity, that is it assumes that primary healthcare facilities are performing identical functions and thus produce similar outputs in order to be identifiable; it also assumes that identical resources are available to all DMUs and that the operate in a similar environment (Hollingsworth B. , 2012).

In order to account for influences of the distribution of health complexity on producer efficiency in the production of health care, involves adding a second stage of analysis to the DEA approach.

The first stage involves running a DEA model based on inputs and outputs to yield efficiency scores for units (say primary healthcare facilities again). The second stage then takes these efficiency scores and statistically regresses them against primary-level variables to assess the impact of the other inputs (number of nurses and doctors, number of patients attended in the year) on the production process and efficiency. This allows the inclusion of variables that do not fall neatly into the input–output analysis to potentially see if they have a significant impact on the efficiency scores obtained in the first stage (Simar & Wilson, 2008).

4.3.1 First stage: Estimation of DEA efficiency score

The original two DEA models for measuring technical efficiency are the Charnes, Cooper and Rhodes (1978) based on the concept by Farell (1957). Farell explained the economic efficiency that could account for multiple inputs. Further, he decomposed the productive efficiency into technical and allocative efficiency. In this approach, the technical efficiency is an ability of a firm to maximize output from a set of given inputs and allocative efficiency is an ability of a firm to use these inputs in optimal proportions, given their respective prices,

The pioneers of DEA based their model on Farell's efficiency. Their model, also known as the CCR model after their names, is based on constant return to scale assumption. The second DEA model is known as BCC model named after the names of Banker, Charnes and Cooper. Coelli et al. (1998) observed that the CRS assumption may not always be appropriate in some real life contexts and therefore cannot be applied in variety of circumstances. To overcome this difficulty, Banker et al. (1984) developed the BCC model which is based on VRS technology assumption.

The efficiency score in the presence of multiple input-output variables is defined as:

Efficiency = weighted sum of outputs/weighted sum of inputs

Suppose we have nDMUs. Each DMUj, j = 1, 2, ..., n produces s different outputs, yrj (r = 1, 2, ..., s), using m different inputs, xij (i = 1, 2, ..., m). Then the primal linear programme for the (input-based) CCR model can be written as (Equation 1):

$$\theta = \min\theta$$

$$\sum_{j=1}^{n} \lambda_j X_{ij} \leq \theta X_{io} \qquad i = 1,2, \ldots \ldots m$$

$$\sum_{j=1}^{n} \lambda_j Y_{rj} \geq Y_{ro} \qquad r = 1,2, \ldots \ldots s$$

$$\lambda_j \geq 0 \qquad j = 1,2, \ldots \ldots n$$

where x_{io} and y_{ro} are, respectively, the *i*th input and *r*th output for DMU_o under evaluation. Associated with the $m + s$ input and output constraints, some non-zero slacks may be identified by utilizing a two-stage process where the efficiency score is first calculated and then the sum of slacks is maximized (Ali & Seiford, 1993). A DMU_o is said to be CCR efficient if and only if (a) y = 1 and (b) all optimum slack values in Eq. (1) are zero.

The second DEA model is referred to as the BCC model and it a modification of the CCR model by applying a much realistic assumption of the variable return to scale (VRS) wherein the DMU is allowed to display different returns to scale because of a different environment. BCC model measures pure technical efficiency

The CCR model measures the relative technical efficiency, while BCC model measures the pure technical efficiency, net of scale-effect - i.e. conversion of inputs into output. Pure technical efficiency is also known as managerial efficiency. If a DMU scores value of both CCR-efficiency and BCC-efficiency equal to one, it is said to operate at the most productive scale size. Actually, BCC model helps to decompose the relative technical efficiency into pure technical efficiency and the scale efficiency.

Scale efficiency of a DMU is measured by dividing the relative technical efficiency from pure technical efficiency. For example, if pure technical efficiency of DMU_1 is equal to 1 and its relative technical efficiency is less than 1, it implies that DMU_1 is able to converts efficiently its inputs into output, however, it is relative technical efficiency is inefficient because it may be too small or too big relative to optimum size. Thus, inefficiency in any firm may occur due to its inefficient operations or due to the disadvantageous conditions under which it operates.

The DEA technique have two basic versions of the CCR model. One is the input-oriented, while the other is output-oriented. The aim of the input oriented CCR model is to minimize inputs for

producing given levels of outputs. The purpose of the output-oriented CCR is to maximize outputs by using no more than given amounts of inputs (Mirmiran, Li, & Ilacqua, 2008).

Let's consider the input-oriented CCR model that assumes that there are *s*-health care facilities (DMUs), each with a set of *j* outputs (y_{ir}, i = 1, 2, ⋯,j,r = 1, 2, ⋯,s) and a set of *l* inputs (x_{kr}, k= 1, 2, ⋯, l, r = 1, 2, ⋯,s). The efficiency parameter (θ_r) of the *r*th DMU is determined by solving the following fractional programming problem to obtain the output weights u_i (i= 1, 2, ⋯,j) and the input weights v_k(k= 1, 2,⋯, l).

Maximise $$\theta_r = \left(\sum_{i=1}^{j} u_r y_{ir}\right) / \sum_{k=1}^{l} v_k x_{kr}$$

Subject to $$\theta_r \leq 1,\ r = 1\text{……..s}$$

$$u_i \geq 0,\ i = 1, 2\ \text{……...j, and}$$

$$v_k \geq 0,\ k = 1, 2\text{……..}l.$$

The input-oriented CCR model, as outlined above, poses no additional constraints. It determines input weights and output weights, which in turn determine the efficiency parameter (θ) of each DMU. When evaluating the efficiency of healthcare facilities, it is not necessary to make any subjective assignment of weights. Healthcare facilities with an efficient score of "1" are said to be efficient and on the efficient frontier; otherwise inefficient and below the frontier. The extent of inefficiency of a DMU is measured by the deviation of its efficiency parameter value from 1.

When the policy directive in a health system is to minimise inputs for given levels of outputs, using the input-oriented CCR is more appropriate for efficiency evaluation. Alternatively, if the aim is to maximise outputs with the given the amount of inputs, the use of output-oriented CCR becomes more suited. In this study, we use both input and output oriented CCR models to test efficiency under either condition, even though the most appropriate for decision making is the output-oriented CCR where we test the levels of outputs given the inputs that are assumed to be homogenous.

This study uses the input oriented model because the primary healthcare facilities serve the public demand as given and must manage their resources accordingly, therefore, they target input minimization rather than output maximization which recommends using the input-oriented DEA models (Ozcan, 2008).

Also, this approach helps to plan ahead should the conditional grant allocation be reduced to a point where the inputs cannot be increased but the need for higher level of production (more people being put on treatment) is imperative.

4.3.2 Second stage: Bootstrap truncated regression

Truncated regression is used to model dependent variables for which some of the observations are not included in the analysis because of the value of the dependent variable.

In this study we applied the two-stage procedure (DEA + truncated regression, bootstrapped) of Simar and Wilson (2008) to examine the determinants of efficiency of primary care services. In the second stage of analysis, we examined potential determinants of efficiency through truncated regression with bootstrapping (following the approach of Simar and Wilson (2008)). Along with dummy variables for number of doctors, number of nurses, number of staff paid under the conditional grant and total headcount), this allowed for the conditioning on other variables that are important for health production and tested their significance.

DEA has been observed to be easily affected by the presence of outliers, especially with a small sample size ((Dyson, et al., 2001). Therefore, the bootstrapping approach was introduced (Simar & Wilson, 2000) to improve the reliability of the efficiency estimates. For this approach, efficiency estimates were regressed on a set of variables then the results were approximated to the sampling distribution of the original estimates. That is, it was able to obtain bias-corrected efficiency estimates and confidence intervals by using the bootstrapping approach.

The main idea in Simar and Wilson (2008) about the data generating process is that efficiency θ_i linearly depends on z_i:

$$\theta_i = z_i \beta + \varepsilon_i$$

where β denotes a vector of coefficients, estimating which is the ultimate objective of the empirical analysis. The disturbances ε_i are assumed to be – conditionally on z_i – independently, truncated normally distributed, with parameters $\mu = 0$ and σ, and left-truncation at $1 - z_i\beta$. This assumption guarantees that θ_i cannot be smaller than unity, irrespective of the values the variables in z_i may take. Though full efficiency ($\theta_i = 1$) is in principle possible, it occurs with zero probability.

4.4 Data and variables

In a DEA model, two main categories are considered: input and output. Health programmes that produce more outputs with given inputs are considered to be relatively more efficient. For the purpose of this research, outputs are a product of the health programme (ART provision), i.e. as antiretroviral treatment services provided to patients. Inputs are factors that significantly affect the production of the outputs.

It is important to choose the output variable/s that best describe the performance of the healthcare system and the input variables that can have a direct impact on the output variables. Since subjective health outcomes such as healthcare satisfaction are not easy to quantify, the objective health outcomes such as total remaining in care (i.e. people on antiretroviral treatment) and total number tested have been widely used in healthcare system efficiency studies (Retzlaff-Roberts, Chang, & Rubin, 2004; Hadad, Hadad, & Simon-Tuval, 2013; Asandului, Roman, & Fatulescu, 2014)

The World Health Organisation (2016) suggests a framework for thinking more clearly about the efficiency measurement. The five key aspects of any efficiency indicators are outlines as follows:

- The **entity** to be measured: the very micro level of analysis must be considered but the principle is that the entity must reflect an entity where there can be clear accountability,
- **Outputs** to be considered: the outputs should be defined in terms of health gains,
- **Inputs** to be considered: efficiency metrics should be developed according to the intention of the analysis. A measure of total costs may be an adequate indicator of an entity's input/s,
- **External influences** on performance: these are influences on the entity beyond its control and may include, for an example, population, intensity of resources, etc,
- **Link to the rest of the health system:** scrutiny of an entity in isolation may ignore important implications of how the entity impacts on the whole system.

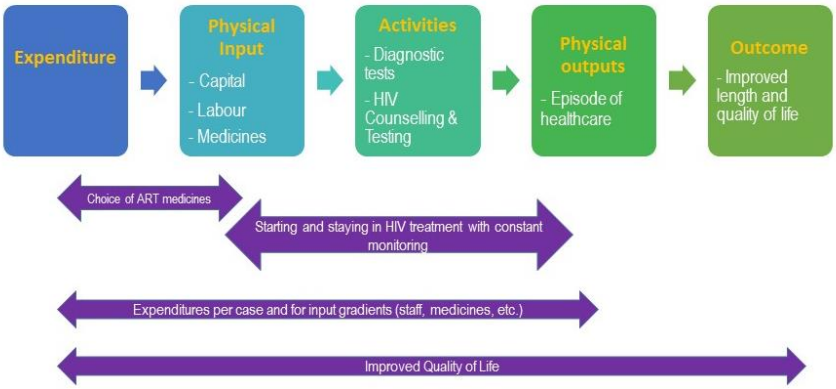


Figure 4.1: The production process in ART delivery with common indicators along the pathway

The figure above (Figure 4.1) illustrates the health production process, represents a typical (but simplified) sequence of processes associated with the treatment of an HIV positive person, along with a set of common indicators associated with a common production pathway. The greatest concern is the cost-efficiency, which is the transformation of expenditures (on the far left) into valued health outputs and outcomes (on the far right

According to Joumard et al (2010), the resources that determine the health status of the population can be measured in monetary terms, or physical terms, lifestyle factors and socio-economic factors. It is important to add here that, human resources fundamental and critical components of the health production. Also, Cooper et al. (2006) suggested a rough rule of thumb which can provide guidance to choose a value of n that satisfies $n \geq \max [m \times s, 3(m + s)]$, where n is the number of DMUs, m is number of inputs and s is number of outputs.

The fundamental assumption in the choice of variables is that the primary healthcare facilities are comparable. It is worth noting that the fact that the difference in output being compared is largely different could be as a result of the quality of care provided.

The output measure represents the intermediate output and not the final health outcome (such as longer life expectancy) in the DEA model. This is not problematic as there is sufficient evidence that it leads to health improvement, i.e. where treatment is taken as directed (Nguyen & Coelli,

2009). The data for the year 2017/18 was collected from the KwaZulu Natal Department of Health reports.

Table 1: Input and Output variables analysed

Primary Healthcare Facilities	Input	Output
	Staff Cost of Employment (CoE), Cost of Laboratory Tests(CLT), and Cost of ARV drugs	Total Remaining on Antiretroviral Treatment (TROA)

The population of primary healthcare faciliies in the district re are 88 primary healthcare facilities and a sample size are 24 primary healthcare facilities (hereunder referred to as Decision Making Units (DMU)) in uMgungundlovu District. Expenditure data for the year 2017/18 on the cost of empoyment, laboratory tests and cost of antitroviral treatment were obtained from the KwaZulu Natal Department of Health's BAS expenditure reports and PERSAL (for the cost of employment). The total remaining on antiretroviral treatment (TROA) was obtained from the Department's Department of Health Information System (DHIS).

The facilities were randomly selected and consist of three community healthcare centres (CHC) that provide similar services as primary healthcare centres including 24 hours maternity service, emergency care and a casualty and a short stay ward. The CHC will refer a patient to a district hospital when necessary. Data were available for all 25 primary healthcare facilities as inclusion criteria; in addition, there were no exclusion criteria and no missing data which is in line with Cooper et al (2006) rule of thumb.

Table 4.2: Facility (DMU) data showing annual spending per unit

	Input variables			Output variable
DMU	**ART**	**COE**	**CLT**	**TROA**
DMU1	6,986,406	7,221,713	1,807,452	3,686
DMU2	31,055,133	8,371,257	5,455,614	1,792
DMU3	22,025,489	12,142,481	3,362,549	6,340
DMU4	2,544,770	1,119,575	388,818	1,694
DMU5	401,945	316,324	73,789	321

DMU6	10,409,224	2,189,491	2,541,040	6,310
DMU7	6,256,429	1,523,008	2,778,788	5,192
DMU8	1,831,865	748,682	9,25,323	1,792
DMU9	1,297,174	662,799	3,68,777	755
DMU10	1,504,968	354,473	411,896	1,154
DMU11	3,428,451	788,443	651,123	1,966
DMU12	2,469,109	1,494,224	721,299	2,193
DMU13	4,868,215	352,041	704,534	2,966
DMU14	3,363,576	799,214	74,081	1,828
DMU15	213,835	2,86,354	1,020,619	159
DMU16	2,385,772	1,368,402	2,033,338	2,479
DMU17	11,454,759	1,644,595	599,861	4,100
DMU18	1,464,046	858,887	187,710	1,465
DMU19	763,839	1,040,141	401,165	537
DMU20	1,455,686	1,732,109	1,128,840	968
DMU21	4,666,853	761,500	845,721	3,763
DMU22	6,296,057	1,687,934	1,342,022	3,597
DMU23	2,283,018	771,927	672,353	1,845
DMU24	911,302	374,246	427,955	902

4.5 Chapter Summary and Conclusion

Hollingsworth (2003) reviewed over 180 papers that measure efficiency and found, among others, that the DEA method is the most popular method of choice in health care and hospitals. There are important lessons to be learnt from prior experience using DEA to measure efficiency in healthcare, particularly with regards to how best to implement and interpret such measures. The DEA is useful when based on sound data, a robust and valid model, and when the limitations of the method are well understood.

Additionally, the DEA is most relevant when the results are presented in a manner that is easily understood by those who are tasked with making the changes in service delivery.

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Introduction

This chapter presents the results of the DEA analysed for DMUs (public clinics) to address the objectives of the study. This study employed the R-Analytics software to measure the relative efficiency of primary healthcare facilities and (in the second stage) used the bootstrap truncated regression to correct for the errors and noise. The chapter starts by describing the descriptive analysis of the input and output data and further provides empirical results for the study.

5.2 Descriptive Statistics

5.2.1 Statistics of DEA variables

Table 5.1 illustrates the descriptive analysis as follows: The average annual ART spending is R5,430,746,71 with the minimum of R213,835 and the maximum expenditure of R31,055,133. The cost of employment average R2,025,409,17 with a minimum of R286, 354 and a maximum of R12, 142,481. It is worth noting that the total amount of R286,354 was the minimum cost spent on employment in the sampled clinics. Only the public laboratories (National Health Laboratory Services) were used, with the cost of laboratory tests used to describe this variable. The mean of R1,205,194,46 was reported for the laboratory tests that were conducted. The maximum and the minimum costs were R5,455,614,00 and R73,739,00 respectively.

The only output that was used in the study was total cost remaining on care, and on average 2,409 was used with the maximum of 6,340 in the clinics. This variation shows a vast difference in performance of DMUs. The disparities are an indication the cost of labour and that of antiretroviral treatment which is an external factor that cannot be controlled by the department and/or the primary healthcare facilities. It may be of interest to see this in comparison to the private sector so as to be able to make a clearer determination if it has to do with efficiency. Also important is the fact that treatment is not accounted in terms of the first, second or third line which affects the unit costs. Again, the cost of labour is taken as a total without clearly defining the category of employees (e.g. nurses, doctors, etc) which may have been a cause for the large disparities in the cost of labour.

Table 5.1: Descriptive statistics for output and input variables

Variable	Mean	Std. Dev.	Min	Max
Annual ART Spending	R5,430,746,71	R7,265,397,85	R213,835	R31,055,133
Cost of Employment	R2,025,409,17	R2,934,572,92	R286,354	R12,142,481
Cost of Laboratory Tests	R1,205,194,46	R1,261,098,75	R73,789,00	R5,455,614,00
Total Remaining on Care (people)	2,409	1,758	159	6,340

As described in the methodology, the results of the DEA were regressed with operational variables that included number of doctors and nurses in primary healthcare facilities, people who visited the clinic regardless of the service they were receiving, and lastly staff in the facilities that were paid using an HIV/AIDS grant. This is the second stage of the DEA where operational variables that are perceived to be important in the health production and are tested for significance. Table 5.2 below provides descriptive statistics of these variables.

Table 5.2: Descriptive statistics for operational variables

Variable	Mean	Std. Dev.	Min	Max
Number of Doctors	0.37	1.05	0	4
Number of Nurses	4.92	5.85	1	26
People Visited Clinic	59,447.75	55,094.75	7,870	209,893
HIV/AIDS Staff paid from grants	8.21	10.29	1	41

These variables were used to get a clear policy recommendation that may improve the operation of primary healthcare facilities. Table 5.2 shows that on average 0.37 doctors were operating in primary healthcare facilities. This observation was accompanied by some facilities with no doctors, while the maximum number of doctors that operated in a facility was four. Doctors were assisted by nurses. In primary healthcare facilities nurses play a pivotal role in the operation (and management) of the clinics.

On average 4.92 nurses were operating in the clinics, with a minimum and maximum of 1 and 26 nurses, respectively in operation. The capacity of nurses in assisting people who visit the clinic is important for operational management. The average number of people regardless of their age that visited the facilities were 594,47.75, with the minimum and maximum of 7,870 and 209,893 per annum respectively. The number of staff that are paid from the HIV/AIDS grants were 8.21 on average, and there was a minimum of one staff per facility and a maximum of 41.

The variations may explain the lag between those who need treatment and those in treatment as being caused by systemic failures do, in part, to staff shortages and unconducive clinic environment (space). Therefore, decision makers must put more attention to resource allocation (with regards to staff in relation to workload and also the size of the clinic). The demand for services may also be inadequate in other clinics especially in the light of changed policy where every person must be offered an HIV test and those testing HIV positive must be started on HIV treatment.

5.3 Results of the DEA Approaches

5.3.1 Input-oriented technical and scale efficiency

This study reported the variable returns to scale technical efficiency and the scale efficiency. Table 5.3 illustrates the frequency distribution of technical and scale efficiency and reveals different variations. The mean technical efficiency score of 86% was reported for primary public facilities, and the mean scale efficiency was 89%. Moreover, out of 24 primary healthcare facilities only 14 DMUs were fully efficient (TE = 1), followed by the group of four that scored mean score between 70 – 99% while the remaining six facilities recorded mean technical efficiency scores ranging between 50 – 70%. The results further reveal that 58% (14) of the primary public health facilities were fully efficient. These findings are important for service delivery in that policy makers are able to know how well the primary healthcare facilities are utilizing the available resources to improve the performance of primary healthcare by 14% so that all facilities are fully efficient.

Table 5.3: Frequency distribution of input-oriented technical and scale efficiency

Frequency (%)	Technical Efficiency (N = 24)	Scale Efficiency (N = 24)
< 10	0	0
10 - 30	0	1
30 - 50	0	0
50 - 70	6	0
70 - 99	4	14
100	14	09
Mean	0.86	0.89
Standard deviation	0.22	0.16
Minimum	0.06	0.29
Maximum	1.00	1.00

From the study we observe that the mean scale efficiency score of 89% and a technical efficiency score of 86%. The study has revealed that 86 percent of primary healthcare facilities in

Umgungundlovu are technically efficient in the period 2017/18. This result implies that the average primary healthcare facilities can produce the same level of output with 14 percent less of inputs. In other words, there was increased scope for the primary healthcare facilities to increase their output if they had operated at the same efficient level as the most efficient primary healthcare facilities in the sample. This means money saved from these facilities could diverted to other areas of need or increase service delivery so as to improve access. Again, these scores are good indicators of the total costs that could be saved since technical and scale efficiency are necessary conditions for cost minimization.

The total of 14 facilities reported a mean score ranging between 70 – 99 scale efficiency, of these 4.16% of the facilities reported mean scores in the range of 10 – 30%, and 58% were of the range between 70 – 99. Overall the results mean that the technical inefficiency of facilities is a result of scale inefficiency rather than pure technical inefficiency. The results mean that most of the facilities are operating at the wrong scale of operations. Specifically, the facilities were operating under decreasing returns to scale, where there is still opportunity to increase operations to obtain optimum scale.

5.3.2 Distribution of returns to scale for input-oriented technical efficiency

Table 5.4: Input-oriented technical efficiency returns to scale

Returns to scale (%)	Percent (N = 24)
Constant return to scale	23.08
Decreasing return to scale	42.30
Increasing return to scale	34.62

Table 5.5 illustrates that 23.08% of primary healthcare facilities performed optimally at constant return to scale - meaning that only 23.08% of primary healthcare facilities where the number of people put on treatment (output) increases as the factors of production increase. While 42.30% of facilities operated at decreasing return to scale, this means that even though the factors of production are increased, the proportion of output increase is not at the same rate or smaller.

In order for primary healthcare facilities to optimize the provision of antiretroviral treatment and increase technical efficiency, there is need for proper allocation of inputs. The total of 34.62% facilities operated in an increasing return to scale and their production is below the optimal scale,

meaning that the output is higher than an increase in the factors of production. This implies that staff could be redeployed to other facilities so as to optimize efficiency across facilities.

5.4 Results of the Bootstrap Truncated Regression Model

In the second stage of the analysis of this study the bootstrap truncated regression model was bootstrapped and regressed against the results of input-oriented technical efficiency to investigate the operational related variables of technical efficiency. Based on the analysis, Tables 5.6 and 5.7 show the results on the bootstrap truncated regression model with both variable return to scale and constant return to scale technical efficiency mean score used at dependent variables to the number of nurses, the total staff that were paid from HIV/AIDS grants and the headcounts of clients who visited the primary healthcare facilities.

Table 5.5: Bootstrap Truncated Regression results for VRS determinants of TE

Variable	Estimate	Standard Error	Pr(>ItI)	Confidence
Nurses	-1,097	0,785	-1,398	0,176
Grants	-0,967	0,417	-2,317	0,030
Headcounts	-0,001	8,057	-1,900	0,070

The findings in Table 5.6. show the negative relationship between the number of nurses, the staff that are funded by grants and the total number of the people visiting the primary healthcare facilities and technical efficiency in providing antiretroviral treatment.

A point to ponder is in relation to the ability of these facilities to provide antiretroviral treatment in relation to their operational abilities. The negative relationship might be attributed to the shortage of nurses, low number of staff that are funded by grants, and the high volumes of people visiting the clinic, and the demand for antiretroviral treatment. Generally, nurses are expected to deliver their services to everyone who visits the faculties, regardless of the services needed. We must stress that all the three variables were not significant at any level.

Table 5.6: Bootstrap Truncated regression results for CRS determinants of TE

Variable	Estimate	Standard Error	Pr(>ItI)	Confidence
Nurses	-2,848	0,627	-4,539	0,000
Grants	-1,777	0,319	-5,557	1,381
Headcounts	-0,002	6,812	-4,327	0,002

The findings in Table 5.7 reveal that the number of nurses was negatively related to constant return to scale technical efficiency in primary healthcare in the Umgungundlovu District. Moreover, like in the VRS model the staff that were paid from grants exhibited the negative relationship with technical efficiency in relation to the provision of antiretroviral treatment of primary public healthcare facilities. Lastly, the total number of people who visited the primary healthcare facilities was also not related to technical efficiency.

CHAPTER 6:
SUMMARY, CONCLUSION AND RECOMMENDATIONS

6.1 Introduction

There has been great progress in the provision of antiretroviral treatments to people who need them in South Africa, and in KwaZulu Natal province in particular. In the thirty years of HIV South Africa has continuously adapted the service delivery models so as to ensure maximum coverage of, and access to, life prolonging treatment. This is evidenced by considerable increase in health budgets and the expansion of the network of health facilities across the country. The policies that the government has pursued over the years have had a direct impact in improving the health status of citizens. These successes notwithstanding, reversals in gains made are possible if the efficient use of resources is not prioritized and made a key managerial function. Reversals are also possible as resources shrink while the demand increases.

In order to address these problems, the Department of Health needs to adopt a smart investment framework with a goal to provide a guide to increasing efficiency in resource allocation, and make this a key strategic imperative. One of the health sector reform objectives needs to be increased levels of efficiency across the health system, and in HIV/AIDS programmes in particular. This study is a start in estimating the level and analysing the trend of efficiency in the ART programme.

The literature indicates that two main approaches are used in estimating the efficiency of a decision making unit, i.e. stochastic frontier approach (SFA), and data envelopment analysis (DEA). The DEA approach was used in this study. A random sample of 24 public clinics (52%) of all public clinics in the uMgungundlovu District was used in the study.

6.2 Summary of Main Findings

The findings showed that 58% (14) of the primary public health facilities were fully efficient. These findings are important for service delivery, for policy makers to note the need to improve the technical efficiency mean score by 14% for all facilities to remain fully efficient.

The mean technical efficiency score of 86% was reported for primary public facilities, and the mean scale efficiency was 89%. Moreover, out of 24 primary healthcare facilities only 14 DMUs were fully efficient (TE = 1), followed by the group of 4 that scored mean score between 70 – 99% while the remaining 6 facilities recorded mean technical efficiency scores ranging between 50 – 70%.

DMU 2 is an outlier. In a paper by Khezrimotlagh (2013) it is suggested that the presence of outliers may be classified in four types. The first type is the outcome of recording/measurement errors. This type is commonly used to introduce outliers in the DEA literature. The second type is from the lack of selecting inputs and/or outputs. In other words, if inputs/outputs of DMUs are not correctly selected, it may cause presenting outliers even if there is no measurement/recording error in the data. In this type, adding an appropriate input/output may rectify the issue. Note that drawbacks of selecting inputs and outputs are not from the structure of DEA, however, DEA is able to measure the impact of each factor on the measured efficiency score by applying key account management (KAM). The third one is when there is a non-homogenous DMU among observations. If data are exact and the inputs/outputs are appropriately selected, a non-homogenous DMU may clearly cause outliers. The fourth type is when a DMU, in comparison with another DMU, uses a much greater (bit less) value of input and only produces a bit greater (much less) value of output. This outlier falls in the fourth type as it uses much greater inputs to achieve bit less output. For an example the input spend for DMU 3 (R37.5 million) is less than that of DMU 2 (R44.8 million) but the output is three times greater in DMU 3 (6,340) than DMU 2 (1,792).

6.3 Policy Recommendations

The challenges facing healthcare delivery are fast becoming more complex; we have less acute diseases and more chronic ones that significantly depend on demographics and lifestyles. As complex problems are caused by multiple factors that interact in complex ways (Horne & Shirley, 2009), linear approaches to service delivery can only be partially effective.

As Burns et al. (2006) indicated, traditionally organizations have been designed for a complicated rather than a complex world. Hierarchical and structures are perfectly designed to break problems down into more manageable fragments. They are not, however, so effective at handling high levels of complexity. For this reason, many of our most long-standing institutions are now struggling to adapt to a more complex world.

Where the executive performance is framed around efficiency, they will guide the organization's culture and also spare no effort to maintain a positive culture where it already exists. The department of health needs to put efficiency at the center of its operations, especially in regards to the HIV/AIDS Conditional grant.

This means that improving efficiency should be coupled with a process of redesigning services so that 'there is more choice, more personalised care, real empowerment of people to improve their health', moving from 'a service that does things to and for its patients to one which is patient-led, where the service works with patients to support them with their health needs' (Burns et al., 2006). This can be realized only at the lowest point of service (i.e. clinics) where the context becomes important, and where managers have both the capacity and authority to make the necessary decisions appropriate in their context.

Therefore, the current arrangement where clinics depend on hospitals, and do not manage their own resources, needs to be reviewed in terms of both resource allocation and management. It may be important to investigate an arrangement where clinics are allocated capacity so that they are directly responsible for their own resources. This means making clinics cost centres where operations managers are responsible for planning, budgeting and management of resources.

The entry point for healthcare delivery is a clinic and therefore the success of the National Health Insurance (NHI) will largely depend on clinics that operate efficiently and optimally. This requires that managers are sufficiently empowered to make on the spot decision and also put in place measures to improve efficiency.

Again, the incremental budgeting that is currently in use makes an erroneous assumption that is not based on operational realities, i.e. it assumes that the primary cost drivers (input costs) do not change from year to year. However, this method perpetuates inefficiencies. For example, if a clinic manager knows that there is an opportunity to grow his/her budget by 10% every year, they will simply take that opportunity to obtain a bigger budget, while not putting effort into seeking ways to cut costs or economise.

Also, it does not take account of the possible increase in headcount (people who will visit the clinic at a certain time) because they may not like the service in their own clinic. This is likely to result in budgetary slack. It is also likely to ignore external drivers of activity and performance. For example, there is very high inflation in certain input costs. Incremental budgeting ignores any external factors and simply assumes the cost will grow by, for example, 10% in 2019/20.

6.4 Limitations of the Study

This study focused only on technical efficiency and did not look into allocative efficiency. Thus the scores do not capture total efficiency or inefficiency.

Since the data are only for a period of 2017/18, the study does not estimate total factor productivity change /and technological changes.

The study used proxy output measure. One of the intrinsic goals of health systems is to improve beneficiaries' health status, and thus the ultimate outcome of the ART programme ought to be its effect on life expectancy and health-related quality of life of all those who receive treatment. Therefore, ideally, the study should have used indices that combine the two dimensions of health into a unitary measure. The readers can refer to Kirigia (1998) for details relating to the application of quality-adjusted life years (QALY) index in Africa; and to Murray (1994) for the disability-adjusted life years (DALY) index. The study opted for proxy outcome measure because of lack of clinic level information needed to calculate QALYs or DALYs for individual clinics.

6.5 Further Areas of Study

There is a need for health systems researchers to conduct detailed investigation studies on a sample of the relatively efficient health facilities to document key attributes of the best practice. Such best practice could then be emulated by the inefficient clinics.

6.6 Conclusion

Using data from 20117/18, we measured the technical efficiency of primary healthcare facilities offering antiretroviral treatment in uMgungundlovu District with an aim of identifying inefficient facilities. We analyzed technical efficiency of 24 facilities and found heterogeneity in technical efficiency scores which is explained by the fact that the facilities are not homogenous and do not have the same productive capacity.

The analysis, it is hoped, will build on a set of variables that conform with the nature of services in the South African HIV/AIDS health services. The output measure conceptualizes the volume of benefit delivered to the public regarding HIV/AIDS treatment programme. The input measures considered the primary healthcare HIV/AIDS operation taking into account the cost of labour, medicines and laboratory tests. The study used the basic data envelopment analysis and further employed a second stage DEA analysis. The work incorporates the environmental factors that may influence the inefficient performance of facilities.

The main findings of the study are: (1) 14 of the 24 (58, 3%) facilities are fully technical efficient and 9 of the 24 (37,5%) are scale efficient which implies that there could be savings in resource

consumption without reducing the volume of services, (2) health managers and policy makers need to constantly use technical efficiency measures in order to facilitate better resource utilization, control of the cost of health services and consequently to provide more affordable healthcare, (3) there is a need to revise resource allocation policies so that where facilities work under increasing returns to scale (34,62%) can share their workload so that they can reach their optimal size of operation, and (4) managers of inefficient facilities must direct their efforts to benchmarking best performers.

Further studies need to be conducted to explore the causes of inefficiency and policy makers can develop context-based strategies for the inefficient facilities to improve efficiency in delivering HIV/AIDS health services which may be used to address the unmet needs and wastages in the delivery HIV/AIDS services.

REFERENCES

Abdool Karim, S., Churchyard, G., Abdool Karim, Q., & Lawn, S. (2009). HIV inefction and tuberculosis in South Africa: an urgent need to escalate the public health response. *Lancet*, 921-933.

Ali, A., & Seiford, L. (1993). The mathematical programming approach to efficiency measurement. In L. S. Fried, *The Measurement of Productive Efficiency: Techniques and Applications.* (pp. 98-123). London: Oxford University Press.

Anderson, S., Cherutich, P. K., Cremin, I., Fetch, D., & Kimanga, D. (2014). Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *Lancet*, 249-256.

Asandului, L., Roman, M., & Fatulescu, P. (2014). The efficiency of healthcare systems in Europe: A data envelopment analysis approach. *Procedia Econ. Financ*, 261–268.

Atun, R., Bennett, S., & Duran, A. (2008). *When do vertical (stand-alone)programmes have a place in health systems?* Copenhagen: WHO.

Babigumira, J. B. (2017). Types of Economic Evaluation in Healthcare . *Global AIDS Research* (pp. 1-26). Washington DC: Global Medicines Program .

Babigumira, J. B., Castelnuovo, B., Lamorde, M., Kambugu, A., Stergachis, A., Easterbrook, J. B., & Garrison, L. P. (2009). Potential impact of task-shifting on costs of antiretroviral therapy and physician supply in Uganda. *BMC Health Services Research 9*, 192.

Babigumira, J., Castelnuovo, B., Stergachis, A., Kiragga, A., Shaefer, P., Lamorde, M., . . . Garrison, L. (2010). Cost effectiveness of a pharmacy-only refill program in a large urban HIV/AIDS clinic in Uganda. *PLoS ONE*, e18193.

Babigumira, J., Sethi, A., Smyth, K., & Singer, M. (2009). Cost effectiveness of facility-based care, home-based care and mobile clinics for provision of antiretroviral therapy in Uganda. *Pharmacoeconomics*, 963–973.

Baily, M. N., Garber, A., Berndt, E., & Cutler, D. (1997). Health Care Productivity. *Brookings Papers on Economic Activity. Microeconomics*, 143-215.

Banker, R., & Thrall, R. (1992). Estimation of returns to scale using data envelopment analysis. *European Journal of Operational Research*, 74-84.

Banker, R., Bardhan, I., & Cooper, W. (1996). A note on returns to scale in DEA. *European Journal of Operational Research*, 583-585.

Banker, R., Conrad, R., & Strauss, R. (1986). A comparative application of data envelopment analysis and translog methods - an illustrative study of hospital production. *Management Science 32(1)*, 30-44.

Barrow, M. (1988). Measuring the performance of local educaion authorities: a frontier approach. *University of Sussex Economic Discussion Paper No. 88/1*, 1-35.

Bautista-Arredondo, S., Sosa-Rubi, S., Opuni, M., Kwan, A., Chaumont, C., Coetzee, J., & Condo, J. (2014). Assessing cost and technical efficiency of HIV prevention interventions in sub-Saharan Africa: the ORPHEA study design and methods. *BMC Health Service Research*, 599-610.

Beck, E., Santas, X., & DeLay, P. (2008). Why and how to monitor the cost and evaluate the cost-effectiveness of HIV services in countries. *AIDS 2008*, S75-S85.

Blecher, M. S., Chlu, C., Abdullah, F., Daven, J., Tavanxi, N., Meyer-Rath, G., . . . Borowitz, M. (2016, June 4). HIV and AIDS financing in South Africa: sustainability and fiscal space. *South African Health Report*, pp. 203-220.

Blecher, M., Chiu, C., Pillay, Y., & al, e. (2016, May 6). HIVand AIDS Financing in South Africa: sustainability and fiscal space. *South African Health Review*, pp. 203-219.

Brent, R. (2010). A social cost-benefit criterion for evaluating Voluntary Counseling and Testing with an application to Tanzania. *Health Economics*, 154-72.

Briggs, A. (2001). The Death of Cost-minimisation Analysis? *Health Economics* , 179-184.

Buttigieg, S. (2012). Benchmarking and Performance Management in Healthcare. *European Health Management Association* (pp. 1-48). London: Health Catalyst.

Charnes, A., Cooper, W., & Lewin, A. (1997). *Data Envelopment Analysis: Theory, Methodology and Applications.* New York: Kluwer Academic Publishers.

Charnes, A., Cooper, W., & Seiford, L. (1978). Measuring the efficiency of decision making units. *European Journal of operational Research*, 429-444.

Chilingerian, J., & Sherman, H. (2011). Health-care Applications: from hospitals to physicians, from productive efficiency to quality frontiers. *Handbook of Data Envelopment Analysis*, 445-493.

Cleary, S., McIntyre, D., & Boulle, A. (2008). Assessing effciency and costs of scalling up HIV treatment. *AIDS 2008*, S35-42.

Coelli, T., Rao, D., O'Donnell, C., & Battese, G. (2005). *An Introduction to Efficiency and Productivity Analysis.* New York: Springer.

Coelli, T., Rao, D., O'Donnell, D., & Battese, C. (1998). *An Introduction to Efficiency and Productivity Analysis.* London: Kulwer Academic Publishers.

Cohen, M., Chen, Y., McCauley, M., Gamble, T., & Hosseinipour, M. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. *New England Journal of Medicine*, 493-505.

Cohen, S., & Guthrie, T. (2013). *Financing the South African National Strategic Plan for HIV, STIs and TB 2012 - 2016.* Pretoria: SANAC.

Cooper, W., Seiford, L. M., & Tone, K. (2006). *Introduction to Data Envelopment Analysis and Its Uses.* London : Kluwer.

Copp, J. H. (1976). Diversity of Rural Society and Health Needs. In E. Hassinger, & L. Whiting, *Rural Health Services.* (pp. 341-352.). Iowa: Iowa State University Press.

Cylus, J., Papanicolas, I., & Smith, P. (2017). *How to make sense of health system efficiency comparisons?* Marmorvej: WHO.

Debreu, G. (1951). The coefficient of resource utilization. *Journal of the Econometric Society*, 273-293.

Department of Health of the Republic of South Africa (DOH) and South African National AIDS Council (NAC). (2019, May 08). *South African HIV and TB Investment Case - Summary Report 1.* Retrieved from South African National AIDS Council: http://sanac.org.za/wp-content/uploads/2016/03/1603-Investment-Case-Report-LowRes-18-Mar.pdf.

Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency in post offices. In M. Marchand, P. Pestieau, & H. Tulkens, *Performance of public enterprises: concepts and measurements* (pp. 108-123). Amsterdam: North-Holland.

Doyle, P., Broomberg, J., Steinberg, M., Masobe, P., & Behr, G. (1991). *AIDS in South Africa: the demographic and economic implications Paper No. 23.* Johannesburg: Centre for Health: University of Witwatersrand.

Drummond, M., Weatherly, H., & Ferguson, B. (2008). Economic evaluation of health interventions. *British Medical Journal*, a1204.

Dyson, R., Allen, R., Camanho, A., Podinovski, V., Sarrico, C., & Shale, E. (2001). Pitfalls and protocols in DEA. *European Journal of Operational Research*, 245–259.

Ettorchi-Tardy, A., Levif, M., & Phillipe, M. (2012). BEnchmarking: A Method for Continous Quality Improvement in Health. *Healthcare Policy*, e101-e119.

Fare, R., Grosskopf, S., & Lovell, C. (1985). *The Measurement of Efficiency of Produclion.* Boston: Kluwer-Nijhoff.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of the Royal Statistical Society*, 253-281.

Forsythe, S. (2006, November 12). Guidelines for Assessing the Economic and Financial Costs of HIV/AIDS Prevention and Care Programs. *Evaluating Programs for HIV/AIDS for HIV/AIDS Prevention and Care in Developing Countries: A Handbook for Program Managers and Decision Makers*, pp. 526-548.

Fraser, N., Kerr, C., Harouna, Z., Alhousseini, Z., Cheikh, N., & Gray, R. (2015). Reorientating the HIV response in Niger towards sex work intervention: from better evidence to targeted and expanded practice. *Journal of Acquired Immunodeficienct Syndrom*, S213-220.

Ganley, A., & Cubbin, J. (1992). *Public Sector Efficiency Measurement: Applications of Data Envelopment Analysis.* Amsterdam: Elsevier.

Gilliam, B., Fiscus, S., Marcus, C., & Zhou, S. (1997). Effects of reverse transcriptase inhibitor therapy on HIV-1 viral burden in semen. *Journal on Acquired Immune Dificiency Syndrom*, 54-60.

Global Fund Advocates Network. (2018). *Get Back on Track.* New York: Global Fund.

Gonzalez Lopez-Valcarcel, B., & Barber Perez, P. (1996). Changes in the efficiency of Spanish public hospitals after the introduction of program-contracts. *Investicaionnes Económicas 20 (3)*, 377-402.

Goodacre, S., & McCabe, C. (2002). An introduction to economic evaluation. *Emergency Medical Journal*, 198–201.

Granich, R., Gilks, C., Dye, C., De Cock, K., & Williams, B. (2009). Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. *Lancet*, 48-57.

Grosskopf, S., & Valdmanis, S. (1987). Measuring hospital performance: A non-parametric approach. *Journal of Health Economics 6 (1)*, 89-107.

Hadad, S., Hadad, Y., & Simon-Tuval, T. (2012). Determinants of healthcare system's efficiency in OECD countries. *European Journal of Health Economics*, 253–265.

Health System Trust. (2016). *South African Health Review 2016.* Durban: Health System Trust.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science,*, 203–218.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, 1107-1128.

Hollingsworth, B. (2012). Revolution, evolution, or status quo? Guidelines for efficiency measurement in health care. *Journal of Productivity Analysis*, 1-5.

Hollingsworth, B., Dawson, P., & Maniadakis, N. (1999). Efficiency measurement of health care: A review of non-parametric methods and applications. *Health Care Management*, 161-172.

Holtgrave, D. R., Qualls, N. L., Curran, J. H., Valdiserri, R. O., Guinan, M. E., & Parra, W. C. (1995). AN overview of the effectiveness and efficiency of HIV Prevention Programs. *Public Health Reports*, 134-146.

Hontelez, J. C., de Vlas, S. J., Tanser, F., Bakker, R., Barninghausen, T., Newell, M.-L., . . . Lurie, M. N. (2011). The ipact of the New WHO Antiretroviral Treatment Guidelines on HIV Epidermin Dynamics and Cost in South Africa. *PLoS ONE*, 1-9.

Hontelez, J., de Vlas, S., Tanser, F., Bakker, R., & Barnighausen, T. (2009, January 19). *The impact of the new WHO antiretroviral treatment guidelines on HIV epidemic dynamics and costs in South Africa.* Retrieved from PLoS ONE 6: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021919

HSRC. (2012). *South African National HIV Prevalence, Incidence and Bahaviour Survey.* Cape Town: HSRC.

HSRC. (2017). *South African National HIV Prevalence, Incidence, Behaviour and Communication Survey.* Pretoria: Human Science Research Council.

Jacobs, R. (2001). Alternative methods to examine hospital efficiency: data envelopment analysis and stochastic frontier analysis. *Health Care Management Science*, 103-115.

Johnson, L. (2012). Access to antiretroviral treatment in South Africa 2004 - 2011. *Southern African Journal of HIV Medicine*, 22-27.

Johnson, L. F., Dorrington, R. E., & Moolla, H. (2017). HIV epidemic drivers in South Africa: A model-based evaluation of factors accounting for inter-provincial differences in HIV prevalence and incidence trends . *Southern African Journal of HIV Medicine*, 695-713.

Joumard, I., André, C., & Nicq, C. (2010). *Health care systems: efficiency and institutions.* Kampala: University of Uganda.

Khezrimotlagh, D. (2019, July 2). How to Detect Outliers in Data Envelopment Analysis by Kourosh and Arash Method. Kuala Lumpur, Malaysia.

Kilmann, R. H. (1985). Corporate culture: managing the intangible style of corporate life may be the key to avoiding stagnation. *Psychology Today*, 62-68.

Koopmans, T. C. (1951). An analysis of production as an efficient combination of activities. In T. C. Koopmans, *Activity Analysis of Production and Allocation* (pp. 35-66). London: John Wiley and Sons, Incorporated.

Kozak, M. (2002). Destination benchmarking: Concepts, practices and operations. *Annals of tourism research*, 497-519.

Kuznik, A., Lamorde, M., Hermans, S., Castelnuovo, B., Auerbach, B., Semeere, A., . . . Manabe, Y. (2012). Evaluating the cost-effectiveness of combination antiretroviral therapy for the prevention of mother-to-child transmission of HIV in Uganda. *Bulletin of the World Health Organisation* , 595–603.

KwaZulu-Natal Department of Health. (2018). *Grant SOPS-HIVAIDS Conditional Grant.* Pietermaritzburg: Government Printers.

KZN Government. (2017). *Provincial Strategic Plan 2018 - 2022.* Pietermaritzburg: KZN Government.

Lacey, L., Mauskopf, J., Lindrooth, R., Pham, S., Saag, M., & Sawyer, W. (1999). A Prospective Cost Consequence of Adding Lamivudine to Zidovudine-Containing Antiretroviral Treatment Regimen in the U.S. *Pharmaeconomics*, 23-37.

Latruffe, L., Davidora, S., & Balcombe, K. (2008). Application of a double bootstrap to investigation of determinants of echnical efficiency of farms in Central Europe. *Journal of productivity Analysis*, 183-191.

Lovell, C. (1993). Production Frontiers and Productive Efficiency. In H. O. Fried, C. A. Lovell, & S. S. Schmidt, *The Measurement of Productive Efficiency: Techniques and Applications.* (pp. 256-366). Oxford: Oxford University Press.

McIntyre, D., Bloom, G., Doherty, J., & Brijlal, P. (2007). *Health Expenditure and Finance in South Africa* . Durban: HST and World Bank.

Meyer-Rath, G. (2018). *Can we afford the HIV programme.* Johanneburg: University of Witwatersrand.

Meyer-Rath, G., & Over, M. (2012). HIV Treatment as Prevention: Modelling the Cost of Antiretroviral Treatment - State of Art and Future Directions. *PLoS Medicine*, 1-10.

Mikkelsen, E., Hontelez, J. A., Jansen, M. P., Barnighausen, T., Hauck, K., Johansson, K. A., . . . Baltussen, R. M. (2017). Evidence for Scaling up HIV treatment in sub-Saharan Africa: A call for incorparating health systemm constrainst. *PLoS Medicine*, 1-5.

Mirmiran, S., Li, H., & Ilacqua, L. (2008). Health Care Efficiency In Transition Economies: An Application Of Data Envelopment Analysis. *International Business & Economics Research Journal* , 47-56.

Montaner, J., Lima, V., Barrios, R., Yip, B., & Wood, E. (2010). Association highly active antiretroviral therapy coverage, population viral load and yearly new HIV diagnoses in British Columbia, Canada: a population-based study. *Lancet*, 532-539.

Morita, H. (2003). Analysis of economies of scope by data envelopment analysis: Comparison of efficient frontiers. *International Transmission Operation Res*, 393-402.

Mortimer, D., & Peacock, S. (2002). *Hospital Efficiency Measurement: Simple Ratios vs Frontier Methods.* Victoria: Centre for Health Program Evaluation.

Murillo-Zamorano, L. R. (2001). Economic efficiency and frontier techniques. *Journal of Economic Surveys*, 33-77.

National Department of Health and SANAC. (2016). *SOUTH AFRICAN HIV AND TB INVESTMENT CASE.* Pretoria: Department of Health, South Africa, and South African National AIDS Council.

National Department of Treasury. (2019). *Estimates of National Expenditure.* Pretoria: government Printers.

National Treasury. (2019). *Estimates of National Expenditure.* Pretoria: National Treasury.

Nayar, P., Ozcan, Y., & Nguyen, A. (2013). Benchmarking urban acute care hospitals. *Health Care Management Review*, 137-145.

Nedelea, I., Fannin, J., & Barnes, J. (2010). Analysing dfferences in rural hospital efficiency: a data envelopment analysis approach. *Agriculture and Applied Economics Association* (pp. 1-45). Washington DC: Agriculture and Applied Economics Association.

Nguyen, K., & Coelli, T. (2009). Quantifying the effects of modelling choices on hospital efficiency measures: A meta-regression analysis. *Working Paper Series No WP07/2009*, 1-56.

ÖNDER, A. Ö., DELIKTAŞ, E., & KARADAĞ, M. (2013). THE COMPARASION OF DEA AND SFA METHODS IN THE EFFICIENCY OF THE TURKISH MANUFACTURING INDUSTRY . *Dokuz Eylül Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi*, 71-92.

Ozcan, Y. (2008). *Health Care Benchmarking and Performance Evaluation: An Assessment using Data Envelopment Analysis.* New York: Springer.

Parkin, D., & Hollingsworth, B. (1997). Measuring production efficiency of acute hospitals in Scotland, 1991-94: Validity issues in data envelopment analysis. *Applied Economics 29*, 1425-1433.

Peacock, S., Chan, C., Mangolini, M., & Johansen, D. (2001). *Techniques for Measuring Efficiency in Health Settings.* Belconnen: Productivity Commission.

PEPFAR. (2018). *South Afric Country Operational Plan (COP) 2018.* Pretoria: PEPFAR SA.

Q, A. K., Abdool Karim, S., Singh, B., Short, R., & Ngxongo, S. (1992). Seroprevalence of HIV infection in rural South Africa. *AIDS*, 1535-2539.

Retzlaff-Roberts, D., Chang, C., & Rubin, R. (2004). Technical efficiency in the use of health care resources: A comparison of OECD countries. *Health Policy*, 55–72.

Rogers, M. (1998). The defination and measurement of productivity. *Melbourne Insitutte of Applied Economics and Social Research*, 1-32.

SANAC. (2012, May 9). *National Strategic Plan on HIV, STIs and TB 2012 -2016.* Retrieved from South African National AIDS Concil: https://sanac.org.za/wp-content/uploads/2015/11/4-National-Strategic-Plan-for-HIV-STIs-and-TB-2012-2016.pdf

SANAC. (2017). *National Strategic Plan 2017-2022.* Pretoria: SANAC.

Schimdt, P. (1986). Frontier production functions. *Econometric Reviews*, 289-328.

Silkman, R. (1986). *Measuring Efficiency: An Assessment of Data Envelopment Analysis* . San Francisco: Jossey Bass.

Simar, L., & Wilson, P. (2000). general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, 779–802.

Simar, L., & Wilson, P. (2007). Estimation and inference in two stage, semi parametric models of production processes. *Journal of Econometrics*, 31-64.

Simar, L., & Wilson, P. (2008). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, 779-802.

South African Department of Health. (2010). *National HIV and syphillis prevalence survey in South Africa.* Pretoria: Department of Health.

South African Department of Health. (2014, May 12). *The 2012 National Antenatal Sentinel HIV and Herpes Simplex type-2 prevalence Survey.* Retrieved from National Department of Health: http://www.nicd.ac.za/wp-content/uploads/2017/03/3Final-25-April-2017_Revised-NAS_v5_NICD.pdf

South African Health Review. (2016). *HIV and AIDS financing in South Africa : sustainability and fiscal space.* Durban: Health Systems Trust.

StatsSA. (2016). *Community Survey.* Pretoria: Statistics South Africa.

StatsSA. (2017). *Mid-Year Population Estimates.* Pretoria: Statistics South Africa.

Sweeney, S., Obure, C. D., Maier, C. B., Greener, R., Dehne, K., & Vassall, A. (2012). Costs and efficiency of integrating HIV/AIDS services with other health services: a systematic review of evidence and experience. *Sex Tranformission Infection*, 85-99.

U.S. Agency for International Dvelopment. (2016). *HEALTH FINANCING PROFILE: South Africa.* Wasington DC: U.S. Agency for International Development.

Valdmanis, V. (1990). Ownership and technical efficiency of hospitals. *Medical Care*, 552-561.

Valdmanis, V. (1992). Sensitivity analysis for DEA models. An empirical example using public vs. NFP hospitals. *Journal of Public Economics 48 (2)*, 185-205.

Wagner, B., & Blower, S. (2010). Costs of eliminating HIV in South Africa have been underestimated. *Lancet*, 953.

Wagstaff, A. (1988). Estimating efficiency in the hospital sector: a comparison of three statistical frontier models. *Applied Economics*, forthcoming.

Wagstaff, A. (1992). Quantitative techniques for investigating health service. *Seminar on Economia y Salud del Sistema Sanitario Espanol.* Santander.

Wagstaff, A., & Wang, L. (2011, August 3). A hybrid approach to efficiency measurement with empirical illustrations from education and health. *World Bank: Policy in Research Working Paper*, pp. 1-39.

Wibowo, D., & Tisdell, C. (1992). Health Production Functions For Preventive Health Programs. *Fourteenth Australian Conference of Health Economists* (pp. 106-133). Melbourne: Faculty of Economics, Commerce and Management Monash University and National Centre for Health Program Evaluation Fairfield Hospital.

Williams, A. (1992). Cost effectiveness analysis, is it ethical? *The Journal of Medical Ethics*, 7-11.