UNIVERSITY OF CAPE TOWN

Measuring hospital efficiency using DEA:

An investigation into the relationship between scale and efficiency within the South African private hospital environment

Andrew Linden

LNDAND006

Dissertation submitted in partial fulfilment of the requirements for the degree of Master of Business Science in Actuarial Science

27 April 2013

Abstract

This paper investigates the relationship between scale and efficiency through the application of Data Envelopment Analysis (DEA) to a set of South African private hospitals over the three year period from 2007 to 2009. As part of the investigation, this paper provides a description of the current research into scale and efficiency with a focus on definition and measurement. It also provides an introduction to DEA as a tool for measuring the relationship between hospital scale and efficiency.

Based on the underlying set of private hospitals, this investigation found that scale efficiency improvements are likely to be possible. On average, these improvements could have produced input savings of 6.9% in 2007, 6.8% in 2008, and 6.2% in 2009. Most hospitals were found to operate under non-increasing returns to scale; with hospitals being more likely to operate under decreasing returns to scale than increasing returns to scale. This, together with relatively low occupancy rates, reinforces the general criticism that excess capacity exists within the South African private hospital industry. However, excess capacity may be appropriate given the operational goals, nature of ownership, and role of private hospitals within the South African healthcare system. There was also evidence that smaller hospitals, when measured in terms of number of beds, are more likely to operate with higher scale efficiency.

DEA model specification was found to have a significant impact on the results of the investigation. In particular, hospital scale efficiency and return to scale classification were significantly impacted by the selection of different combinations of input and output variables. Additionally, this paper demonstrates how oversimplified approaches to scale analysis can lead to incorrect conclusions.

Plagiarism declaration

- I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is my own.
- I have used the Harvard referencing guide for citation and referencing. Each significant contribution to this dissertation from the work of other people has been cited and referenced.
- This dissertation is my own work.
- I have not allowed, and will not allow, anyone to copy my work.

Table of contents

Table of abbreviations and acronyms

AE	Allocative Efficiency
BCC model	Banker, Charnes & Cooper model
BDRG	Basic Diagnostic Related Group
CCR model	Charnes, Cooper & Rhodes model
CE	Cost Efficiency
CEO	Chief Executive Officer
CRS	Constant Returns to Scale
DEA	Data Envelopment Analysis
DRG	Diagnostic Related Group
DRS	Decreasing Returns to Scale
ERS	Efficiency Reference Set
FTE	Full Time Equivalent
HIV/AIDS	Human Immunodeficiency Virus / Acquired Immunodeficiency Syndrome
ICU	Intensive Care Unit
IRS	Increasing Returns to Scale
MPSS	Most Productive Scale Size
NDRS	Non-Decreasing Returns to Scale
NHI	National Health Insurance
NIRS	Non-Increasing Returns to Scale
OECD	Organization for Economic Cooperation and Development
RTS	Returns To Scale
SE	Scale Efficiency
SFA	Stochastic Frontier Analysis
TE	Technical Efficiency
VRS	Variable Returns to Scale
WHO	World Health Organisation

1. Introduction

1.1. Context of the investigation

Healthcare consumes a vast quantity of South Africa's resources, yet South Africa's health outcomes remain poor when compared to similar middle income countries (Department of Health, 2011a). The provision of healthcare has a significant financial and social impact on South Africa, and this warrants research into improving the national healthcare system. Furthermore, the Department of Health (2011b) has indicated that reducing the high costs of private healthcare is critical to the successful implementation of National Health Insurance (NHI). This means that the measurement, demonstration, and improvement of efficiency in the private sector will become increasingly important in the future South African healthcare environment. Improving scale efficiency is one approach that could assist with reducing costs in the private sector. However, a better understanding of the relationship between scale and efficiency is needed in order to determine whether scale inefficiencies exist, and whether they are able to be addressed.

1.2. Research objectives

The high level aim of this paper is to investigate the relationship between scale and efficiency within the South African private hospital environment.

In more detail, the objectives of this paper are to:

1. Describe the current research into scale and efficiency, focusing on definition and measurement.
2. Provide an introduction to Data Envelopment Analysis (DEA), highlighting its use as a tool for measuring the relationship between hospital scale and efficiency.
3. Investigate the relationship between scale and efficiency through the application of DEA to a set of South African private hospitals.

1.3. Overview of the investigation

This paper consists of 7 chapters. The current chapter forms the introduction to the investigation. Chapter 2 provides the background and context of the investigation. This chapter focuses on providing an overview of the South African healthcare environment, the financing of the healthcare system, the introduction of the NHI, and the private hospital industry. In order to investigate the relationship between scale and efficiency, clear definitions of the concepts of scale and efficiency are required. These definitions are provided in chapter 3. The various methods available to measure efficiency are also discussed in chapter 3. One such method, Data Envelopment Analysis (DEA), is expanded upon

in chapter 4. This includes a discussion of DEA model specification, variables and limitations. Additionally, chapter 4 describes the practical applications of DEA in the hospital industry, with particular focus on the major studies of scale efficiency within the South African hospital industry. Chapter 5 describes the data and methodology used in this paper to investigate the relationship between scale and efficiency within the South African private hospital environment. This is followed, in chapter 6, by a discussion of the results of the investigation. Chapter 7 then presents the conclusions that can be drawn from this investigation, followed by recommendations for further research.

1.4. Scope and limitations

This paper investigates the relationship between scale and efficiency by analysing data from a set of South African private hospitals for the three year period from 2007 to 2009. This dataset was sourced from a single provider, which may be materially different from the other providers in the industry. Therefore the results and conclusions drawn from this investigation may not be representative of the South African private hospital industry as a whole.

Additionally, this investigation cannot be used to draw conclusions regarding public sector hospitals, as these hospitals differ significantly from private hospitals in terms of operating constraints and objectives. However, given an appropriate dataset, it is envisioned that the methodology adopted in this paper could be applied to the public sector.

While this investigation attempts to identify and discuss scale inefficiencies, it does not attempt to examine the feasibility and practical difficulties of implementing scale improvements.

2. Background to the investigation

2.1. The South African healthcare environment

The Constitution of the Republic of South Africa (1996) grants every South African the right to have access to healthcare services. The South African Department of Health is tasked with ensuring that this right is met. However, this must be done within a socio-economic environment that is plagued by poverty and inequality.

The Department of Health (2011c) states that it strives to provide a long and healthy life for all South Africans. This organisation aims to improve the general health of the South African population through the prevention of illnesses and the promotion of healthy lifestyles, and to consistently improve the healthcare delivery system by focusing on access, equity, efficiency, quality and sustainability. In order to meet these aims the Department of Health is proposing a radical overhaul of the current South African healthcare system, in the form of National Health Insurance (NHI). It is envisioned that NHI will provide universal coverage for all South Africans.

The current South African healthcare system is characterised as a two-tier system, split between the public and private sectors. The public sector is criticised for providing poor health outcomes due to bad management and design, while the private sector serves its customers well (Centre for Development and Enterprise, 2011). However, the private sector services a minority of the population who are able to afford access, while the majority of the population depends on the public sector. This has resulted in South Africa's healthcare system being fragmented along socio-economic lines (McIntyre, Thiede, Nkosi, Mutyambizi, Castillo-Riquelme, Gilson, Erasmus & Goudge, 2007). The private healthcare system is mainly funded by medical schemes[1], various hospital care plans and out-of pocket-payments. The public sector is mainly funded by the state through the fiscus.

The imbalances between the availability of resources in the private and public sectors are well documented (McIntyre *et al*, 2007; McIntyre, 2010; Department of Health, 2011a). Of South Africa's total healthcare expenditure, 44% is attributable to medical schemes (McIntyre, 2010). The majority of this expenditure goes towards private for-profit providers, particularly private hospitals, specialists and pharmacies. However, only 15% of the population are members of medical schemes and benefit from this expenditure (McIntyre, 2010). The private sector has a disproportionate share of financial

[1] Medical schemes are South African specific, not-for-profit schemes that provide healthcare insurance to members in return for private contributions. Medical schemes are regulated by the Medical Schemes Act (1998) and are characterised by the principles of open enrolment, community-rating and prescribed minimum benefits. The reader is directed to Ramjee & McLeod (2007) for further details regarding medical schemes.

and human resources given the size of the population that it serves. On the other hand, the public sector is under-resourced given the size of the population that it serves and its burden of disease (Department of Health, 2011a).

Coovadia, Jewkes, Barron, Sanders & McIntyre (2009) describes the burden of disease in South Africa as being quadruple (divided into four clear health problems). The first health problem is the HIV/AIDS pandemic. South Africa has 17% of the HIV infected people in the world while only having 0.7% of the world's population. HIV/AIDS and tuberculosis co-infection is also common, and South Africa has one of the highest tuberculosis infection rates in the world. The second problem is maternal, infant and child mortality. This is driven by poor access to healthcare, as well as the HIV/AIDS pandemic. The third health problem is non-communicable diseases. These include high blood pressure, diabetes, heart disease, lung diseases, cancer and mental illness. These diseases have four main risk factors: alcohol, smoking, poor diet, and lack of exercise; all of which are widespread in South Africa. The fourth problem is injury and violence. A significant portion of injury is due to road accidents and inter-personal violence, particularly violence against women and children. These health problems are exacerbated by widespread poverty and unemployment, and a lack of basic infrastructure that is necessary for a healthy life (Centre for Development and Enterprise, 2011).

2.2. Healthcare financing in South Africa

Total government spending on public health services has increased strongly from R63 billion in 2007 to R121 billion projected for 2012 (Gordhan, 2011; Gordhan, 2012). Public healthcare spending is approximately 11% of projected total government spending for 2012.

According to the World Health Organisation (2012), South Africa's combined private and public sector healthcare expenditure in 2010 was 8.9% of gross domestic product. Of this expenditure 55.9% was spent in the private sector, and 44.1% was spent in the public sector.

Schieber, Baeza, Kress & Maier (2006) estimated that high income countries spend on average 7.7% of their gross domestic product on healthcare, while middle income countries spend 5.8%, and low income countries spend 4.7%. Even though South Africa spent more on healthcare in 2010, than any of the above averages, the provision of healthcare for the South African population remains poor when compared to similar middle income countries (Department of Health, 2011a).

2.3. The introduction of National Health Insurance

In 2005, the member states of the World Health Organisation (WHO) adopted a resolution which encouraged its members to develop health financing systems aimed at providing universal coverage (Carrin, Mathauer, Ke Xu & Evans, 2005). Norman & Weber (2009) define universal coverage, as

envisioned by the WHO, as providing access to key promotive, preventive, curative and rehabilitative health interventions for all at affordable cost, thereby achieving equity in access. They elaborate on this definition by explaining that universal coverage means that every citizen must be provided with access to necessary care and essential service. For the citizen, universal coverage should ideally guarantee the provision of quality healthcare, together with protection against the financial risks associated with healthcare costs.

The Department of Health (2011a) has gazetted a policy dealing with the proposed National Health Insurance (NHI) system. This policy claims that the NHI model of healthcare delivery is based on the WHO's universal coverage model. The Department of Health (2011a) states that the NHI will ensure that all South African citizens have access to appropriate, affordable, efficient and quality healthcare services. This access will be granted regardless of employment status or ability to make contributions to the NHI fund. Implementation of the NHI will require a major overhaul of the current healthcare system, and is expected to be phased in over a period of 14 years. However, the private sector and other stakeholders have expressed concerns over the proposed timelines for the implementation of NHI and universal coverage (Ramjee & McLeod, 2010).

In a press release, the Department of Health (2011b) stated that it is intended that the NHI will build on the strengths of both the public and private sectors. The Department of Health cited the quality of healthcare in the private sector as a strength, but criticised the costs associated with delivering these services. The high cost of healthcare in the private sector was claimed to distort prices across the entire healthcare market, impacting the public sector. This claim is supported by McIntyre (2010), who has shown that per capita spending (in real terms) in the public sector has been relatively constant, while increasing rapidly in the private sector. The Department of Health (2011b) identified two goals that must be achieved in order for the NHI to succeed. Firstly, the quality of service in the public sector must be improved and, secondly, the high costs of private healthcare must be addressed. In order to contain costs and ensure financial sustainability, it is likely that payment mechanisms to healthcare providers will change under NHI (Econex, 2010a). Efficient resource use, through the minimisation of inputs and the maximisation of outputs, will thus be even more important in an NHI driven healthcare market.

The Department of Health (2011a) has stated that the first five years of NHI will focus on testing its implementation through pilot projects, as well as strengthening the healthcare system. Particular attention will be devoted to improving the management of healthcare facilities and the quality of their services. Human resources planning, as well as investment in infrastructure, equipment, systems and data management are all necessary for the successful implementation of NHI.

Various steps and commitments to the implementation of NHI have already taken place. The National Treasury of South Africa (Gordhan, 2011) noted that the phasing in of NHI would require substantial reforms to address imbalances across the public and private sectors. Gordhan, in the 2011 budget, allocated R8 billion for laying the foundations of NHI. The Department of Health (2011d) has gazetted a new policy regarding the management of public hospitals, which will facilitate the implementation of NHI. This new policy aims to improve the management, governance and functionality of hospitals through explicitly defining responsibilities and accountabilities of management, and through the effective recruitment of hospital CEOs and board members. Human Resources for Health South Africa (2011) has begun planning to meet the resource demands of the NHI, with their publication of a human resources strategy. This strategy extends to 2017 and focuses on increasing the training of healthcare professionals, while recognising that it takes many years to adequately train these professionals.

Ramjee & McLeod (2010) found that the private sector – represented by stakeholder groups from the hospital industry, pharmaceutical industry, the medical scheme industry, and the actuarial profession – expressed a commitment to the goal of achieving universal access to quality healthcare for all South Africans. The private sector also recognised that the current healthcare system is unsustainable, and that reform is necessary to improve the system (Ramjee & McLeod, 2010).

2.4. The South African private hospital industry

The South African private hospital industry is highly concentrated, with three main players dominating the industry, namely Netcare, Life Healthcare and Mediclinic. These organisations are large, for-profit, public companies that are listed on the Johannesburg Stock Exchange. Together they account for 77% of private hospital beds in the industry (McIntyre, 2010). Netcare accounts for 31% of private hospital beds, Life Healthcare accounts for 25%, and Mediclinic accounts for 21%. The remaining 23% of private hospital beds are accounted for by independent hospitals, which are often owned by groups of doctors (McIntyre, 2010). This high level of industry concentration is due to considerable industry consolidation which took place in the 1990s and continued until 2006 (Matsebula & Willie, 2007). Geographically, private hospitals are concentrated in major urban areas with the largest number of hospitals located in Gauteng, KwaZulu-Natal and the Western Cape (Council for Medical Schemes, 2011).

The average private hospital has less than 200 beds, and most patients are admitted for less than 30 days (Matsebula & Willie, 2007). When compared to the public sector, private hospitals control a considerable proportion of South Africa's healthcare resources. Private hospitals account for 21.0% of total hospital beds (Matsebula & Willie, 2007), 38.1% of the total number of general practitioners and 55.6% of the total number of medical specialists in South Africa (Econex, 2010b). However, the

Centre for Development and Enterprise (2011) has found evidence that estimates of the resources available in private hospitals are often overstated. For example, estimates of private hospital resource use that were provided in the Development Bank of Southern Africa's Health Roadmap (2008) were found to be significantly overstated. The Centre for Development and Enterprise (2011) also points out that the use of these resources in the private sector greatly reduces the burden on the public sector.

There is criticism that excess capacity exists within the private hospital industry (African National Congress, 2010). The average bed occupancy rate in South African private hospitals is around 65%; while the average occupancy rate in the 30 OECD countries in 2005 was around 75% (African National Congress, 2010; Centre for Development and Enterprise, 2011). International best practice recommends that occupancy rates should not exceed 85% as this compromises infection control and the ability to cope with emergencies (Keegan, 2008). Reducing excess capacity would increase hospital efficiency and allow a significantly larger number of health outcomes to be produced given the current level of inputs.

A serious problem facing private hospitals is rapid cost escalation (McIntyre, 2010). The Centre for Development and Enterprise (2011) has found that the cost of private hospital healthcare has increased 12 times in real terms over the last 30 years. One potential driver of this is private industry concentration and market power (Centre for Development and Enterprise, 2011). Mergers and acquisitions by all of the three large firms have been taken to the Competition Tribunal of South Africa. For example, cases involving Netcare, Mediclinic and Life Healthcare were heard in 2006, 2009 and 2010 respectively (Competition Tribunal of South Africa). Matsebula and Willie (2007) describe how the private sector's cost structures and pricing practices may be the cause of cost escalation. In particular, the fee-for-service method of payment adopted by medical schemes creates an incentive for over-servicing. It also reduces the incentive for hospitals to innovate with regard to healthcare delivery (Centre for Development and Enterprise, 2011). However, these cost structures are largely driven by the dynamics of the current healthcare environment and the way in which medical schemes operate. It is possible that this will be addressed under NHI, where private hospitals may need to adopt new methods of payment when contracting with the NHI (Econex, 2010a). Additionally, private hospitals, by law, cannot employ doctors and specialists directly. In order to attract these resources, private hospitals compete with each other by investing in facilities and equipment in excess of their needs (Centre for Development and Enterprise, 2011). This cost escalation is largely driven by the legislation governing the current healthcare environment, which could also be addressed under NHI (Centre for Development and Enterprise, 2011). The shortage of healthcare professionals in South Africa also contributes to rising costs. Other explanations of rising costs that are not under the control of hospital management include the general depreciation of the

Rand, the increasing number of inpatient days, changes in case-mix, and the introduction of community rating by medical schemes (Matsebula & Willie, 2007; Medical Schemes Act, 1998).

The private hospital industry consumes a large proportion of South Africa's healthcare resources. It is therefore important to measure and understand the productivity of these resources. Additionally, any attempt to address excess capacity in the private hospital industry requires a better understanding of the relationship between scale and efficiency within the context of the South African healthcare system. This warrants further investigation.

3. Defining and measuring efficiency

3.1. Introducing efficiency

Defining and measuring efficiency is a long standing problem in economics. In the context of the firm, the term *efficiency* refers to the best use of resources in production (Hollingsworth *et al*, 1999). Nguyen & Coelli (2009) provide a more general definition of efficiency; they interpret efficiency as the extent to which objectives are achieved in relation to the economic resources used.

One may be tempted to think of efficiency as having a much narrower definition. For example, it would be reasonable to think of efficiency as the ability to produce the maximum possible outputs given a set of inputs; or alternatively as the ability to produce a set of outputs using the minimum possible level of inputs. This way of thinking about efficiency is not incorrect; however it is only able to capture one of many types of efficiency, namely technical efficiency.

In fact, the literature identifies various types of efficiency. Separate analysis of the different types of efficiency can provide further insight into the economics of efficiency and productivity. From the works of Sherman & Zhu (2006) and Coelli, Rao, O'Donnell & Battese (2005), five types of efficiency are identified, namely:

- Price efficiency,
- Technical efficiency,
- Allocative efficiency,
- Cost efficiency, and
- Scale efficiency.

Note that the focus of this investigation will be on scale efficiency and, to a lesser extent, technical efficiency. The concept of productivity is also widely used in the literature. Productivity and the five types of efficiency are described in the sections that follow.

3.2. Productivity

Productivity is generally defined as the ratio of units of outputs to units of inputs (Coelli *et al*, 2005; Sherman & Zhu, 2006). Note that inputs used to calculate the productivity ratio refer to inputs arising from all factors of production. Hollingsworth *et al* (1999) explain that calculating the productivity ratio is simple in the case of a single-input, single-output firm. For a single-input, single-output firm productivity can be defined as:

$$Productivity = \frac{y}{x} \tag{1}$$

where,

y is the number of units of the firm's single output, and

x is the number of units of the firm's single input.

However, for the more realistic case of a multiple-input, multiple-output firm, calculating the productivity ratio is significantly more difficult and less objective. This is because the inputs and outputs cannot be simply summed; they must be aggregated into a single index representing total output and a single index representing total input. This can be achieved by weighting the outputs and inputs before summation. Deciding the values of the weights can be a subjective and difficult process. Therefore, for a multiple-input, multiple-output firm productivity is defined as:

$$Productivity = \frac{\sum_{r=1}^{s} u_r y_r}{\sum_{i=1}^{m} v_i x_i} \tag{2}$$

where,

u_r is the weighting for output r,

y_r is the number of units of output r,

v_r is the weighting for input i,

x_i is the number of units of input i,

s is the total number of outputs, and

m is the total number of inputs.

Note that the productivity ratio concerns itself with units of inputs and outputs, and does not necessarily take prices into account.

All hospitals are multiple-input, multiple-output firms. Inputs would include, *inter alia*, number of doctors, nurses, beds, pharmaceuticals, equipment and facilities. Outputs would include, *inter alia*, number of treated patients, inpatient days, outpatient cases and surgical procedures. When compared with the manufacturing industry, outputs in a service industry, such as the hospital industry, are more difficult to define (Sherman & Zhu, 2006). For example, there are few objective ways of determining the quality of service in a hospital, such as whether a patient requires one more day of hospitalisation. This aspect of the service industry, particularly the hospital industry, introduces additional complications when dealing with matters of efficiency.

A graphical interpretation of productivity is provided in section 3.8.

3.3. Price efficiency

A firm is price efficient when it purchases all its inputs (capital, labour and production materials) at the lowest possible price (Sherman & Zhu, 2006). This means that a firm could increase its efficiency if it can purchase an input at a lower price, without compromising on the quality of that input. However, there are many factors that influence the price efficiency of a firm. In the hospital industry, the degree of competition could influence price efficiency. Price efficiency could also be impacted by the relative bargaining powers of a hospital, its suppliers and the other hospitals in the industry. Furthermore, the public and private sectors may impact industry prices in different ways.

3.4. Technical efficiency

A firm is technically efficient when it produces the maximum possible outputs given a set of inputs. Alternatively, efficiency is the ability to produce a set of outputs using the minimum possible level of inputs. This definition of technical efficiency appears in the works of Farrell (1957), Coelli *et al* (2005) and Sherman & Zhu (2006). Intuitively, a technically efficient firm avoids waste by using resources in the most technologically efficient manner (Nguyen & Coelli, 2009).

Conversely, technical inefficiency will exist when it is possible to produce more outputs with the current levels of inputs (or when it is possible to produce the current levels of outputs with fewer inputs). For example, in the context of a hospital, the implementation of a more efficient staffing management system may reduce resource inputs while maintaining the same level of outputs, thereby reducing technical inefficiency.

Hollingsworth *et al* (1999) and Coelli *et al* (2005) identify a technically efficient firm as operating on the production frontier. This is illustrated graphically in sections 3.8 and 3.9.

Note that a technically efficient firm is not necessarily a firm that maximises its productivity, as it may still be able to improve its productivity, say by exploiting scale economies (Coelli *et al*, 2005). Furthermore, technical efficiency is concerned with production and does not take the prices of inputs and outputs into account.

3.5. Allocative efficiency

A firm is allocatively efficient when, given a set of required outputs and prevailing input prices, the firm adopts the input mix that minimises its production costs (Linna, 1998; Coelli *et al*, 2005). Alternatively, a firm is allocatively efficient when it produces the mix of outputs that maximises its revenue, given a set of inputs and prevailing output prices. Allocative efficiency thus refers to the use of inputs, or production of outputs, in optimal proportions.

Price information must be available in order to calculate allocative efficiency. Furthermore, a behavioural assumption is needed, such as profit maximisation or cost minimisation (Nguyen & Coelli, 2009). Profit maximisation is likely to be the appropriate assumption for private hospitals, particularly private hospitals in the current South African healthcare environment. However, in the future NHI environment, private hospitals may have to accept new payment mechanisms, shifting their behaviour closer to that of cost minimisation (Econex, 2010a). For example, this could occur in the future NHI environment if private hospitals are required to contract on a payment mechanism that fixes their revenue. In this case, profits will be maximised through cost minimisation.

Allocative efficiency naturally leads to the classic economic question of capital versus labour (Sherman & Zhu, 2006). In a hospital context, for example, there may be a trade-off between the number of nursing staff and medical monitoring equipment. The trade-off will depend on the relative prices of the resources, but could also be influenced by the efficacy and quality of the care provided by each of the alternatives.

A graphical interpretation of allocative efficiency is provided in section 3.9.

3.6. Cost efficiency

A firm is cost efficient when it is technically and allocatively efficient, and cost efficiency is defined as the product of technical and allocative efficiency (Coelli *et al*, 2005). Cost efficiency is thus a measure of total efficiency, keeping the firm's scale and external prices constant.

Since price information must be available in order to calculate allocative efficiency, it follows that price information is also needed to calculate cost efficiency. The distinction between technical efficiency, allocative efficiency and cost efficiency is illustrated graphically in section 3.9.

3.7. Scale efficiency

A firm is scale efficient when it operates at a point on the production frontier which maximises its productivity (Coelli *et al*, 2005). Banker (1984) labelled this point, or set of points, as the most productive scale size (MPSS), which represents the optimal scale given the current production technology. A firm may be technically efficient (operating at a point on the production frontier) but may still be able to improve productivity by exploiting scale economies. Scale efficiency is a simple concept that is easy to understand in the single-input, single-output case, but it is more difficult to conceptualise in a multi-input, multi-output situation (Coelli *et al*, 2005). The concepts of scale efficiency and MPSS are illustrated graphically in section 3.8.

Returns to scale can be classified as three distinct types: constant returns to scale (CRS), increasing returns to scale (IRS), and decreasing returns to scale (DRS). Different regions of the production frontier can exhibit different return to scale classifications. This type of production frontier is sometimes referred to as having variable returns to scale (VRS).

A firm is said to exhibit IRS when a proportionate increase in inputs results in a larger than proportionate increase in outputs (Banker & Thrall, 1992). These firms can increase their productivity by increasing the scale of their operations. Similarly, a firm is said to exhibit DRS when a proportionate increase in inputs results in a less than proportionate increase in outputs. These firms can increase their productivity by decreasing the scale of their operations. A firm is said to exhibit CRS when a proportionate increase in inputs results in a proportionate increase in outputs. For these firms production size does not matter (Nguyen & Coelli, 2009). By definition, a firm that is operating at its MPSS must exhibit CRS, else it would not be scale efficient.

In practice, inefficiencies due to scale may exist for many reasons. For example, scale inefficiencies may arise when fixed costs are high or when additional capital investment can only be made in indivisible units. Sherman & Zhu (2006) note that a hospital will tend to need at least one administrator regardless of how small it is, which may increase the relative cost of administering a small hospital compared to a larger hospital. Nguyen & Coelli (2009) identify potential sources of scale inefficiencies in the hospital industry. For example, imperfect competition or government regulations may reduce economic incentives for hospitals to operate at optimal scale efficiency. Additionally, financial constraints or valid social objectives may impose external constraints on the level of scale efficiency that is possible to achieve. Given these complexities, improvements in scale efficiency may be difficult to achieve in the short term and may only be possible to achieve in the long run where all inputs, including capital inputs, are variable. Sherman & Zhu (2006) claim that scale issues can be easily oversimplified or misunderstood, and that hospital scale research often focuses on the number of beds as a way to measure scale size. The results of such analyses tend to suggest that scale effects are small. However, these studies often do not take into account other scale factors such as the optimal number of specialist facilities, equipment, etc.

3.8. Graphical representations of productivity, technical efficiency and scale efficiency

Figure 1, which is adapted from Nguyen & Coelli (2009), is an illustration of a simple single-input (X), single-output (Y) production technology. The production frontier, representing the maximum output attainable from each level of input, reflects VRS in this example. All points between the VRS frontier and the x-axis are input-output combinations that are capable of being produced under the current production technology.

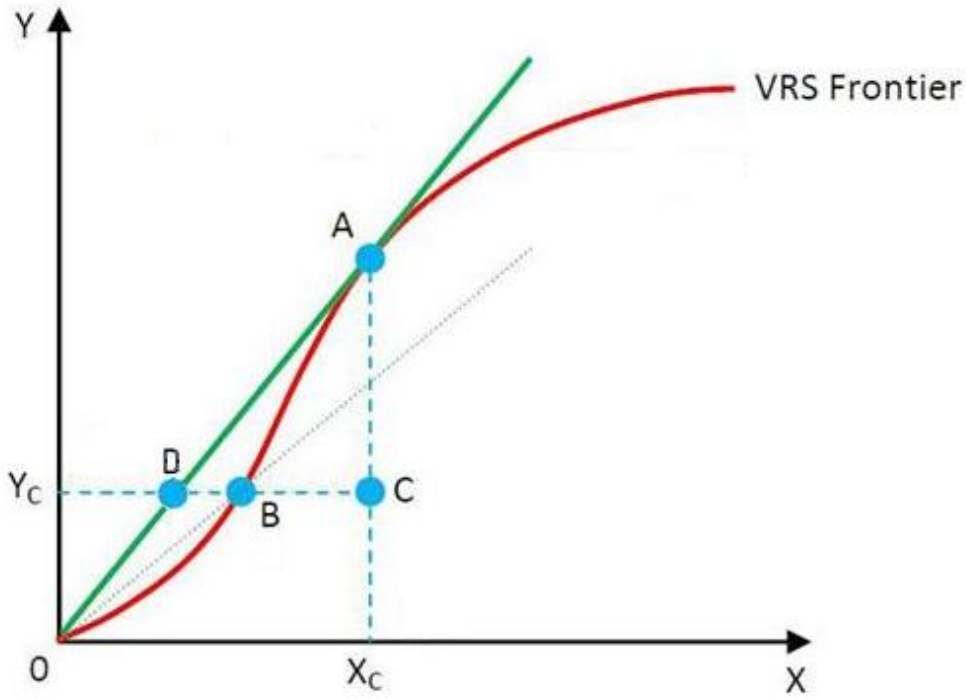

Figure 1: A production frontier illustrating productivity, technical efficiency and scale efficiency
Source: adapted from Nguyen & Coelli, 2009

Firms operating on the production frontier are technically efficient, while those operating below the production frontier are technically inefficient. In Figure 1, firms A and B are technically efficient while firm C is technically inefficient. It is therefore possible for C to produce the same level of output while using fewer inputs (or produce more outputs using the same level of input). The technical inefficiency of firm C can be interpreted graphically as the distance from point B to point C. The technical efficiency of firm C (TE_C) is measured as the ratio of the inputs that would be required to produce the same level of output if the firm operated on the production frontier (i.e. if firm C reduced its input use to point B), to its current level of inputs. Therefore technical efficiency of firm C is defined as:

$$TE_C = \frac{BY_C}{CY_C} \tag{3}$$

where,

$0 \leq TE_C \leq 1$

A technical efficiency (TE) score of one indicates that the firm lies on the production frontier and that no contraction of inputs is feasible i.e. the firm is technically efficient. A firm with a TE score of less than one is inefficient; and the more inefficient the firm, the lower the TE score. The above definition of technical efficiency refers to input-orientated technical efficiency as the inputs have been reduced

in order to move the firm (currently operating at point C) to the production frontier (point B). It is also possible to define technical efficiency in terms of an output-orientated technical efficiency, whereby the firm's outputs are increased in order to move the firm to the production frontier, in this case from point C to point A.

Even though firms A and B are technically efficient, they are not equally productive. Productivity was defined in section 3.2 as the ratio of outputs to inputs. In the simple single-input (X), single-output (Y) case, the productivity ratio is equal to the gradient of the ray from the origin to the point of interest. The gradient of the ray from the origin (and hence productivity) is maximized when the ray is tangential to the production frontier. This occurs at point A. Since the gradient of ray OB is less than the gradient of ray OA, firm B is less productive than firm A.

In section 3.7, a firm was defined as being scale efficient and producing at the most productive scale size (MPSS) when it operates at a point on the production frontier which maximises its productivity. It has been shown that this point is point A. Firms producing less output than point A, such as point B, are operating in the IRS region of the production frontier. These firms could increase their productivity by increasing the scale of their operations towards the MPSS (point A). Similarly, firms producing more output than point A are operating in the DRS region of the production frontier. These firms could increase their productivity by decreasing the scale of their operations towards the MPSS. Banker, Charnes & Cooper (1984) noted that tangents to points on the production frontier that exhibit CRS intercept the y-axis at the origin, while tangents to points on the production frontier that exhibit IRS intercept the y-axis below the origin (i.e. the tangents have negative y-intercepts). Similarly, tangents to points on the production frontier that exhibit DRS intercept the y-axis above the origin (i.e. the tangents have positive y-intercepts). These observations can be generalised to the multiple-input, multiple-output case.

The scale efficiency of a firm can be defined as the ratio of the productivity of that firm to the maximum possible productivity given the current production technology. For firm B, this is the ratio of the gradient of ray OB, to the gradient of the ray OA. Therefore, the scale efficiency of firm B (SE_B) can be defined as:

$$SE_B = \frac{Gradient\ of\ OB}{Gradient\ of\ OA} \tag{4}$$

where,

$0 \leq SE_B \leq 1$

A scale efficiency (SE) score of one indicates that the firm is producing at the MPSS and that no scale inefficiencies exist. Since points A and D lie on the same tangent, the gradient of the ray OA is equal to the gradient of OD. The scale efficiency of firm B can therefore also be defined as:

$$SE_B = \frac{Gradient\ of\ OB}{Gradient\ of\ OD} = \frac{OY_C/BY_C}{OY_C/DY_C} = \frac{DY_C}{BY_C} \tag{5}$$

where,

$0 \leq SE_B \leq 1$

Note that, given the current production technology, production at point D is not possible. However, it is often more convenient to define scale efficiency in this way – as a ratio of inputs. Graphically, this definition of scale inefficiency for firm B can be interpreted as distance from point B to point D.

Alternatively, scale efficiency of firm B can also be defined as the ratio of $TE_{B(CRS)}$ to $TE_{B(VRS)}$; where $TE_{B(CRS)}$ is the technical efficiency score of firm B under the assumption of a CRS production frontier (calculated using ray OA as the production frontier), and $TE_{B(VRS)}$ is the technical efficiency score of firm B, calculated using the actual VRS production frontier. The scale efficiency of firm B can therefore also be defined as:

$$SE_B = \frac{TE_{B(CRS)}}{TE_{B(VRS)}} = \frac{DY_C/BY_C}{BY_C/BY_C} = \frac{DY_C}{BY_C} \tag{6}$$

where,

$0 \leq SE_B, TE_{B(CRS)}, TE_{B(VRS)} \leq 1$

3.9. Graphical representations of technical, allocative and cost efficiencies

Figure 2, which is adapted from Nguyen & Coelli (2009), is an illustration of a simple two-input (X_1 and X_2), one-output production technology. The unit isoquant represents the different combinations of the two inputs that can be used by a fully efficient firm to produce one unit of output (Coelli *et al*, 2005). The isoquant can be interpreted as a cross-section of a two-input production frontier surface at a given level of output.

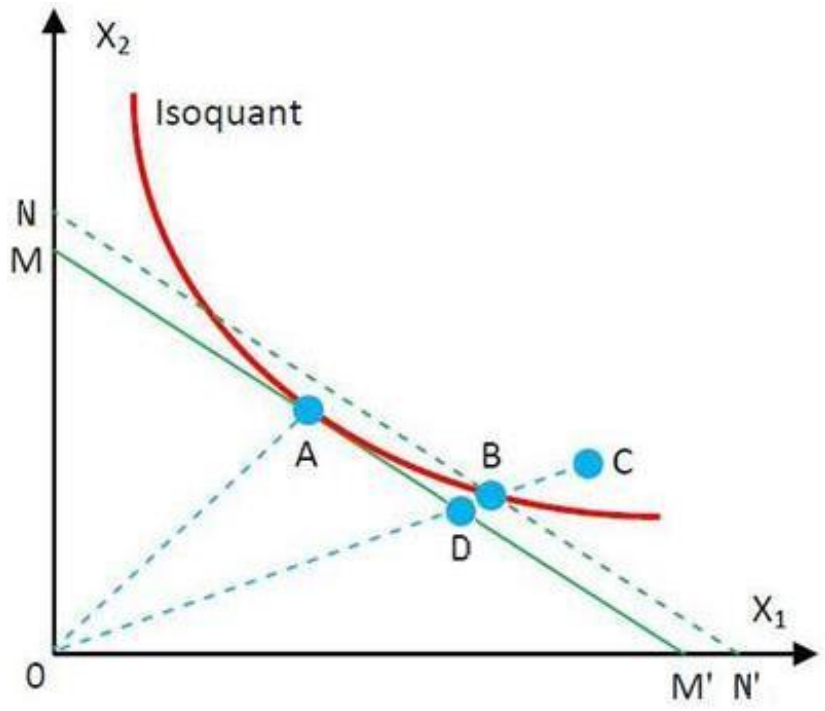

Figure 2: A production frontier illustrating technical, allocative and cost efficiencies
Source: adapted from Nguyen & Coelli, 2009

Firms A and B lie on the production frontier (represented by the isoquant) and are therefore technically efficient. Firm C lies to the right of the isoquant in an area where a greater number of inputs are required to produce the same level of output produced by firms A and B. Firm C is therefore technically inefficient and could produce the same level of output while using fewer inputs. This technical inefficiency is represented graphically as the distance from point B to point C. The technical efficiency of firm C (TE_C) can be defined by the following ratio:

$$TE_C = \frac{OB}{OC} \tag{7}$$

where,

$0 \leq TE_C \leq 1$

A TE score of one implies that a firm is fully technically efficient. A firm with a TE score of less than one is inefficient; and the more inefficient the firm, the lower the TE score. Farrell (1957), using a similar graphical representation, claimed that the equation specified in (7) is the natural definition of technical efficiency. The TE score is the amount by which all inputs could be proportionally reduced without a reduction in output (Coelli *et al*, 2005).

The isocost line (MM') represents the different combinations of the two inputs that have the same total cost. The slope of the isocost reflects the relative prices of the two inputs, X_1 and X_2 (Nguyen &

Coelli, 2009). The isocost MM' is tangential to the isoquant at point A, implying that a firm operating at point A will produce the isoquant level of output at the minimum cost. Under the behavioural assumption of cost minimisation, a firm operating at point A is said to be allocatively efficient as it adopts the input mix that minimises its production costs given a particular level of output. Firm B is technically efficient, but is not allocatively efficient as it lies on a higher isocost line (NN'). Firm B could decrease its production costs by changing its mix of inputs, thereby increasing its allocative efficiency. This allocative inefficiency is represented graphically as the distance from point B to point D. The allocative efficiency of firm B (AE_B) can be defined by the following ratio:

$$AE_B = \frac{OD}{OB} \tag{8}$$

where,

$0 \leq AE_B \leq 1$

As is the case with the other efficiency scores, an AE score of one implies that a firm is fully allocatively efficient.

Cost efficiency is measure of total efficiency, keeping the firm's scale and external prices constant. Since firm C is neither technically nor allocatively efficient, it provides a good illustration of the concept of cost efficiency. The cost inefficiency of firm C is represented graphically as the distance from point C to point D. Line segment CD can be broken into line segment CB, representing technical inefficiency, and line segment BD, representing allocative inefficiency. The cost efficiency of firm C (CE_C) can therefore be defined as:

$$CE_C = \frac{OD}{OC} \tag{9}$$

where,

$0 \leq CE_C \leq 1$

The cost efficiency of firm C can also be defined as the product of its technical efficiency (TE_C) and allocative efficiency (AE_C), which is equivalent to the above definition:

$$CE_C = TE_C . AE_C = \frac{OB}{OC} . \frac{OD}{OB} = \frac{OD}{OC} \tag{10}$$

where,

$0 \leq CE_C \leq 1$

Note that, in the above example, firms B and C both lie on the same ray from the origin (OC). This results in the allocative efficiency of firm B being equal that of firm C ($AE_B = AE_C$).

It should also be noted that the types of efficiencies discussed in this chapter are not exhaustive. For example, one could examine the financing efficiency of a firm. The change in efficiency over time due to technological changes could also be examined. These changes could occur due to advances in technology, such as diagnostic and treatment methods or administration systems, which cause an upward shift in the production frontier. The reader is directed to Coelli *et al* (2005) and Sherman & Zhu (2006) for further information on these and other types of efficiencies.

3.10. Introducing efficiency measurement

Efficiency is generally unobservable and therefore has be measured indirectly using appropriate observable components (Nguyen & Coelli, 2009). These components often include the relationship between inputs and outputs, and their prices. Due to the importance of measuring efficiency, a wide range of techniques have been developed. None are completely satisfactory, but each attempts to capture different aspects of productivity and may be appropriate in different situations. Using a combination of techniques may provide additional insight into efficiency. Sherman & Zhu (2006) refer to a large number of techniques that can be used for measuring efficiency. These techniques include, *inter alia*, data envelopment analysis (DEA), ratio analysis, stochastic frontier analysis (SFA), standard cost systems, activity based costing, best practice analysis, comparative efficiency analysis, peer and management reviews, balanced scorecards, return on investment measures, and budgeting measures. Zere, McIntyre, Addison (2001) state that hospital efficiency may be measured by ratio analysis or production frontier models.

Techniques that specify production frontiers, such as SFA and DEA, can be grouped into two main categories, namely parametric and non-parametric techniques (Seiford & Thrall, 1990). Parametric techniques generally involve econometric estimation of the production frontier and *a priori* modelling decisions for which there is no widely accepted methodology (Smith & Street, 2005). Non-parametric techniques usually require less *a priori* decision making, and are therefore more objective (Sherman & Zhu, 2006). Production frontiers techniques measure the efficiency of a firm as the distance between the firm's observed level of inputs and outputs and the best practice production frontier (Linna, 1998). The differences between a parametric and a non-parametric technique are illustrated in Figure 3, which has been sourced from Nguyen & Coelli (2009). For illustrative purposes, SFA has been used as an example of a parametric technique and DEA has been used as an example of a non-parametric technique. Points A-F represent the unadjusted input-output combinations of firms A-F; while points A′, C′, D′ & E′ represent the error (or noise) adjusted input-output combinations of firms A, C, D & E. The parametric technique illustrated in Figure 3 specifies, *a priori*, the functional form

of the production frontier (such as the Cobb-Douglas functional form) as well as an error distribution (such as the half-normal or exponential distributions). The firms' input-output combinations are adjusted according to the error distribution. The functional form is fitted to the adjusted input-output combinations such that all combinations lie on or below the frontier. The functional form is usually fitted using econometric techniques such as corrected ordinary least squares or maximum likelihood (Nguyen & Coelli, 2009). This yields the smooth parametric frontier. The non-parametric frontier is constructed in a piecewise manner by linearly connecting the firms with the highest output-input combinations.

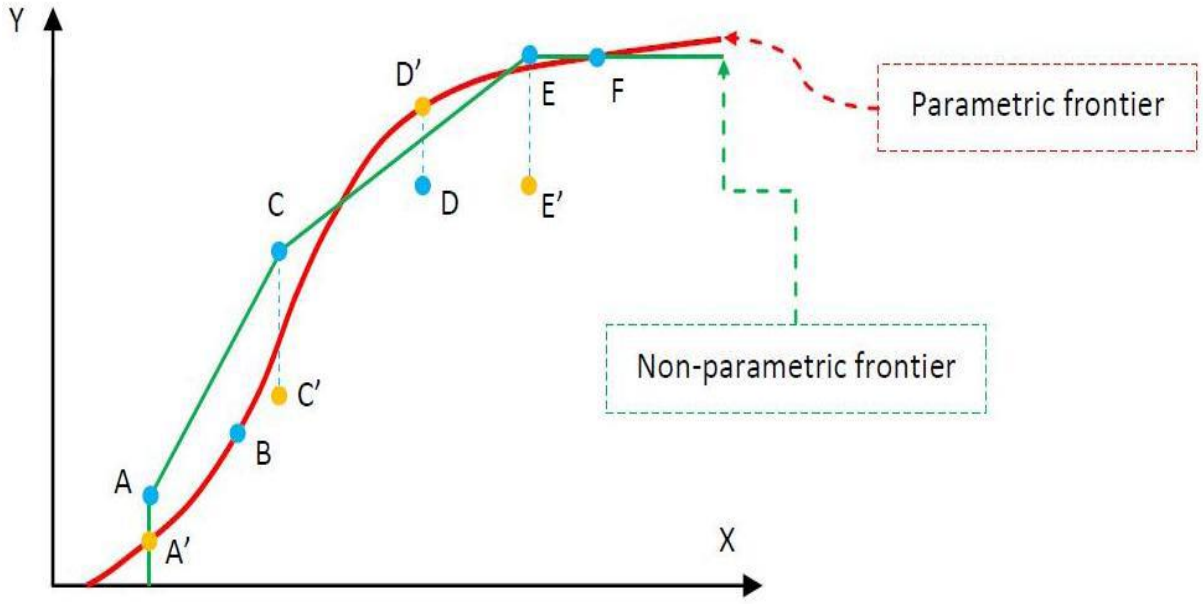

Figure 3: Production frontiers illustrating parametric and non-parametric techniques
Source: Nguyen & Coelli, 2009

This paper focuses on three efficiency measurement techniques, namely DEA, ratio analysis and SFA. These three techniques are expanded upon below.

3.11. Data Envelopment Analysis

Farrell, in his 1957 paper, developed the groundwork for Data Envelopment Analysis (DEA). DEA is a model that applies linear programming techniques to the data to construct a non-parametric, piecewise, linear production frontier. The DEA model is empirical and non-parametric as it does not require the specification of a functional form, error distribution or the relationships between inputs and outputs (Kibambe & Koch, 2007). According to Nguyen & Coelli (2009), the non-parametric nature of DEA is its most attractive feature. This is because, by not specifying a functional form, DEA avoids the possibility of incorrectly ascribing the effects of functional form misspecification as inefficiency. Since Farrell's paper (1957) there has been much research into DEA, resulting in a large

body of literature. Notable papers include Farrell (1957), Charnes, Cooper & Rhodes (1978) and Banker *et al* (1984). Seiford (1997) has compiled a bibliography of over 800 articles and dissertations (published between 1978 and 1996) which relate to DEA. Emrouznejada, Parkerb & Tavaresc (2008) identified more than 4,000 research articles published since the inception of DEA. Of these articles, banking, education and healthcare were found to be the most popular areas of research.

DEA is a model that "envelops" the dataset to identify best practice firms. The DEA model uses linear programming to allocate weights to each input and output in a way that maximises the productivity ratio of each firm (Sherman & Zhu, 2006). The weights are chosen in such a way that no other weights will result in a greater efficiency score for each firm (Charnes *et al*, 1978). This means that DEA tends to understate the inefficiency of a firm. This is desirable from the perspective that an efficient firm will not be labelled as inefficient. However, understating the inefficiency of each firm may result in many of the firms in the sample being labelled as efficient (Sherman & Zhu, 2006). DEA assigns each firm an efficiency score from zero to one, where an efficiency score of one is a fully efficient firm. The efficiency scores can be used to rank the efficiencies of the group of firms. Note that DEA does not require the existence or specification of an absolute efficient standard, as the model determines the best practice firms by comparing the actual operating results of a group of firms (Sherman & Zhu, 2006). This may be viewed as an advantage of the DEA model as it may not be possible to specify an efficient standard; however it may be viewed as a disadvantage as the model cannot determine whether a firm that lies on the DEA production frontier could further improve its efficiency. A DEA efficient firm is efficient relative to the other firms in the group, but it may still be inefficient when compared to an absolute efficiency standard. However, within a hospital context, an absolute efficiency standard is unlikely to exist because of the difficulties associated with measuring health outcomes as well as the complex nature of the production process.

DEA has the advantage of being able to cope with multiple-inputs and multiple-outputs simultaneously, as well as allowing each of the inputs and outputs to be measured in different units (Valdmanis, Rosko & Mutter, 2008). This allows DEA to incorporate inputs or outputs which do not have a clear price or market value, such as staff training or research and development activities (Sherman & Zhu, 2006).

Once the best practice firms have been identified, the firms that are not best practice can be benchmarked relative to the best practice firms. DEA allows the possible drivers of inefficiency to be analysed and quantified for each firm which is not best practice (Valdmanis *et al*, 2008). Expressed differently, DEA quantifies the reduction in inputs (or the increase in outputs) that is necessary for an inefficient firm to move onto the efficient frontier.

Importantly, DEA allows the calculation of technical efficiency, allocative efficiency, cost efficiency and scale efficiency (Coelli *et al*, 2005). However, price information and a behavioural assumption (such as profit maximisation or cost minimisation) are required in order to calculate allocative efficiency and hence cost efficiency (Nguyen & Coelli, 2009).

A major disadvantage of the DEA model is that, since it is a deterministic model, it does not account for measurement error or random noise (Zere *et al*, 2001). This means that the DEA model ascribes any deviation from the production frontier as being due to inefficiency, when in fact it may be due to measurement error or random noise. Random noise could simply be a result of random fluctuations or once-off impacts that are outside of the control of management (Worthington, 2004). Examples of uncontrollable events in a hospital environment include the cost of unexpected hospital repairs, the outbreak of epidemics, and the resignation of key staff (Jacobs, 2001).

3.12. Ratio analysis

Ratio analysis is an intuitive and simple method of measuring efficiency (Sherman & Zhu, 2006), and involves assessing the efficiency of a firm by analysing different key ratios (Zere *et al*, 2001). One such ratio (discussed in section 3.2) is the productivity ratio, which is the ratio of outputs to inputs. Many different ratios exist, and are intended to focus on different aspects of a firm's operations. A set of ratios for each firm can be compared with other similar firms to assess their performance. Alternatively, ratios can be compared within a single firm over different time periods (Sherman & Zhu, 2006).

Sherman (1984) explains how ratio analysis is particularly useful at identifying relationships that are abnormally high or low, such as a very high cost per unit output. These abnormal relationships can be identified, for example, as large deviations from the mean. Management can then focus their attention on correcting these abnormal relationships. However, ratio analysis does not provide an objective way of identifying inefficient firms, such as an objective cut-off point that classifies a firm as inefficient. Arbitrary cut-off points are often used, such as one standard deviation above the mean (Sherman & Zhu, 2006). This involves an element of judgement and reduces the credibility of the method, as there is no way to ensure that firms operating below the cut-off point are in fact efficient. This is in contrast to DEA, which can objectively identify a firm as inefficient relative to its peers without the need for an efficiency standard (Sherman & Zhu, 2006).

Each ratio is limited to a single-input and a single-output and cannot easily be extended to the multiple-input, multiple-output case. Multiple-inputs and multiple-outputs are sometimes aggregated using weights, which can be subjective and arbitrary. According to Coelli *et al* (2005), using ratio analysis for multiple-inputs and multiple-outputs may not be useful and may in fact be misleading.

Furthermore, ratio analysis requires homogeneous measurement units for the aggregation of inputs and outputs (Zere *et al*, 2001). DEA has the advantage that it does not require homogeneous measurement units and can accommodate multiple-inputs and multiple-outputs without the need for subjective weights. This means that DEA can identify efficiency improvements that cannot be identified by ratio analysis (Sherman & Zhu, 2006). Nevertheless, ratio analysis is often complementary when used with other efficiency measures, such as DEA or SFA. In particular, ratio analysis can provide an intuitive check on the results of other methods.

Since each ratio looks at a single component of efficiency, a large number of ratios are often developed (Sherman, 1984). A study done by Cleverley, Stanko, & Zeller (1997) found that typical financial reports may contain as many as 30 ratios. The simultaneous interpretation of a large number of ratios is a complex process that is likely to involve judgement. This reduces the simplicity of the ratio analysis method, which is arguably its greatest appeal (Sherman & Zhu, 2006).

Ratio analysis involves simple mathematical concepts which are easy to understand, while DEA involves more complex mathematics. Management may not be comfortable with the DEA model, which may be viewed as a black box (Sherman & Zhu, 2006). This means that the DEA model may be accompanied by the additional cost of training management in its use.

3.13. Stochastic Frontier Analysis

Stochastic frontier analysis (SFA) is a parametric technique that is used to specify the production frontier. Unlike DEA, SFA makes use of an econometric model that can accommodate random noise (Jacobs, 2001). SFA is therefore able to isolate deviations from the production frontier into two components, one representing random noise and the other inefficiency; while DEA assumes that all deviations from the production frontier are due to inefficiency (Linna, 1998). This also means that DEA is more sensitive to outliers than SFA (O'Neill, Raunerb, Heidenbergerb & Krausc, 2008). The stochastic nature of the SFA model provides a basis for statistical inference, such as hypothesis testing on the efficiency scores (Nguyen & Coelli, 2009). As with DEA, SFA allows the calculation of technical efficiency, allocative efficiency, cost efficiency and scale efficiency (Coelli *et al*, 2005).

Within the SFA model, the production frontier and the random error distribution must be specified *a priori*. The production frontier assumes a functional form that specifies the relationship between inputs and outputs. Commonly assumed functional forms include the Cobb-Douglas, normalised quadratic and translog functional forms (Smith & Street, 2005). Once the functional form has been chosen, it needs to be parameterised using econometric techniques. Common parameterisation techniques include corrected ordinary least squares, feasible generalised least squares, and maximum likelihood (Nguyen & Coelli, 2009). Each of these techniques requires various modelling decisions,

which involve judgement (Smith & Street, 2005). Note that the functional form is fitted to the adjusted input-output combinations so that all combinations lie on or below the fitted production frontier (Coelli *et al*, 2005).

The SFA random noise (error) distribution is commonly assumed, *a priori*, to follow a half-normal, truncated normal, exponential or gamma distribution (Nguyen & Coelli, 2009). As with the production frontier functional form, there is no way to test whether the random noise distribution has been appropriately specified (Jacobs, 2001). It is likely that a different error distribution, or a different functional form, will result in different efficiency scores (Coelli *et al*, 2005). The *a priori* specification of an error distribution and a functional form are both strong assumptions which make SFA vulnerable to model misspecification (O'Neill *et al*, 2008). Furthermore, the inappropriate specification of a functional form or an error distribution may confound inefficiency with model misspecification (Lovell, 1996). This detracts from the advantage of SFA being able to distinguish between random noise and inefficiency (Lovell, 1996). DEA, in contrast, does not require these *a priori* assumptions.

DEA and SFA can be used as complementary techniques, as well as a useful check on different efficiency measures (Jacobs, 2001). However, both techniques are fairly complicated and possibly resource intensive. Coelli *et al* (2005) explains that the context of any investigation should guide the selection of an efficiency measure.

3.14. Selecting an efficiency measurement technique

From the above discussions and comparisons in sections 3.10 to 3.13, it can be concluded that ratio analysis fails to adequately capture the dynamics of a multiple-input, multiple-output production process; while SFA relies heavily on *a priori* assumptions regarding the functional form of the production frontier and the random error distribution. Within a South African context, there is a lack of research that can be used to inform any *a priori* assumptions relating to the hospital production process. Furthermore, there is no absolute efficiency standard for the hospital industry that can be used to inform the functional form of the production process. DEA is a multiple-input, multiple-output efficiency measurement technique that does not require the existence of an absolute efficiency standard nor does it require *a priori* assumptions. DEA is therefore the most appropriate efficiency measurement technique for this paper's investigation.

Additionally, DEA facilitates the calculation of both technical efficiency and scale efficiency, which is essential for this paper's investigation. DEA also provides additional information which is useful to management, such as the potential resource savings of inefficient hospitals. This can be combined with qualitative management input to further increase the usefulness of the DEA results. For these

reasons, DEA was selected as the efficiency measurement technique for this paper. Further details and the practical applications of the DEA model are discussed in the next chapter.

Note that, since the data analysed in this paper are relatively clean, the inability of DEA to account for data measurement issues was less of a concern. The reader is directed to chapter 5 for information regarding the data used in this investigation.

4. Details and practical applications of the DEA model

4.1. Overview of the details and practical applications of the DEA model

The details of the DEA model are expanded upon in this chapter. Section 4.2 discusses the input-orientated and output-orientated DEA models. This is followed in section 4.3 by the mathematical formulation of the DEA model, which is extended to include returns to scale in section 4.4. The specification of input and output variables and the limitations of DEA are covered in sections 4.5 and 4.6 respectively. This chapter concludes with a discussion of the practical application of DEA in the hospital industry, with particular focus on the South African hospital industry.

4.2. Model orientation

The DEA model can be specified as an input-orientated model or an output-orientated model. An input-orientated model identifies efficiency improvements as a proportional reduction in input usage; while an output-orientated model identifies efficiency improvements as a proportional increase in output production (Coelli, 1996).

Both the input-orientated model and output-orientated model will identify the same firms as being efficient, producing the same efficient frontier (Coelli *et al*, 2005). In fact, the two models produce equivalent results under the constant returns to scale assumption. However, under a variable returns to scale assumption, the efficiency scores of inefficient firms may differ between the two models (Nguyen & Coelli, 2009).

The selection of a particular model orientation depends on the dynamics of the industry being modelled. An input-orientated model is most appropriate for an industry that can manage its resource use, but has little or no control over the demand for its outputs (Sherman & Zhu, 2006). An input-orientated model is therefore appropriate when management is concerned with minimising the resources required to produce a target level of output. Similarly, an output-orientated model is most appropriate when management is required to maximise outputs for a given level of input. This can occur when a firm is faced with input constraints (Nguyen & Coelli, 2009). In practice, management may wish to reduce inputs while simultaneously increasing outputs. However, the main consideration when selecting a model orientation for a DEA study should be whether management has more control over the inputs or more control over the outputs used in the production process (Coelli *et al*, 2005).

Within the healthcare environment, O'Neill *et al* (2008) claim that the objective of most studies is to reduce costs rather than to increase the provision of services, which naturally leads to the selection of an input-orientated DEA model. Sherman & Zhu (2006) explain that hospital management can

influence the amounts of inputs that are used to provide services to patients, but generally have little control over the demand for healthcare services and hence the number of outputs that a hospital produces. This too naturally leads to the selection of an input-orientated DEA model for studies of hospital efficiency. Within a South African context, by law, doctors cannot be directly employed by hospitals. However, these doctors are largely responsible for admission decisions. This means that the admission decisions, which influence the demand for healthcare services, are also outside of the direct control of hospital management. For these reasons, an input-orientated DEA model was selected for this paper's hospital efficiency analysis.

4.3. Model specifications

The DEA models described in this section are input-orientated models. The output-orientated models are not discussed here, but have similar specifications. For a full description of the output-orientated models, the reader is directed to Sherman & Zhu (2006). The specification of DEA models in this section and section 4.4 has been aided by the works of Banker (1984); Banker *et al* (1984); Banker, Bardhan & Cooper (1996a); Banker, Chang & Cooper (1996b); Banker, Cooper, Seiford, Thrall & Zhu (2004); Banker & Thrall (1992); Charnes & Cooper (1989); Charnes *et al* (1978); Coelli *et al* (2005); Mehrabian, Jahanshahloo, Alirezaee & Amin (2000); Ramanathan (2003); Sherman & Zhu (2006); and Tone (1995).

The DEA model, proposed in an influential paper by Charnes, Cooper & Rhodes (1978), has been widely applied and extended. This DEA model is commonly referred to as the CCR model. The CCR model is a constant returns to scale model, that was first proposed in its input-orientated form. Note that all CCR models discussed in this section are constant returns to scale models. Variable returns to scale models are discussed in the next section.

The CCR model can be expressed, most intuitively, in its ratio form. At a high level, the model attempts to maximise the productivity ratio of each firm in the analysis, which is the ratio of outputs to inputs. This is done by assigning weights to the inputs and outputs that are determined by linear programming and the observed data (Charnes *et al*, 1978). Note that the weights are not specified *a priori*, but are derived directly from the observed data. The productivity ratio is used as a measure of technical efficiency and is constrained so that it is less than or equal to one. This constraint can be interpreted as ensuring that each firm lies on or below the efficient frontier (Zere *et al*, 2001). The CCR ratio model is specified by the following linear programme:

$$\max \ \theta_i = \frac{\boldsymbol{u_i y_i}}{\boldsymbol{v_i x_i}}$$

subject to,

$$\frac{\boldsymbol{u_i y_j}}{\boldsymbol{v_i x_j}} \leq 1, \qquad j = 1, 2, \dots, n$$

$$\boldsymbol{u}, \boldsymbol{v} \geq 0 \tag{11}$$

where,

θ_i is the technical efficiency score for firm i,

$\boldsymbol{u_i}$ is a $1 \times s$ vector of output weights for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

$\boldsymbol{v_i}$ is a $1 \times m$ vector of input weights for firm i,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

The model is run for each firm in the dataset ($i = 1, 2, \dots, n$), resulting in: a set of technical efficiency scores ($\theta_1, \dots, \theta_n$), a set of vectors of output weights ($\boldsymbol{u_1}, \dots, \boldsymbol{u_n}$), and a set of vectors of input weights ($\boldsymbol{v_1}, \dots, \boldsymbol{v_n}$).

However, this particular ratio form of the CCR model is not used in practice as the model has an infinite number of solutions (Coelli *et al*, 2005). If $(\boldsymbol{u_i^*, v_i^*})$ represents an optimal solution to the linear programme then any scalar multiple, $(\alpha \boldsymbol{u_i^*}, \alpha \boldsymbol{v_i^*})$, is also an optimal solution (Coelli *et al*, 2005). Unlike the above ratio model, the CCR models that are discussed below do not have an infinite number of solutions.

Charnes *et al* (1978) go on to define the multiplier form and envelopment form of the CCR model. These two models are dual linear programmes. This means that they are equivalent and have the same optimal solutions. The CCR multiplier model is specified by the following linear programme:

$$\max \ \theta_i = \boldsymbol{u_i y_i}$$

subject to,

$$\boldsymbol{u_i Y - v_i X \leq 0}$$
$$\boldsymbol{v_i x_i} = 1$$
$$\boldsymbol{u_i, v_i \geq 0} \tag{12}$$

where,

θ_i is the technical efficiency score for firm i,

$\boldsymbol{u_i}$ is a $1 \times s$ vector of output weights for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

$\boldsymbol{Y}$ is a $s \times n$ matrix comprised of n column vectors, $\boldsymbol{y_1}, \dots, \boldsymbol{y_n}$,

$\boldsymbol{v_i}$ is a $1 \times m$ vector of input weights for firm i,

$\boldsymbol{X}$ is a $m \times n$ matrix comprised of n column vectors, $\boldsymbol{x_1}, \dots, \boldsymbol{x_n}$,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

In the CCR multiplier model, the constraint $\boldsymbol{v_i x_i} = 1$ ensures that the model does not have an infinite number of solutions. The output and input weights ($\boldsymbol{u_i}$ and $\boldsymbol{v_i}$ respectively) are determined in such a way that maximises the efficiency score for firm i (θ_i), given the constraints of the linear programme and the observed dataset. A firm is only labelled as inefficient after all possible weights have been considered, and no other weights will provide a higher efficiency rating (Charnes *et al*, 1978). DEA can be thought of as giving the benefit of the doubt to each firm when calculating their efficiency score (Sherman, 1984). Sherman (1984) claims that this is a strength of DEA because an efficient firm will not be labelled as inefficient. However, this could also lead to firms being labelled as more efficient than they actually are.

The input and output weights are sometimes used to derive the rate of substitution of inputs and the rate of substitution of outputs (Sherman & Zhu, 2006). However, Coelli *et al* (2005) caution that the rates of substitution derived in this manner do not necessarily apply in practice, and may even be unrealistic for some firms.

The dual of the CCR multiplier model is the envelopment model. The envelopment model provides the user with the efficiency reference set for each inefficient firm. This allows the user to quantify the level of inefficiency of the firm in terms of units of inputs and outputs. The CCR envelopment model is specified by the following linear programme:

$$\min \ \theta_i$$

subject to,

$$\theta_i \boldsymbol{x_i} \geq \boldsymbol{X\lambda_i}$$
$$\boldsymbol{y_i} \leq \boldsymbol{Y\lambda_i}$$
$$\boldsymbol{\lambda_i} \geq \boldsymbol{0} \tag{13}$$

where,

θ_i is the technical efficiency score for firm i,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

$\boldsymbol{X}$ is a $m \times n$ matrix comprised of n column vectors, $\boldsymbol{x_1}, \dots, \boldsymbol{x_n}$,

$\boldsymbol{\lambda_i}$ is a $n \times 1$ vector of lambda weights for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

$\boldsymbol{Y}$ is a $s \times n$ matrix comprised of n column vectors, $\boldsymbol{y_1}, \dots, \boldsymbol{y_n}$,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

The model is run for each firm in the dataset ($i = 1, 2, \dots, n$), resulting in a set of technical efficiency scores ($\theta_1, \dots, \theta_n$), and a set of lambda weights ($\boldsymbol{\lambda_1}, \dots, \boldsymbol{\lambda_n}$). The CCR envelopment linear programme can be interpreted as attempting to minimise the efficiency score of firm i, subject to three constraints: the inputs of firm i must be greater than or equal to the lambda weighted inputs of the other firms; the outputs of firm i must be less than than or equal to the lambda weighted outputs of the other firms; and the lambda weights must be non-negative (Sherman & Zhu, 2006).

Note that in the multiplier model for firm i, an individual weight (i.e. a component of vectors $\boldsymbol{u_i}$ and $\boldsymbol{v_i}$) is assigned to each input and output variable; while in the envelopment model for firm i, an individual weight (i.e. a component of vector $\boldsymbol{\lambda_i}$) is assigned to each firm in the dataset.

Intuitively, the envelopment model minimises the efficiency score of a firm i, while assigning a non-negative lambda weight to each of the firms in the dataset. In order to minimise the efficiency score of firm i, the model radially contracts the firm's input vector, $\boldsymbol{x_i}$, until the firm lies on the efficient frontier (Coelli *et al*, 1995). The efficiency score is therefore a measure of the radial distance from the efficient frontier to firm i. It quantifies the minimum proportional reduction in inputs that is required for an inefficient firm to operate on the efficient frontier (Anderson & Peterson, 1993). While minimising the efficiency score for firm i, the lambda weights (i.e. the components of vector $\boldsymbol{\lambda_i}$) are simultaneously assigned non-negative values. The firms that are not assigned zero weights are all

efficient firms (Sherman & Zhu, 2006). Collectively, these firms are referred to as the efficiency reference set (ERS) of firm i. The ERS can be used to project the current production point of firm i onto the efficient frontier. The projected point corresponds to the point that was identified using the efficiency score of firm i (the production point that requires the minimum proportional reduction in inputs for firm i to operate on the efficient frontier). This projected point is a linear combination of the production points of the firms in the ERS, and is identified by weighting the input and output vectors of each firm in the ERS by that firm's corresponding lambda value. The resulting input and output vectors are then summed, producing a vector of inputs and outputs that lie on the efficient frontier. In matrix notion, firm i is projected onto the efficient frontier at point $(\boldsymbol{X\lambda_i}, \boldsymbol{Y\lambda_i})$ (Coelli *et al*, 2005). Furthermore, lambda vectors can be used to define the entire production possibility set. Tone (1995) defines the production possibility set as follows:

$$P = \{(\boldsymbol{x}, \boldsymbol{y}): \boldsymbol{x} \geq \boldsymbol{X\lambda}, \boldsymbol{y} \leq \boldsymbol{Y\lambda}, \boldsymbol{\lambda} \geq 0\} \tag{14}$$

where,

P is the production possibility set,

$\boldsymbol{x}$ is a $m \times 1$ vector of inputs,

$\boldsymbol{y}$ is a $s \times 1$ vector of outputs,

$\boldsymbol{X}$ is a $m \times n$ matrix comprised of n column vectors, $\boldsymbol{x_1}, \dots, \boldsymbol{x_n}$,

$\boldsymbol{\lambda}$ is a $n \times 1$ vector of lambda weights that satisfies the CCR linear programme in (13),

$\boldsymbol{Y}$ is a $s \times n$ matrix comprised of n column vectors, $\boldsymbol{y_1}, \dots, \boldsymbol{y_n}$,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

The particular forms of the multiplier and envelopment CCR models, which have been discussed above, do not account for any *slacks* that may be present in the linear programme. An input slack exists when an input of a firm can be reduced without impacting the efficiency score of the firm or violating the constraints of the linear programme (Charnes *et al*, 1978). Similarly, an output slack exists when an output of a firm can be increased without impacting the efficiency score of the firm or violating the constraints of the linear programme. The CCR envelopment model that accounts for slacks is specified by the following linear programme:

$$\min \ \theta_i - \varepsilon(\boldsymbol{I_x s_x} + \boldsymbol{I_y s_y})$$

subject to,

$$\begin{gathered}\theta_i \boldsymbol{x_i} = \boldsymbol{X\lambda_i} + \boldsymbol{s_x} \\ \boldsymbol{y_i} = \boldsymbol{Y\lambda_i} - \boldsymbol{s_y} \\ \boldsymbol{\lambda_i, s_x, s_y} \geq \boldsymbol{0}\end{gathered} \tag{15}$$

where,

θ_i is the technical efficiency score for firm i,

ε is a non-Archimedean infinitesimal,

$\boldsymbol{I_x}$ is a $1 \times m$ vector of ones,

$\boldsymbol{s_x}$ is a $m \times 1$ vector of input slacks,

$\boldsymbol{I_y}$ is a $1 \times s$ vector of ones,

$\boldsymbol{s_y}$ is a $s \times 1$ vector of output slacks,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

$\boldsymbol{X}$ is a $m \times n$ matrix comprised of n column vectors, $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$,

$\boldsymbol{\lambda_i}$ is a $n \times 1$ vector of lambda weights for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

$\boldsymbol{Y}$ is a $s \times n$ matrix comprised of n column vectors, $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

The multiplier form of the CCR model, which was specified by the linear programme in (12), is easily adapted to account for slacks. In order to achieve this, the constraints $\boldsymbol{u_i, v_i} \geq \boldsymbol{0}$ in (12) are replaced with $\boldsymbol{u_i} \geq \varepsilon \boldsymbol{I_y}$ and $\boldsymbol{v_i} \geq \varepsilon \boldsymbol{I_x}$ (where $\varepsilon, \boldsymbol{I_x}$ and $\boldsymbol{I_y}$ have the same meaning as defined in (15)).

The inclusion of the input and output slacks in the linear programme ensures that the observed inputs and outputs for an efficient firm do in fact lie on the efficient frontier. Each input is associated with an input slack, and each output with an output slack. Each slack is constrained so that it is non-negative. This means that a slack represents the absolute amount that an input of a firm can be reduced, or an output increased. Slacks are a consequence of the piecewise linear nature of the DEA model, which enables the model to define sections of the efficient frontier that run parallel to the axes (Coelli *et al*, 2005). Slacks may occur in these sections of the frontier. In certain cases, a firm may appear efficient but may be able to decrease its input usage (or increase its output production) by moving along the parallel section of the efficient frontier.

In the above linear programme (15), the non-Archimedean infinitesimal, ε, is a theoretical construct that is defined to be less than any real number (Mehrabian *et al*, 2000). The non-Archimedean infinitesimal was introduced into the CCR models to solve the problem that, under certain circumstances, the models rated a firm as efficient even though non-zero slacks existed (Ramanathan, 2003).

The inclusion of slacks in the envelopment model means that the DEA efficiency score for firm i, θ_i, is no longer a sufficient measure of efficiency. This means that the measure of DEA efficiency must account for any slacks that are present (Charnes *et al*, 1978). Therefore firm i is defined as fully DEA efficient if the following two conditions must be met:

1. $\theta_i = 1$, and
2. The slack variables are all zero.

An inefficient firm's projection onto the efficient frontier must also be updated to reflect the possibility that slacks may be present. When not accounting for slacks, firm i is projected onto the efficient frontier at point $(\theta_i \boldsymbol{x_i}, \boldsymbol{y_i})$. When slacks are accounted for, the point of projection onto the efficient frontier for firm i becomes $(\theta_i \boldsymbol{x_i} - \boldsymbol{s_x}, \boldsymbol{y_i} + \boldsymbol{s_y})$. Note that when the slack variables are all zero, the two projection points coincide. In this paper all efficiency scores have been adjusted for the presence of slacks.

4.4. Returns to scale

The Charnes, Cooper & Rhodes (1978) model is limited by the fact that it can only model technology that exhibits constant returns to scale. Banker, Charnes & Cooper (1984) extended the CCR model to allow for variable returns to scale. This DEA model is commonly referred to as the BCC model. The BCC model is widely used, particularly for the study of returns to scale.

Banker *et al* (1984) showed that the BCC model could be used to classify firms as exhibiting constant, increasing, or decreasing returns to scale. Banker & Thrall (1992) conducted extensive research on returns to scale, and made more precise the work of Banker *et al* (1984). However, the concept of returns to scale is only well defined for efficient points on the production frontier, with points lying within the interior of the production frontier being more difficult to classify. The term *global returns to scale* refers to the return to scale classifications of efficient firms operating on the frontier; while *local returns to scale* refers to the return to scale classifications of inefficient firms operating within the interior of the frontier. Färe, Grosskopf & Lovell (1985) and Banker *et al* (1996b) proposed solutions to the local returns to scale classification problem, which were extended and simplified by Tone (1996).

The starting point for analysing returns to scale is the BCC model. Like the CCR model, the BCC model can be specified in both envelopment and multiplier forms. The BCC envelopment model is specified by the following linear programme:

$$\min \ \theta_i - \varepsilon(\boldsymbol{I_x s_x} + \boldsymbol{I_y s_y})$$

subject to,

$$\begin{gathered} \theta_i \boldsymbol{x_i} = \boldsymbol{X\lambda_i} + \boldsymbol{s_x} \\ \boldsymbol{y_i} = \boldsymbol{Y\lambda_i} - \boldsymbol{s_y} \\ \boldsymbol{I_n \lambda_i} = 1 \\ \boldsymbol{\lambda_i, s_x, s_y} \geq \boldsymbol{0} \end{gathered} \tag{16}$$

where,

θ_i is the technical efficiency score for firm i,

ε is a non-Archimedean infinitesimal,

$\boldsymbol{I_x}$ is a $1 \times m$ vector of ones,

$\boldsymbol{s_x}$ is a $m \times 1$ vector of input slacks,

$\boldsymbol{I_y}$ is a $1 \times s$ vector of ones,

$\boldsymbol{s_y}$ is a $s \times 1$ vector of output slacks,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

$\boldsymbol{X}$ is a $m \times n$ matrix comprised of n column vectors, $\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}$,

$\boldsymbol{\lambda_i}$ is a $n \times 1$ vector of lambda weights for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

$\boldsymbol{Y}$ is a $s \times n$ matrix comprised of n column vectors, $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$,

$\boldsymbol{I_n}$ is a $1 \times n$ vector of ones,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

The above BCC envelopment linear programme is almost identical to the CCR linear programme specified in (15), the difference being that (16) includes an additional constraint, namely $\boldsymbol{I_n \lambda_i} = 1$. This constraint is referred to as the convexity constraint (Coelli *et al*, 2005). The convexity constraint allows the BCC model to construct an efficient frontier that allows for variable returns to scale (VRS). In Figure 4, points A, B, C, D & E represent firms in the dataset that use input x to produce output y. Under constant returns to scale (CRS), the efficient frontier is represented by line segment OF. While, under VRS, the efficient frontier is represented by the piecewise line segments joining points A, B, C & D. As can be seen in Figure 4, the VRS efficient frontier follows the data points more closely than

the CRS frontier. This means that DEA efficiency scores obtained under VRS will be greater than or equal to the efficiency scores obtained under CRS (Coelli *et al*, 2005). The increasing returns to scale (IRS) section of the efficient frontier is represented by line segment AB, and the decreasing returns to scale (DRS) section by line segment CD. The IRS and CRS sections of the frontier, represented by the line segments joining points A, B & C, are together referred to as non-decreasing returns to scale (NDRS) section of the frontier. Similarly, the DRS and CRS sections of the frontier, represented by the line segments joining points B, C & D, are together referred to as the non-increasing returns to scale (NIRS) section of the frontier.

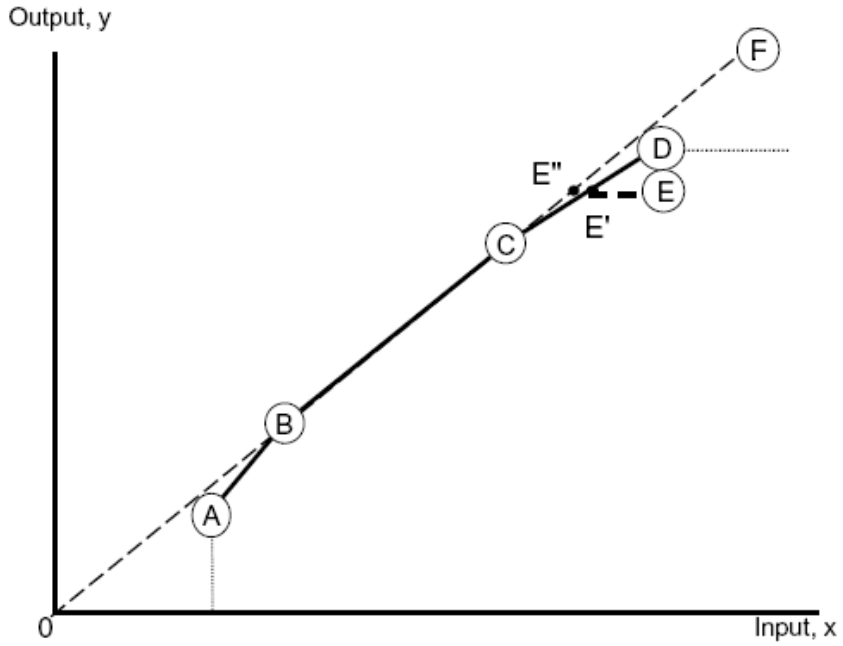

Figure 4: Production frontiers under different returns to scale assumptions
Source: adapted from Banker *et al*, 2004

A VRS model allows inefficient firms to be benchmarked against firms of similar size, while a CRS model benchmarks inefficient firms against firms operating at optimal size. Under CRS, firm E is projected onto the efficient frontier at point E", which is operating at optimal scale. While under VRS, firm E is projected onto the efficient frontier at point E', which exhibits DRS. An assumption of CRS is likely to be appropriate when all firms are operating at optimal scale, and an assumption of VRS is likely to be appropriate when all firms are not operating at optimal scale. If a CRS specification is used when all firms are not operating at optimal scale, then a firm may be inappropriately benchmarked against firms that are operating at a substantially different size. This may result in the effects of technical inefficiency and scale inefficiency being confounded (Coelli *et al*, 2005). Using a VRS specification will avoid this problem.

The BCC envelopment model, specified by the linear programme in (16), can be modified to enforce NIRS and NDRS. This is achieved by altering the convexity constraint – the constraint that ensures that the lambda weights sum to one ($\boldsymbol{I_n\lambda_i} = 1$). The returns to scale of the envelopment model can be specified by replacing the convexity constraint in (16) by any one of the following constraints:

$$\boldsymbol{I_n\lambda_i} = 1 \Rightarrow VRS$$

$$\boldsymbol{I_n\lambda_i} \leq 1 \Rightarrow NIRS$$

$$\boldsymbol{I_n\lambda_i} \geq 1 \Rightarrow NDRS$$

$$\boldsymbol{I_n\lambda_i}\ is\ unconstrained \Rightarrow CRS \tag{17}$$

where,

$\boldsymbol{I_n}$ is a $1 \times n$ vector of ones,

$\boldsymbol{\lambda_i}$ is a $n \times 1$ vector of lambda weights for firm i, and

n is the total number of firms in the dataset.

Note that when $\boldsymbol{I_n\lambda_i}$ is unconstrained, the BCC envelopment model (16) reduces to the CCR envelopment model (15). The relationships in (17) are a result of the works by Banker *et al* (1984) and Banker & Thrall (1992). The reader is directed to these papers for the technical derivations of these relationships, as well as further information. Note that the converses of the relationships in (17) do not necessarily hold. If the value of $\boldsymbol{I_n\lambda_i}$ is calculated using the CCR model, where $\boldsymbol{I_n\lambda_i}$ is unconstrained, the resulting values of $\boldsymbol{I_n\lambda_i}$ imply the following:

$$\boldsymbol{I_n\lambda_i} = 1\ for\ \boldsymbol{any}\ alternate\ optima \Rightarrow CRS$$

$$\boldsymbol{I_n\lambda_i} > 1\ for\ \boldsymbol{all}\ alternate\ optima \Rightarrow DRS$$

$$\boldsymbol{I_n\lambda_i} < 1\ for\ \boldsymbol{all}\ alternate\ optima \Rightarrow IRS \tag{18}$$

where,

$\boldsymbol{I_n}$ is a $1 \times n$ vector of ones,

$\boldsymbol{\lambda_i}$ is a $n \times 1$ vector of lambda weights for firm i, and

n is the total number of firms in the dataset.

Note that the above conditions only apply to firms operating on the efficient frontier, or projections of inefficient firms onto the efficient frontier. Note also that for DRS or IRS to prevail, the inequalities must be true for all optimal solutions of the CCR linear programme. The process of examining all the alternate optima can be complicated and laborious. However, the reader is directed to Banker *et al* (1996b) for a method that reduces the computational complexity of this task.

The BCC envelopment model can also be specified as a multiplier model. The following linear programme specifies the BCC multiplier model:

$$\max \ \theta_i = \boldsymbol{u_i y_i} - u_0$$

subject to,

$$\begin{gathered} \boldsymbol{u_i Y} - \boldsymbol{v_i X} - u_0 \boldsymbol{I_n} \leq \boldsymbol{0} \\ \boldsymbol{v_i x_i} = 1 \\ \boldsymbol{u_i} \geq \varepsilon \boldsymbol{I_y} \\ \boldsymbol{v_i} \geq \varepsilon \boldsymbol{I_x} \end{gathered} \tag{19}$$

where,

θ_i is the technical efficiency score for firm i,

$\boldsymbol{u_i}$ is a $1 \times s$ vector of output weights for firm i,

$\boldsymbol{y_i}$ is a $s \times 1$ vector of observed outputs for firm i,

u_0 is a scalar,

$\boldsymbol{Y}$ is a $s \times n$ matrix comprised of n column vectors, $\boldsymbol{y_1}, \dots, \boldsymbol{y_n}$,

$\boldsymbol{v_i}$ is a $1 \times m$ vector of input weights for firm i,

$\boldsymbol{X}$ is a $m \times n$ matrix comprised of n column vectors, $\boldsymbol{x_1}, \dots, \boldsymbol{x_n}$,

$\boldsymbol{I_n}$ is a $1 \times n$ vector of ones,

$\boldsymbol{x_i}$ is a $m \times 1$ vector of observed inputs for firm i,

ε is a non-Archimedean infinitesimal,

$\boldsymbol{I_y}$ is a $1 \times s$ vector of ones,

$\boldsymbol{I_x}$ is a $1 \times m$ vector of ones,

m is the total number of input variables used in the model,

s is the total number of output variables used in the model, and

n is the total number of firms in the dataset.

The BCC multiplier linear programme is similar to the CCR multiplier linear programme, which was specified in (12), the difference being that the BCC multiplier linear programme now includes u_0 (a scalar). In this model, the value of u_0 is unconstrained which results in a VRS specification. The returns to scale of the multiplier model (19) can be specified by the addition of any one of the following constraints on the value of u_0:

$$u_0 \text{ is unconstrained} \Rightarrow VRS$$

$$u_0 \leq 0 \Rightarrow NIRS$$

$$u_0 \geq 0 \Rightarrow NDRS$$

$$u_0 = 0 \Rightarrow CRS \tag{20}$$

where,

u_0 is a scalar.

Note that when the value of u_0 is set to zero, the BCC multiplier model (20) reduces to the CCR multiplier model (12). The relationships in (20) are a result of the works by Banker *et al* (1984) and Banker & Thrall (1992). The reader is directed to these papers for the technical derivations of these relationships, as well as further information.

When u_0 is unconstrained, as is the case in (19), the outputted value of u_0 implies the following:

$$u_0 = 0 \text{ for } \textbf{any} \text{ alternate optima} \Rightarrow CRS$$

$$u_0 < 0 \text{ for } \textbf{all} \text{ alternate optima} \Rightarrow DRS$$

$$u_0 > 0 \text{ for } \textbf{all} \text{ alternate optima} \Rightarrow IRS \tag{21}$$

where,

u_0 is a scalar.

Note that the above conditions only apply to firms operating on the efficient frontier, or projections of inefficient firms onto the efficient frontier. Note also that for DRS or IRS to prevail, the inequalities must be true for all optimal solutions of the BCC linear programme. As was the case in (18), the process of examining all alternate optima can be complicated and laborious. Again, the reader is directed to Banker *et al* (1996a) for a method that reduces the computational complexity of this task.

As described in section 3.8, an intuitive measure of scale efficiency for a firm is the ratio of its technical efficiency score under CRS to its technical efficiency score under VRS. The scale efficiency score of firm i is defined as:

$$SE_i = \frac{TE_{i(CRS)}}{TE_{i(VRS)}} \Rightarrow TE_{i(CRS)} = TE_{i(VRS)}.SE_i \tag{22}$$

where,

$0 \leq SE_i, TE_{i(CRS)}, TE_{i(VRS)} \leq 1$,

SE_i is the scale efficiency score of firm i,

$TE_{i(CRS)}$ is the technical efficiency score of firm i under CRS, and

$TE_{i(VRS)}$ is the technical efficiency score of firm i under VRS.

This measure of scale efficiency allows the technical efficiency score of a firm under CRS to be split up into the product of its technical efficiency score under VRS and its scale efficiency score. However, this measure of scale efficiency does not identify whether the firm is operating in a region of IRS, CRS or DRS.

This problem is solved for firms operating on the efficient frontier (for global returns to scale) using the criteria specified in (18) or (21). However, these methods are computationally intensive. An alternative method is to calculate the different technical efficiency scores of a firm under different returns to scale model specifications (using the constraints in (17) or (20)). The resulting technical efficiency scores can then be used to classify the returns to scale of a firm operating on the VRS efficient frontier. Let such a firm be firm i. Since firm i is operating on the VRS efficient frontier, $TE_{i(VRS)} = 1$. Three possible cases exist:

1. Firm i exhibits CRS. This occurs if $TE_{i(CRS)} = 1$.
2. Firm i exhibits IRS. This occurs if $TE_{i(NDRS)} = 1$ and $TE_{i(CRS)} < 1$.
3. Firm i exhibits DRS. This occurs if $TE_{i(NIRS)} = 1$ and $TE_{i(CRS)} < 1$.

Alternatively, case 3 occurs when firm i does not exhibit CRS or IRS i.e. when case 1 and case 2 do not apply. This alternative for case 3 avoids the need for running both the NIRS and NDRS models.

As mentioned above, the concept of returns to scale is only well defined for efficient points on the production frontier, with points lying within the interior of the production frontier being more difficult to classify. However, the classification of local returns to scale is also required in order to fully investigate scale. For example, the number of inefficient firms in a dataset may be large relative to the number of efficient firms. If there is no way of classifying local returns to scale, then it will not be possible to determine the return to scale classifications of these inefficient firms.

Determining local returns to scale often involves projecting inefficient firms onto the efficient frontier (the reader is directed to Färe *et al* (1985), Banker *et al* (1996a), and Banker *et al* (1996b) for further

information). Once these projections are completed, the classification of local returns to scale can be performed in the same way as the classification of global returns to scale. However, different methods used to project the inefficient firms onto the efficient frontier may result in different return to scale classifications (Tone, 1996). Furthermore, these projection methods are also computationally intensive. Tone (1996) proposes an alternative, simpler, more computationally efficient solution to the problem of local returns to scale. In this method, the returns to scale of an inefficient firm can be determined directly from its efficiency reference set (ERS). This is done by examining the return to scale classifications of the firms in the ERS. The following relationships between a firm's ERS and its return to scale classification are proved in Tone (1996):

$$All\ firms\ in\ the\ ERS\ have\ CRS \Rightarrow\ CRS$$

$$All\ firms\ in\ the\ ERS\ have\ DRS\ \Rightarrow\ DRS$$

$$The\ ERS\ contains\ a\ mixture\ of\ firms\ with\ CRS\ and\ DRS \Rightarrow\ DRS$$

$$All\ firms\ in\ the\ ERS\ have\ IRS\ \Rightarrow\ IRS$$

$$The\ ERS\ contains\ a\ mixture\ of\ firms\ with\ CRS\ and\ IRS \Rightarrow\ IRS$$

4.5. Input and output variables

The choice of DEA input and output variables is an important decision that can have a significant impact on efficiency estimates (Smith, 1997). Ideally, the set of inputs and outputs should capture the dynamics of the production process (Nguyen & Coelli, 2009). However, the specification of input and output variables may be limited by the dataset, or by the characteristics of the DEA model. The choice of variables may also be influenced by the nature of the problem that is being investigated (Sherman & Zhu, 2006). Importantly, variable specification involves aggregating the underlying data to an appropriate level (Coelli *et al*, 2005). This may include combining inputs, or outputs, in order for the analysis to have sufficient degrees of freedom for meaningful empirical analysis (McCallio, Glass, Jackson, Kerr & McKillops, 2000).

When performing a DEA investigation, model misspecification must be a central concern (Smith, 1997). The choice of input and output variables is an area of potential model misspecification. This can occur in the form of omitted variables or the inclusion of extraneous variables, which may bring the credibility of DEA efficiency estimates into question (Nguyen & Coelli, 2009).

The problem of model misspecification is compounded in DEA because no objective tests are available that can assess the suitability of a particular model specification (Smith, 1997). However, testing the validity of the chosen model specification is of the utmost importance if the results are to be applied to industry – with the aim of increasing the efficiency of firms. Hollingsworth (2008)

discusses the application of DEA to the healthcare environment, and recommends that sensitivity analysis is used to validate the DEA model. Sensitivity analysis cannot guarantee the absence of model misspecification, but it is a useful check on the robustness of the results (Parkin & Hollingsworth, 1997). Industry expertise and management can also help to mitigate the risk of selecting inappropriate input and output variables.

A large number of DEA inputs and outputs can result in an excessive number of firms being ranked as efficient, which reduces the ability of DEA to identify inefficient firms (McCallion *et al*, 2000). However, a large number of input and output variables may be favoured as this may better capture the dynamics of the production process. An increase in the number of variables must therefore be weighed against the reduction in power of DEA.

4.6. Limitations of DEA

Since the DEA model is deterministic, it cannot account for measurement error or random noise (Zere *et al*, 2001). The reader is directed to section 3.11 for a discussion of the limitations of the deterministic nature of DEA.

There is no way to ensure that a firm that is operating on the DEA efficient frontier is in fact efficient when compared to an absolute efficiency standard (Sherman & Zhu, 2006). This is because DEA measures the relative, not absolute, efficiency of a firm. In the case where all firms in the dataset are inefficient, the DEA model will still identify at least one of the firms as being efficient. However, absolute efficiency measures often do not exist, leaving relative efficiency measures as the best available option.

The lack of an absolute efficiency standard also complicates the analysis of efficiency changes over time. For example, the efficiency of a firm relative to the other firms in the dataset may have increased over time, while the absolute efficiency of the firm has decreased. Therefore, the direct comparison of the DEA efficiency scores of a particular firm at two points in time may lead to incorrect conclusions (Sherman & Zhu, 2006). However, techniques have been developed to facilitate these types of comparisons. For example, Malmquist productivity indices can be used to analyse changes in efficiency over time and to decompose these efficiency changes into changes in technical efficiency, changes in scale efficiency, and changes in efficiency due to technological changes (Coelli *et al*, 2005). For further details of efficiency changes over time and Malmquist productivity indices, the reader is directed to Coelli *et al* (2005).

The efficiency scores, or mean efficiency scores, from two different DEA studies cannot be compared with each other (Coelli *et al*, 2005). This is because DEA is a measure of relative efficiency within a

specific dataset. This means that any comparisons with efficiency scores that are not derived from the same dataset may not be meaningful.

If a firm is very different to the other firms, it may be assigned an efficiency score of one as the firm will lie on its own frontier (Valdmanis, 2008). It is therefore important to ensure the firms in the dataset are sufficiently homogeneous. In particular, not accounting for environmental differences between firms may result in misleading efficiency measurements (Coelli *et al*, 2005).

As discussed in section 4.5, the results of the DEA model are sensitive to the selection of inputs and outputs. In particular, the exclusion of important input and output variables may bias the results (Coelli *et al*, 2005). Furthermore, no tests are available to determine the optimal selection of inputs and outputs.

It is also important to ensure that inputs and outputs are homogeneous across firms. Treating heterogeneous inputs and outputs as homogeneous may bias the results (Coelli *et al*, 2005). For example, in a hospital analysis, defining an output variable as the total number of patients treated is likely to result in heterogeneous output variables across hospitals. This is because the output variable (the total number of patients treated) needs to be adjusted by the severity of the treated illness before it can be compared across hospitals. Differences in the quality of outputs produced by different firms also introduce heterogeneity into output variables. Furthermore, it is very difficult to measure and control for the heterogeneity of outputs across firms (Linna, 1998). This means that unmeasured output heterogeneity may impact the reliability of DEA efficiency estimates. However, it should be noted that the difficulties associated with heterogeneity are not unique to DEA and also apply to other non-parametric and parametric efficiency measurement techniques.

The number of efficient firms on the production frontier tends to increase with the number of inputs and output variables (Coelli *et al*, 2005). Therefore, the power of DEA reduces when a large number of inputs and outputs are specified. This may be a significant limitation when trying to model a complex process with many inputs and outputs. Similarly, the power of DEA decreases when the dataset is small (Sherman & Zhu, 2006).

Finally, the results of all models depend on good quality, appropriate, credible data. DEA is no exception.

4.7. Practical applications of DEA in the hospital industry

There is a large body of literature focusing on the application of DEA and frontier methods to hospital and healthcare services. For example, Hollingsworth (2003) reviews 188 published papers that focus on the measurement of healthcare efficiency.

At the time of writing this paper, Zere *et al* (2001) and Kibambe & Koch (2007) were the major studies that investigate scale efficiency within South African hospitals. These studies analyse the efficiency of public sector hospitals. Therefore the results of these studies cannot be compared directly with the results of this paper because of the significant differences that exist between the public and private healthcare sectors. In particular, significant differences include: objectives (social and for-profit objectives); degree of competition; sources of financing; available resources including the availability of expensive medical technologies; attractiveness to human resources; case-mix profiles which differ with the socio-economic status of patients; quality of data systems and consequent availability and reliability of data. Some of the salient points of Zere *et al* (2001) and Kibambe & Koch (2007) are outlined in sections 4.7.2 and 4.7.3 respectively.

The specification of input and output variables is a key methodological decision of DEA. Drawing on the available literature, the next section examines the specification of DEA input and output variables for studies involving the hospital industry.

4.7.1. Input and output variables relating to the hospital industry

In the DEA literature, there are many examples of input and output variables specifications relating to the hospital industry. For example, Nguyen & Coelli (1996) investigate the modelling choices of 95 hospital efficiency studies. For each of these studies, they provide a list of the input and output variables. Worthington (2004) reviews 38 studies focusing on efficiency measurement within healthcare services. For each of these studies, the input and output variables specifications are provided. Sherman & Zhu (2006) also provide many examples of input and output variables relating to healthcare applications of DEA.

Input variables should reflect the factors of production, namely capital, labour and production materials (Nguyen & Coelli, 2009). Hospital labour inputs include doctors, specialists, nurses, administrative and operational staff. These categories are usually aggregated into two main labour inputs, doctors and nurses, using weights to reflect the seniority of staff (Nguyen & Coelli, 2009). Materials used in the hospital production process, such as pharmaceuticals and consumable items, are usually measured by the cost of these inputs (Nguyen & Coelli, 2009). The theoretically correct measure of capital is the flow of capital services over the period rather than capital stock (Worthington, 2004). The flow of capital services is the amount of capital consumed over the

production period, while the capital stock is the total quantity of infrastructure and equipment on hand during the production process (Nguyen & Coelli, 2009). The flow of capital services could be estimated by the market rental price of the capital for the particular production period (Barro, 2008). However, estimating the flow of capital or even measuring the existing capital stock is a challenging task. The majority of DEA studies focusing on hospital efficiency therefore use total number of beds as a proxy for the flow of capital services (Nguyen & Coelli, 2009). The use of total number of beds as a proxy for the flow of capital services may be a reasonable assumption because the total number of beds in a hospital is typically correlated with the stock of other capital infrastructure and equipment (Clement, Vivian, Vladmanis, Bazzoli, Zhao & Chukmaitov, 2008). Alternatively, the value of property, plant and equipment sourced from the hospital's financial accounts could be used as a proxy for the flow of capital services (Clement *et al*, 2008). However, these figures will also be subject to measurement difficulties and assumptions.

The following are typical examples of the sets of input variables that have been used in various DEA studies:

- Clement *et al* (2008) specified four input variables: full time equivalent (FTE) registered nurses, FTE licensed practical nurses, other FTEs, and licensed staff beds.
- Kibambe & Koch (2007) specified three input variables: medical doctors and specialists, active beds, and nurses.
- Parkin & Hollingsworth (1997) specified six input variables: staffed beds, number of nurses, number of non-nursing medical and dental staff, other staff, the cost of the drug supply for the hospital, and the hospital's capital charge.
- Zere *et al* (2001) specified two input variables: recurrent expenditure, and beds.

All of the above studies include an input variable that represents the flow of capital services, namely number of beds. Labour is usually included directly in terms of number of medical professionals or their FTEs; however the study by Zere *et al* (2001) combines labour and materials into a single input variable, namely recurrent expenditure. The studies by Clement *et al* (2008) and Kibambe & Koch (2007) do not include input variables that represent the materials used in the production process. For this paper's investigation, the intention is to select input variables that are most representative of the factors of production. However, the set of variables selected for any investigation is highly dependent on the available dataset.

Grosskopf & Valdmanis (1987) explain that it is difficult to define and accurately measure real hospital output. Theoretically, hospital output should be defined as the change in health status of a patient due to receiving hospital treatments (Grosskopf & Valdmanis, 1987). This is echoed by Nguyen & Coelli (2009) who define the theoretical measure of hospital output as the difference between the health status of a patient with treatment and without treatment. In practice, this measure

will not be available and hospital output must be measured by proxies that are assumed to be related to improved health status (Grosskopf & Valdmanis, 1987). Common proxies include the number of cases, number of inpatient days, and the number of outpatient cases (Worthington, 2004). However, cases are not homogeneous and differ significantly by severity and therefore resource intensiveness (McCallion *et al*, 2000). Without accounting for this heterogeneity, all else equal, a hospital with a more severe case-mix will be penalised and appear relatively less efficient than a hospital with a less severe case-mix. Case heterogeneity can be dealt with by splitting the cases into sufficiently homogeneous groups based on gender, age or broad clinical groupings, such as surgeries, emergency room visits, referrals to specialists, laboratory work, etc. (Sherman & Zhu, 2006). These groups should be sufficiently similar to make efficiency comparisons across hospitals meaningful. Alternatively, case heterogeneity can be accounted for by adjusting the number of cases by a factor that represents the severity of the underlying case-mix (Clement *et al*, 2008). Case-mix adjustment factors can be calculated by weighting each of the underlying cases by the ratio of the average cost of that particular type of case to the average cost of all cases (McCallion *et al*, 2000). Diagnostic related groups[2] (DRGs) can be used to define fairly homogeneous groupings of cases for the purposes of calculating case-mix adjustment factors (Nguyen & Coelli, 2009). After case-mix adjustment, this method facilitates the aggregation of a large number of homogeneous groups of cases (for example, hundreds of DRGs) into one or two output variables (Nguyen & Coelli, 2009). Together with case outputs, quality of care is also not uniform across hospitals (Zuckerman, Hadley & Iezzoni, 1994). Quality of care is a complex, multidimensional issue that needs to be considered in order to perform a comprehensive hospital efficiency analysis (Sherman & Zhu, 2006). If quality is not accounted for, a hospital will be penalised when it uses more resources to produce higher quality health outcomes, even if this is done in an efficient manner. However, higher quality does not necessarily mean greater resource utilisation (Clement *et al*, 2008). Quality for a specific case could be measured by comparing the actual health improvements of a patient due to treatment against the expected best practice health improvements (McCallion *et al*, 2000). However, it is questionable whether any quality measure is comprehensive or even meaningful (Sherman & Zhu, 2006). Eckermann & Coelli (2008) investigate various methods for accounting for quality: quality as an input variable, quality as a good output, quality as a bad output, quality as an exogenous factor, and the option to ignore quality. Other methods have been developed to account for quality. The reader is directed to Sherman & Zhu (2006) for further details. McCallion *et al* (2000) argues that DEA can still provide meaningful insights for policymakers and hospital management, even if quality of care is not adequately accounted for.

[2]Diagnostic related groups (DRGs) form part of a hospital case classification system whereby cases are classified into a limited number of groups based on certain common characteristics. The reader is directed to Scheller-Kreinsen, Geissler & Busse (2009) for further details regarding DRGs.

The following are typical examples of the output variables that have been used in various DEA studies:

- Clement *et al* (2008) specified five output variables: number of births, outpatient surgeries, emergency room visits, outpatient visits, and case-mix adjusted admissions.
- Kibambe & Koch (2007) specified four output variables: outpatient visits, total admissions, inpatient days, and theatre cases / surgeries.
- Parkin & Hollingsworth (1997) specified six output variables: acute discharges (medical), acute discharges (surgical), accident and emergency attendances, outpatient attendances, obstetrics and gynaecology discharges, and other speciality discharges.
- Zere *et al* (2001) specified two output variables: outpatient visits, and inpatient days.

All of the above studies attempt to include output variables that capture the health outcomes produced by each hospital. This is done by including variables that represent the number of cases treated by each hospital, sub-divided into homogeneous groups. Note that the power of DEA to identify inefficient hospitals decreases as the number of case sub-divisions increases. Of the above studies, only Clement *et al* (2008) attempt to directly adjust for case-mix differences across hospitals, which was done in order to better reflect the actual health outcomes produced by each hospital. By adjusting for case-mix, a hospital with a more severe, resource-intensive case-mix is not penalised in terms of efficiency when it is compared to a hospital with a less severe, less resource-intensive case-mix. For this paper's investigation, the intention is to select output variables that are most representative of the health outcomes produced by each hospital. In particular, this will involve adjusting cases to reflect differences in case-mix severity across hospitals. However, the set of variables selected for any investigation is highly dependent on the available dataset.

It is possible that a variable can be classified as either an input or an output (Cook & Zhu, 2007). This usually occurs when a resource can also represent a production output of the hospital, such as medical interns, nurse trainees or research funding. The intermediate nature of health outcomes can also lead to input-output ambiguity. For example, the number of days that a patient spends in hospital could be interpreted as an input into the production of health outcomes. However, it could also be interpreted as a proxy for health outcomes and therefore as an output of the production process, the latter being the usual interpretation (Nguyen & Coelli, 1996). The reader is directed to Cook & Zhu (2007) for more information on how to account for these types of variables within DEA models.

Hollingsworth, Dawson & Maniadakis (1999) claim that, due to the complications involved in modelling healthcare services, it is likely that most studies of healthcare efficiency suffer from bias caused by omitted variables. Smith (1997) has shown that the omission of important variables appears to have a large impact on efficiency estimates, and that this impact reduces with increasing complexity of the production process. If this is indeed the case, then the omission of variables when

modelling a complex hospital production process may be less of an issue than when modelling a simpler production process. Smith (1997) also found that the inclusion of extraneous variables appears to have a small impact on efficiency estimates, regardless of the complexity of the production process. In practice this means that, when in doubt, a variable should be included rather than excluded, particularly for a simple production process or when there is a low correlation between inputs. Again, the inclusion of additional variables must be weighed against the reduced power of DEA to identify inefficient hospitals (McCallion *et al*, 2000). Nguyen & Coelli (2009) and Smith (1997) note that, due to the lack of quality data within the healthcare environment, the omission of variables is much more likely than the inclusion of extraneous variables.

4.7.2. The study by Zere, McIntyre & Addison (2001)

The objectives of the study by Zere *et al* (2001) were to examine the technical efficiency and productivity of South African public hospitals. This was done by analysing the technical efficiency, and the drivers of inefficiency, of a sample of 86 public hospitals situated in the Northern, Eastern and Western Cape over the 1992/93 year. The changes in productivity over the period between 1992/93 and 1996/97 were then investigated for a sample of Western Cape hospitals. The study was limited by poor availability and quality of the data. An input-orientated, two-input, two-output DEA model was used to calculate the technical efficiency scores of the hospitals. The two inputs were: number of beds (which was used as a proxy for capital) and recurrent expenditure (which was used as a proxy for labour and production materials). The two outputs were: outpatient visits and inpatient days. Zere *et al* (2001) found that only 13% of the hospitals were technically efficient, with an average inefficiency lying between 35% and 47%. Zere *et al* (2001) also found that smaller hospitals were relatively more scale efficient than larger hospitals. Approximately 50% of the hospitals exhibited DRS, 13% exhibited CRS, and 37% exhibited IRS. Zere *et al* (2001) go on to explain that addressing scale inefficiencies is a complex issue that involves practical difficulties, as well as the consideration of both the demand and the supply of healthcare services. The possible drivers of inefficiency were identified using a Tobit model. Zere *et al* (2001) found that low occupancy rates as well as a high proportion of inpatient days to outpatient days had a negative impact on efficiency. Changes in productivity over time were examined using the Malmquist productivity index. The results showed that total factor productivity had decreased over time, mainly due to technical regress. Zere *et al* (2001) conclude that improving the technical efficiency of the inefficient hospitals would result in an immense savings of all key resources. Efficiency improvements could potentially generate recurrent expenditure savings of between 26% and 33%. Additionally, the number of beds could potentially be reduced by 30% to 39%. These savings could provide resources that are necessary to extend and improve the quality of South Africa's primary healthcare services.

4.7.3. The study by Kibambe & Koch (2007)

Kibambe & Koch (2007) applied DEA to a sample of public hospitals situated in Gauteng. An incomplete sample of approximately 50% of Gauteng public hospitals was used in the analysis. The dataset included 14 hospitals and spanned the period from 1999 to 2004; however not all hospitals provided data for all of these years. The major limitation of the study was the poor availability and quality of the data, to the extent that the data may not be sufficiently representative of the production process or the public hospital population (Kibambe & Koch, 2007). The difficulty obtaining data was ascribed to reluctant participation in the study as well as inadequate information systems. Kibambe & Koch (2007) express concern over the lack of South African public hospital data, and the impact of this on research and policy. In order to apply DEA to the dataset, three input and four output variables were specified. The three inputs were: medical doctors and specialists, active beds, and nurses. The four outputs were: outpatient visits, total admissions, inpatient days, and theatre cases / surgeries. Efficiency scores were then calculated for numerous input-output combinations. Kibambe & Koch (2007) found that average technical efficiency, calculated under the assumption of a CRS production technology, ranged from 70% to 90%; while the average technical efficiency, calculated under the assumption of a VRS production technology, ranged from 83% to 99%. The average technical efficiency was calculated for various multiple-input, multiple-output models. In each of these models, the number of hospitals included in the analysis was subject to data availability. It was observed that a greater number of public hospitals in Gauteng exhibit DRS rather than IRS. Kibambe & Koch (2007) explain that this may be a result of hospitals having too few medical professionals per bed. Kibambe & Koch (2007) also suggest that small hospitals and hospitals offering more technical services, such as surgeries, appear to deliver healthcare services in a more efficient manner.

4.7.4. The need for further research

Zere *et al* (2001) note that the measurement of hospital efficiency within Sub-Saharan Africa is an under-researched area of investigation. As a result, there is very little credible information regarding the nature and extent of the inefficiency present within these healthcare systems. This paper attempts to contribute to the limited body of research by using DEA to investigate the relationship between scale and efficiency within the South African private hospital environment.

5. Data and methodology

5.1. Data

The discussion in chapter 2 highlighted the importance of using healthcare resources efficiently, in both the public and private sectors. This paper focuses on examining efficiency and scale within the private hospital industry, rather than the public sector. The primary reason for examining private hospitals is due to data limitations in the public sector. This is well documented in the study performed by Kibambe & Koch (2007). However, given an appropriate dataset, it is envisioned that the methodology adopted in this paper could be applied to the public sector.

The dataset underlying this investigation was sourced from a large private hospital company operating within the South African healthcare environment. Although it cannot be said that the dataset is representative of the South African private hospital industry, the dataset is large and the results derived from the dataset may provide useful insight into the operations and efficiency of the private hospital industry. At the very least, the results will highlight some issues that should be considered when investigating the relationship between hospital scale and efficiency.

Data were received for 52 hospitals administered by the private hospital company. For each hospital, the dataset contained case information, human resources information, and information regarding the operations of the hospital (for example, number of beds) for the three year period from 2007 to 2009. The dataset can be described as being of good quality, barring the limitations described in section 5.2. The human resources data contained payroll information, which included number of staff, their employment titles, and their salaries. The case data consisted of an exhaustive list of cases for each hospital spanning 2007 to 2009. Details of each case were provided in a number of fields. Fields of interest included patient age, date of birth, date of admission, number of calendar days spent in hospital, number of billed days spent in hospital, hospital billed amount, pharmacy billed amount, surgical theatre minutes, an admission category grouping and various case-mix indicators (for example, diagnostic related group (DRG) codes and basic diagnostic related groups (BDRG) codes).

5.2. Limitations of the data

Various adjustments were made to the dataset during the data cleaning process. At a hospital level, five hospitals had incomplete datasets due to a significant number of missing records. In some cases, the datasets did not contain records for all three years. For consistency, these hospitals were excluded from the analysis. It was also necessary to exclude these hospitals in order to facilitate efficiency comparisons, and to avoid misrepresenting the efficiency of these hospitals. Three hospitals operated outside of South Africa, and were therefore excluded from the analysis. Additionally, a day hospital

was excluded as this hospital was atypical when compared with the rest of the hospitals in the dataset. Furthermore, data from two hospitals were merged as these hospitals operated under the same management. After all adjustments, the final dataset consisted of 42 hospitals.

At a case level, cases with zero and negative billed amounts were excluded. Some cases did not result in hospital admissions – such as trauma cases, pharmacy only cases, and cases involving the use of catheterisation laboratories. These cases are referred to as *partial accounts* and are atypical cases, often having small or volatile billed amounts. For these reasons, partial accounts were excluded from the analysis. Cases used purely for management controls were also excluded. Various spot checks were then performed to ensure the consistency of the excluded cases. Using zero billed amount cases as an example, it was checked that these cases as a percentage of total cases remained fairly stable across years for a particular hospital. Although these checks were not performed for all excluded cases for all hospital across all years, the spot checks did provide some comfort that these cases were appropriately and consistently excluded. Some cases in the unadjusted dataset had unreasonably small billed amounts. However, through the implementation of the above adjustments these small cases were removed. This provided further comfort in the above data cleaning process. Cases with large billed amounts were not removed as such cases do occur in the normal course of hospital operations. Additionally, the removal of these cases may distort efficiency comparisons across hospitals.

Although the data contained a sufficient level of detail, the analysis could have benefited from a greater level of detail. For example, case-specific details of expenditure on production materials, in addition to pharmaceutical expenditure, would have provided a more accurate description of the inputs used in the production process. In the same vein, further details regarding capital inputs, such as the amount of equipment utilised by each hospital, would have facilitated a better estimation the flow of capital services for each hospital. Additionally, the analysis would have benefited from further details regarding the quality of healthcare outputs. These data, if available for this investigation, would have provided a more comprehensive description of the hospital production process.

5.3. Overview of the methodology

The remainder of this chapter discusses the methodological decisions of the investigation. DEA was selected as the efficiency measurement technique, which forms the methodological core of the analysis. The reader is directed to section 3.14 for an explanation of why DEA is the most appropriate efficiency measurement technique for this investigation. In more detail, an input-orientated DEA model was used in the analysis. A discussion as to why input-orientated models are more appropriate than output-orientated models for the analysis of hospital efficiency is provided in section 4.2.

Three DEA models were then specified and applied to the data. This involved specifying possible input and output variables, and selecting different combinations of these variables in order to specify complete DEA models. This process is discussed in sections 5.4 and 5.5. The software used for the analysis and outputs of the analysis are then discussed in section 5.6. This chapter concludes with a discussion of the limitations of the methodology.

5.4. Input and output variables

The reader is directed to sections 4.5 and 4.7.1 for detailed discussions of the specification of DEA input and output variables. The input and output variables for this investigation were chosen so as to be representative of the hospital production process. In order to derive these variables from the dataset, it was necessary to perform certain adjustments and aggregations. Proxies were used where the dataset did not allow direct measurement of a particular variable. The specification of the input and output variables used in this investigation is discussed in detail below.

After adjustments and aggregation of the data, the final set of relevant output variables include:

- Total number of billed days spent in hospital,
- Total number of theatre minutes,
- DRG case-mix adjusted total number of cases, and
- BDRG case-mix adjusted total number of cases.

Total number of billed days spent in hospital, total number of theatre minutes, and total number of cases all represent outputs that will impact health outcomes. They are therefore proxies for health outcomes produced by a hospital. Each of these outputs has been aggregated into a single variable for each hospital.

However, the total number of cases cannot be used directly as it does not account for case-mix differences between hospitals. For example, a hospital that treats many resource intensive, severe cases may appear less efficient than a hospital that treats relatively less resource intensive cases. Note that case-mix severity is already reflected, at least partially, in the total number of billed days spent in hospital and the total number of theatre minutes. The total number of cases per hospital therefore needs to be adjusted to reflect the case-mix of that hospital. Diagnostic related groupings (DRGs) of cases can provide information regarding the case-mix of each hospital. It was decided to calculate a case-mix adjustment factor for each hospital, which would adjust the total number of cases to reflect the case-mix differences between hospitals. For this investigation 1,000 different DRGs were used to calculate each hospital's adjustment factor. For each DRG, the average billed amount per case across all hospitals was calculated. These figures were then normalised using the average billed amount per case across all hospitals and DRGs. The resulting ratio for each DRG represents the severity of that

DRG relative to the other DRGs. The case-mix adjustment factor for each hospital was then calculated as the weighted average of the total number of cases in each DRG for that particular hospital, weighted by the DRG specific severity ratio. These case-mix adjustment factors were calculated for each of the three years. For comparative purposes, this process was repeated using basic diagnostic related groupings (BDRGs) instead of DRGs.

The case-mix adjustment factors were then multiplied by the total number of cases for each hospital. Where a hospital treated a greater number of severe cases, relative to the average severity of cases, the case-mix adjustment increased the total number of cases. Similarly, when a hospital treated a lower number of severe cases, the case-mix adjustment decreased the total number of cases. Adjusting total number of cases by case-mix provides a more appropriate measure of the actual number of healthcare outputs produced by a hospital. It also facilitates meaningful comparisons of healthcare outputs across hospitals. The DRG case-mix adjustment factors and number of operational beds for each hospital in 2009 are displayed in Figure 5 below.

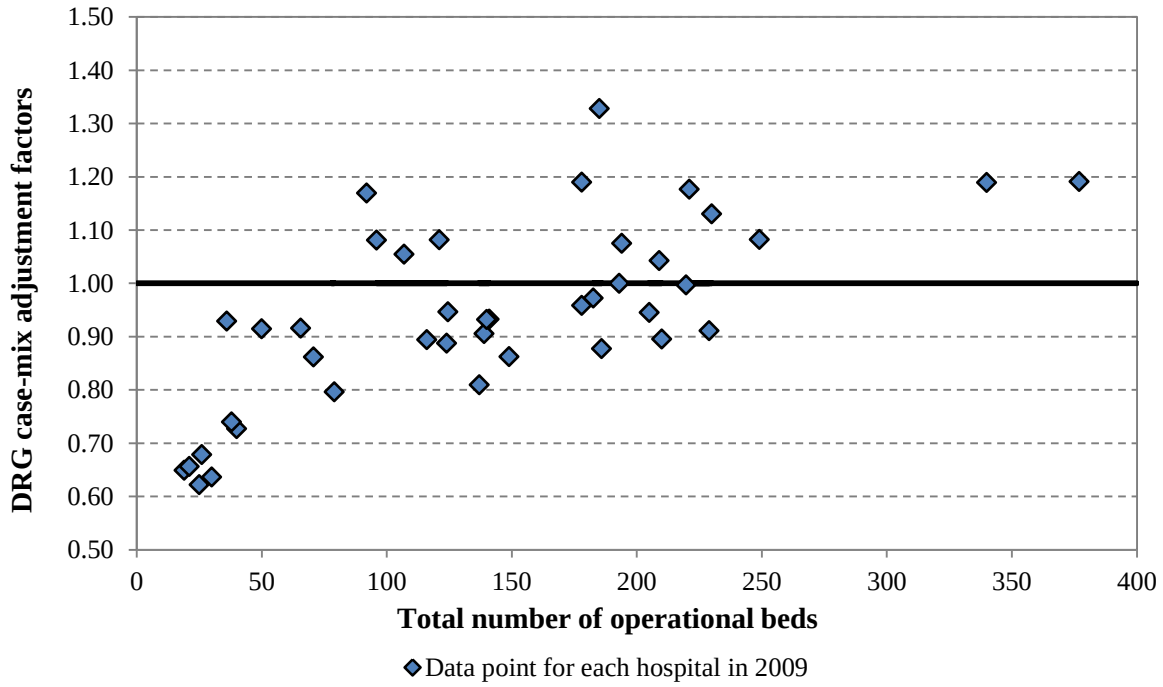

Figure 5: DRG case-mix adjustment factors and number of beds for each hospital in 2009

Using number of beds as a proxy for size, it can be seen in Figure 5 that smaller hospitals tend to have lower DRG case-mix adjustment factors than larger hospitals. This relationship was also observed in 2007 and 2008. One interpretation is that smaller hospitals tend to treat a higher concentration of simpler, less resource intensive cases; while larger hospitals trend to treat a higher concentration of more complex, resource intensive cases. This could occur, for example, because complex cases are referred by smaller hospitals to larger hospitals where the necessary resources and specialist facilities are available to treat these cases.

In Figure 5 it can be seen that smaller hospitals tend to have DRG case-mix adjustment factors that are less than one. For these smaller hospitals, the adjustment factors decrease the number of cases in order to arrive at the case-mix adjusted number of cases. Similarly, some of the larger hospitals have DRG case-mix adjustment factors that are greater than one. These adjustments could potentially amplify any scale differences that exist between smaller and larger hospitals. However, the case-mix adjustments are necessary in order to better reflect the actual healthcare outputs produced by a hospital, relative to the other hospitals in the dataset. This results in the output variable, case-mix adjusted number of cases, being more consistent with the inputs used to produce this output. For these reasons, the DRG case-mix adjustment used in this investigation may in fact result in a more appropriate representation of hospital scale.

The correlations between the four output variables for 2009 are provided in Table 1 below.

Table 1: Correlations between the four output variables in 2009

2009	Billed days	Theatre minutes	DRG adjusted cases	BDRG adjusted cases
Billed days	1			
Theatre minutes	0.897	1		
DRG adjusted cases	0.960	0.956	1	
BDRG adjusted cases	0.957	0.959	0.999	1

The correlation between DRG and BDRG case-mix adjusted number of cases is very high (0.997 in 2007, 0.999 in 2008 and 0.999 in 2009). This means that the impact of using BDRG and DRG groupings to adjust for differences in case-mix across hospitals was expected to be small. This small impact was confirmed by quantifying the actual impact on the results of using BDRG instead of DRG to adjust for case-mix. Note that this was confirmed for each of the three years. Based on the above, it was decided to exclude BDRG case-mix adjusted number of cases and focus on DRG case-mix adjusted number of cases. The DRG adjustment was chosen as it is calculated at a more granular level.

The remaining three output variables are all highly correlated. This is as expected because each is an output of the same production process. As such, it can be expected that each output from the same hospital would be proportional in some way to the other outputs, resulting in high correlations between these outputs. It was verified that the correlations remained fairly stable across the three

years. This is as expected because the relationship between outputs is not expected to change dramatically within a three year period.

Now consider the set of input variables. After adjustments and aggregation of the data, the final set of relevant input variables include:

- Total number of operational beds,
- Salary adjusted number of nurses, and
- Total billed pharmacy amount.

The three inputs were used as proxies for the key economic inputs of capital, labour and production materials.

Total number of operational beds was used as a proxy for the flow of capital services over the year. Measures of the total amount of infrastructure and equipment used in the production process were not available. If these measures were available, then they could have been used to improve the proxy measurement for the flow of capital services. However, it is likely that the total number of operational beds is highly correlated with the total amount of infrastructure and equipment used in the production process. Under this assumption, the number of operational beds is a reasonable proxy for the flow of capital services. The reader is directed to section 4.7.2 for further details of this.

Salary adjusted number of nurses was used as a proxy for labour. This was derived using Human Resource data for each hospital and year. Note that, by law, doctors and specialists cannot be employed by hospitals and therefore information regarding doctors and specialists did not form part of the data made available for this investigation. The data did however provide details of the nurses employed at each hospital at an individual level. This included the seniority and salary of each nurse. Nurses form a large part of each hospital's human resource costs and are therefore an appropriate basis for the derivation of a proxy for labour. Furthermore, many studies of hospital efficiency include nurses as an input variable – the reader is directed to section 4.7.2 for further details of this. Using the data, the nurses employed at each hospital were subdivided by seniority. Nurses' seniority categories included: enrolled nurses, nursing auxiliaries, professional nurses, pupil nursing assistants, pupil enrolled nurses, nurses, senior enrolled nurses, senior professional nurses and senior registered nurses. The relatively large number of categories meant that it was not feasible to have each category as a separate input. Therefore it was necessary to aggregate the number of nurses in each category into a single measure of nursing input. The average salary for each nursing category was then calculated. These figures were used to calculate an average salary ratio for each nursing category. This was done by normalising the average salary for each category by the average salary across all nursing categories. The resulting ratio for each category represents the seniority of that nursing category relative to the other nursing categories. The salary adjusted number of nurses was then calculated, for

each hospital, by weighting the number of nurses in each category by the category specific average salary ratio. Salary adjusted total number of nurses provides an appropriate measure of the actual number of nursing inputs used by a hospital. This approach also facilitates meaningful comparisons of these inputs across hospitals and years. Importantly, this approach controls for geographical differences between nurses' salaries which, if not controlled for, could distort efficiency measurements.

Total billed pharmacy amount was used as a proxy for the materials used in the production of health outcomes. The total amount spent on pharmaceuticals forms a large part of each hospital's operational costs and is therefore an appropriate input into this process. Furthermore, this was the only data made available for the investigation that could be used to derive a proxy for materials. After consideration, the total monetary amount of the billed pharmacy was specified as the input variable. This was done under the assumption that each hospital is subject to the same pharmacy inflation in one particular year.

The correlations between the three input variables for 2009 are provided in Table 2 below.

Table 2: Correlations between the three input variables in 2009

2009	Number of beds	Salary adjusted nurses	Billed pharmacy amount
Number of beds	1		
Salary adjusted Nurses	0.971	1	
Billed pharmacy amount	0.927	0.953	1

The three relevant input variables are all highly correlated. This is as expected because each is an input of the same production process. As such, it can be expected that each input from the same hospital would be proportional in some way to the other inputs, resulting in high correlations between these inputs. It was verified that these correlations remained fairly stable across the three years. This is as expected because the relationship between inputs is not expected to change dramatically within a three year period.

5.5. Model specification

In order to facilitate returns to scale classification, the BCC envelopment (16) and multiplier (19) models were selected for the analysis. Through different combinations of the available three input and three output variables, it is possible to specify 49 different DEA models. However, not all of these 49 models will produce results that are relevant to the analysis of the relationship between scale and efficiency. After consideration, three models were specified – each with a particular purpose in mind. These three models are discussed in this section.

Note that the models are named using a convention which reflects the number of input and output variables of that particular model. For example, model 3x1y is a DEA model with three input variables (three x-variables) and one output variable (one y-variable). The following three models were specified:

1. Model 3x1y,
2. Model 3x3y, and
3. Model 1x3y.

Examining the results of multiple models can be considered a form of sensitivity testing of the model specifications. Furthermore, it facilitates comparison of the consistency of the results across different models.

Model 3x1y is specified by three input variables:

1. Total number of operational beds,
2. Salary adjusted number of nurses, and
3. Total billed pharmacy amount;

and one output variable:

1. DRG case-mix adjusted total number of cases.

Given the available data, it was expected that this model is most representative of the hospital production process. The three inputs can be viewed as a proxy for the key economic inputs of capital, labour and production materials; and the output as the best available proxy for the health improvements produced by hospital services.

Two output variables (total number of billed days spent in hospital, and total number of theatre minutes) were excluded from this model. These variables were excluded because they are partially accounted for by the case-mix adjustment applied to the total number of cases. Therefore the inclusion of these additional two output variables could lead to double counting of outputs. This is elaborated on in the paragraph below.

The additional two output variables, total number of billed days spent in hospital and total number of theatre minutes, already account for differences in case-mix. This is because a more severe case will, generally, require more days in hospital or more time in theatre. The DRG case-mix adjusted total number of cases accounts for case-mix difference across all cases, including those cases that require a greater number of days spent in hospital or more time in theatre. This means that, by including total number of billed days spent in hospital and total number of theatre minutes, it is likely that the impact of cases involving a greater number of days spent in hospital or more time in theatre will be double counted. Therefore the single output of DRG case-mix adjusted total number of cases appears to best capture the health outcomes produced by a hospital.

However, the exclusion of these two variables will lead to a loss of information regarding the production process. Including these variables will therefore result in a trade-off between a fuller description of the hospital production process and double counting some cases. It can also be argued that the time spent in theatre is a highly resource intensive activity and should be included in the model. However, this will be partially captured by the DRG case-mix adjusted total number of cases. Smith (1997) proposes that, when faced with a model specification decision, variables should be included rather than excluded. The reader is directed to section 4.7.1 for further detail of this. For these reasons, all three output variables were included in the 3x3y model.

Model 3x3y is specified by three input variables:

1. Total number of operational beds,
2. Salary adjusted number of nurses, and
3. Total billed pharmacy amount;

and three output variables:

1. DRG case-mix adjusted total number of cases,
2. Total number of billed days spent in hospital, and
3. Total number of theatre minutes.

As with model 3x1y, the three inputs can be viewed as a proxy for the key economic inputs of capital, labour and production materials. All three output variables were included in order to capture the greatest quantity of available output information, but at the cost of double counting some outputs. Note that, relative to the other models, the greater number of variables in the 3x3y model is likely to reduce the power of the DEA model to identify inefficient hospitals (McCallion *et al*, 2000). This property of DEA was discussed in section 4.5.

Model 1x3y is specified by one input variable:

1. Total number of operational beds;

and three output variables:

1. DRG case-mix adjusted total number of cases,
2. Total number of billed days spent in hospital, and
3. Total number of theatre minutes.

The 1x3y model only has one input variable, number of operational beds. This model was used to examine the relationship between operational beds and returns to scale and to compare these results with the results of the multiple-input models. As with model 3x3y, all three output variables were included in order to capture the greatest quantity of available output information, but at the cost of double counting some outputs.

The above three models were chosen with the intention of providing insight into the relationship between scale and efficiency within the South African private hospital environment. This will be achieved through analysing and comparing the results of these three models – each of which represent a different variation of the same hospital production process.

5.6. Software and model outputs

Various DEA software packages were considered for this investigation. It was decided that R (version 2.15.2), developed by the R Foundation for Statistical Computing (2012), could provide a comprehensive analysis of the data. In particular, the Benchmarking package in R, developed by Bogetoft & Otto (2011), provided the necessary DEA tools. Additionally, Microsoft Excel 2010 and Microsoft Visual Basic (version 7.0) were used for the data analysis.

For each of the three DEA models specified in section 5.5, these software packages were used to output the following for each hospital and each year from 2007 to 2009:

- The technical efficiency scores under VRS, IRS, CRS, and DRS,
- The scale efficiency scores,
- The global returns to scale,
- The local returns to scale,
- The lambda weights ($\boldsymbol{\lambda}$ vectors), and
- The potential input savings.

These results are presented and discussed in chapter 6.

5.7. Limitations of the methodology

The use of DEA as an efficiency measurement technique is a core methodological assumption of this investigation. The reader is directed to section 4.6 for a discussion of the limitations of DEA. In addition to the limitations of DEA, the methodology adopted in this investigation has a number of further limitations.

The DEA methodology makes use of a set of defined input and output variables for each hospital. At a variable specification level, each variable may not be fully representative of the input, or output, that it is trying to capture. For example, the number of hospital beds may not be a good proxy for the flow of capital services. This is discussed in further detail in sections 4.7.1 and 5.4. Furthermore, the degree of aggregation of the input and output data into input and output variables that can be used in the DEA model may cause a loss of information and, consequently, bias the results of the analysis. For example, in this analysis, a large volume of individual patient data has been aggregated into a few variables with the intention that these variables are representative of the hospital's production process. These aggregated variables include DRG case-mix adjusted total number of cases, total number of billed days spent in hospital, total number of theatre minutes, and total billed pharmacy amount. It is possible that these aggregated variables do not adequately reflect the differences in the underlying individual patient data.

At a model specification level, the set of input and output variables selected for a particular DEA model may not fully capture the production process. For example, in this analysis, model 3x1y does not include total number of theatre minutes as a direct output. Conclusions drawn from the results of this model may be biased if total number of theatre minutes is an essential component of the production process that is not adequately captured by the output variable of this model, namely DRG case-mix adjusted number of cases. The potential for model misspecification is therefore a methodological limitation. However, it should be noted that the potential for model misspecification is not unique to DEA and also applies to other non-parametric and parametric efficiency measurement techniques. Model misspecification is discussed further in in section 4.5.

This methodology accounts for case-mix differences between hospitals by using a simple DRG case-mix adjustment. This adjustment could potentially be improved, for example, by accounting for the age and gender of patients. However, cases grouped by DRG are relatively homogeneous and should adequately reflect the differences in case-mix across hospitals. Additionally, the scope for further case-mix adjustments was limited by the available dataset. The reader is directed to section 4.7.1 for a more detailed discussion of case-mix adjustments. In this methodology, quality of care was assumed to be consistent across hospitals and was not explicitly accounted for. This is because each hospital forms part of the same organisation and is therefore subject to the same set of quality standards and

procedures. Additionally, the explicit modelling of quality presents many challenges, mainly due to dataset limitations but also due to the difficulties associated with accounting for quality (the reader is directed to section 4.7.1 for a discussion of these difficulties). As a consequence of not accounting for all case-mix and quality differences, it is possible that some of the efficiency estimates in this investigation could be biased.

6. Results and discussion of results

6.1. Overview of the results

This chapter presents and discusses the results of the investigation. Section 6.2 provides an in-depth analysis of the results of the 3x1y model. This is followed in section 6.3 by a comparison of the results of the 3x1y model with the results of the 3x3y and 3x1y models. Five hospitals from the 3x1y model were then selected and analysed due to their interesting characteristics. The details of this analysis are provided in section 6.4. The remainder of the current section discusses some of the considerations that must be taken into account when interpreting the results of this investigation. In particular, the discussion focuses on the interpretation of scale efficiency scores and the comparison of efficiency scores over time.

It is useful to note that a change in scale efficiency can be explained in terms of the change in technical efficiency calculated under the assumption of a CRS production technology (TE_{CRS}), and the change in technical efficiency calculated under the assumption of a VRS production technology (TE_{VRS}). From the definition of scale efficiency as specified in equation (6) of section 3.8, it can be seen that an increase in the scale efficiency of a hospital may be the result of one of the following:

- TE_{CRS} of the hospital increases and TE_{VRS} remains constant,
- TE_{CRS} of the hospital increases by more than TE_{VRS} increases,
- TE_{CRS} of the hospital increases and TE_{VRS} decreases,
- TE_{CRS} of the hospital remains constant and TE_{VRS} decreases, or
- TE_{CRS} of the hospital decreases by less than TE_{VRS} decreases.

Changes in a hospital's scale efficiency should be thought of within this context. However, caution must be applied when comparing efficiency results across years. This is because DEA measures the relative, not absolute, efficiency of each hospital within a particular year (Sherman & Zhu, 2006). For example, the efficiency scores of a hospital relative to the other hospitals in the dataset may have increased between two years, but the absolute efficiency of the hospital may have in fact decreased. Techniques such as Malmquist productivity indices have been developed to facilitate the comparison of efficiency scores over time. The application of such techniques is beyond the scope of this investigation, but would be a useful extension in future research. For further details of efficiency changes over time and Malmquist productivity indices, the reader is directed to Coelli *et al* (2005).

6.2. Results for the 3x1y model

As discussed in section 5.5, the 3x1y model was expected to be the most representative of the hospital production process and is therefore the focus of this investigation. The reader is directed to section 5.5 for the specification of the 3x1y model. The results for this model are discussed below.

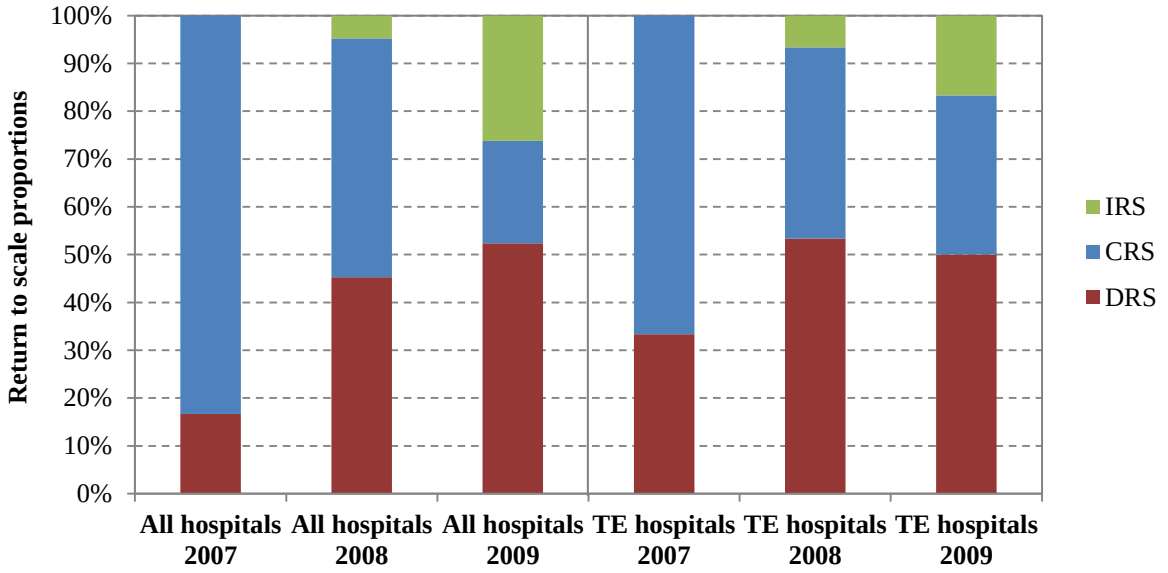

Figure 6: The proportion of the set of all hospitals and the set of technically efficient hospitals operating under each return to scale classification for the 3x1y model

Figure 6 displays the proportion of the set of all hospitals and the set of technically efficient hospitals operating under each return to scale classification for the 3x1y model. For the set of all hospitals, the majority of hospitals operate under CRS in 2007 and 2008, and under DRS in 2009. The proportion of hospitals operating under DRS increases over the three years (17% in 2007, 45% in 2008, and 52% in 2009), while the proportion operating under CRS decreases over the three years (83% in 2007, 50% in 2008, and 21% in 2009). It is interesting to note that no hospitals operate under IRS in 2007, and only a small proportion in 2008 (5%). The proportion of hospitals operating under IRS increases to 26% in 2009, which is a greater proportion than those operating under CRS (21%). The return to scale classifications of the hospitals is discussed in further detail below (the reader is directed to the discussion relating to Table 4).

These results imply that, in 2009, 79% of hospitals could benefit from a change in return to scale classification (while only 50% would benefit from this in 2008, and 17% in 2007). In 2009, 52% of the hospitals operate under DRS, which means that a decrease in their inputs should result in a less than proportional decrease in their outputs. This in theory could occur up to the point where CRS applies. Similarly, for the 26% of hospitals operating under IRS in 2009, an increase in their inputs should result in a more than proportional increase in their outputs. In practice a hospital will be

limited in the extent and speed that it can change its scale, say, because inputs may not be available or may take time to change, such as capital inputs.

There are large differences between the proportion of hospitals operating under each return to scale classification across all three years. However, it was expected that these proportions would be fairly stable across years as changes in scale are expected to happen gradually unless there are large changes in the production process, say, driven by strong management action applied across all hospitals. The dataset, and a general understanding of the operations of the company that provided the dataset, suggest that no large changes have occurred, in any of the three years, that would impact all of the hospitals. A possible explanation of the variability of return to scale classifications could be that a large proportion of hospitals are operating close to the point where their return to scale classifications change. This could cause their classifications to oscillate from year to year. The variability could also be caused by errors in the dataset. However, it is unlikely that errors of the size required to cause such variability would be present throughout the dataset.

For the set of technically efficient hospitals, the majority of hospitals operate under CRS in 2007, and under DRS in 2008 and 2009. The proportion of hospitals operating under CRS decreases over the three years (67% in 2007, 40% in 2008, and 33% in 2009). No technically efficient hospitals operate under IRS in 2007; while 7% operate under IRS in 2008 and 17% in 2009. The relationship between the proportion of technically efficient hospitals operating under each return to scale classification is also not stable across the three years. In particular, 2007 has a notably different return to scale profile than 2008 and 2009. The return to scale profiles of the set of all hospitals and the set of technically efficient hospitals are broadly consistent across the three years.

Table 3: The average technical and scale efficiency scores grouped by each return to scale classification for the 3x1y model

3x1y	**Technical efficiency (VRS)**			**Scale efficiency**		
	2007	**2008**	**2009**	**2007**	**2008**	**2009**
IRS	-	92.8%	87.2%	-	93.1%	96.3%
CRS	88.5%	90.8%	92.4%	94.9%	94.6%	99.2%
DRS	93.9%	89.9%	91.5%	83.8%	91.6%	90.4%
All hospitals	89.4%	90.5%	90.6%	93.1%	93.2%	93.8%
Std dev of all hospitals	12.1%	11.6%	12.0%	8.5%	7.5%	6.7%

Table 3 contains the average technical and scale efficiency scores for the 3x1y model for each return to scale classification and year. The average technical efficiency scores for all hospitals ranged from 89.4% in 2007 to 90.6% in 2009. As expected, these average technical efficiency scores remain fairly stable across all three years. This is because it is likely that the average technical efficiency of a group on hospitals would change gradually over time, unless there have been drastic changes to the

production process. The average technical efficiency scores show that the set of hospitals, on average, could have produced the same level of output by using up to 10.6% less inputs in 2007, 9.5% less in 2008, and 9.4% less in 2009. However, the extent to which these savings could be realised in practice is likely to be limited. The reader is directed to section 7.2 for further details of this. The average technical efficiency score increased from 2007 to 2009. This means that in each year the technical efficiency scores of the set of hospitals, relative to each other, have increased. Note that an increase in average technical efficiency score does not necessarily imply that the set of hospitals is operating more efficiently than the comparative year (Coelli *et al*, 2005). For example, the set of hospitals may have become less efficient but the differences in efficiency between the inefficient and efficient hospitals may have reduced, increasing the average efficiency score but not the absolute efficiency of the set of hospitals relative to the comparative year. Malmquist productivity indices can be used to analyse changes in efficiency over time, as was done in the study by Zere *et al* (2001). For further details of Malmquist productivity indices the reader is directed to Coelli *et al* (2005).

In each year, the relationship between return to scale classification and average technical efficiency is different. In 2007, hospitals exhibiting DRS are the most efficient. In 2008, IRS hospitals are the most efficient; and in 2009, CRS hospitals are the most efficient. It should be noted that the set of hospitals falling into each of the return to scale classifications in each year is not necessarily the same. Furthermore, it is clear that the relationship between return to scale classification and average technical efficiency is not stable across the three years. It was however expected that this relationship would remain relatively stable across all three years. The small proportion of hospitals exhibiting DRS in 2007, IRS in 2008 and CRS in 2009 could be driving the variation in the average technical efficiency scores. This is because the relatively small sample of, say, DRS hospitals in 2007 would have less stable efficiency scores than larger samples. This could impact the credibility of the average efficiency scores and could possibly explain why hospitals operating under IRS have the highest average technical efficiency score in 2008 and the lowest score in 2009.

The relationship between average scale efficiency scores and return to scale classification is relatively stable across all three years. The average scale efficiency scores for all hospitals ranged from 93.1% in 2007 to 93.8% in 2009. This small change is in line with expectations because the average scale efficiency of a group of hospitals should change gradually over time, due to the difficulties associated with changing the scale of production. For example, difficulties could occur because of the indivisible nature of capital inputs. The average scale efficiency scores show that the set of hospitals, on average, could have produced the same level of output by using up to 6.9% less inputs in 2007, 6.8% less in 2008, and 6.2% less in 2009. Across all three years, hospitals operating under CRS are, on average, more scale efficient than those operating under DRS or IRS. This is as expected because only hospitals operating under CRS can be scale efficient and, by definition, hospitals operating under

DRS or IRS cannot operate under CRS. It is important to note that technically efficient hospitals that operate under CRS are also scale efficient; but technically inefficient hospitals that operate under CRS are not scale efficient i.e. a hospital can operate under CRS and not be scale efficient. In 2008 and 2009, it was noted that hospitals operating under IRS have higher average scale efficiency scores than hospitals operating under DRS. No hospitals operate under IRS in 2007.

The standard deviations of technical efficiency scores are relatively stable across all three years, while the standard deviations of scale efficiency scores have decreased from 2007 to 2009. This means that the scale efficiency scores in 2009 exhibit less variation than the scale efficiency scores in 2007. The standard deviation of the scale efficiency scores is lower than that of the technical efficiency scores, indicating that the scale efficiency scores are more stable than the technical efficiency scores.

The relationship between hospital size, return to scale classification, technical efficiency and scale efficiency is summarised in Table 4, as well as Figures 7 and 8 below. Note that, in Table 4, average number of beds refers to the average number of beds across the three year period from 2007 to 2009. As an input variable, total number of hospital beds has been used as a proxy for the flow of capital services. Additionally, the average number of beds, as shown in Table 4, has been interpreted as a proxy for hospital size.

Table 4: Average number of beds, return to scale classification, technical efficiency and scale efficiency under the 3x1y model

Hospital number	Average number of beds	2007	2008	2009	Hospital is TE across all years	Hospital is SE across all years
1	19	CRS	IRS	IRS	TE=1	
2	20	CRS	CRS	IRS	TE=1	
3	25	CRS	DRS	IRS		
4	26	CRS	IRS	IRS		
5	30	CRS	CRS	CRS	TE=1	SE=1
6	36	CRS	DRS	IRS		
7	38	DRS	DRS	IRS		
8	40	CRS	CRS	CRS	TE=1	SE=1
9	48	CRS	CRS	IRS		
10	54	CRS	CRS	IRS		
11	65	CRS	CRS	CRS		
12	71	DRS	DRS	DRS		
13	78	CRS	DRS	IRS		
14	91	CRS	DRS	CRS		
15	96	CRS	DRS	DRS		
16	101	CRS	CRS	DRS		
17	107	CRS	DRS	DRS		
18	116	DRS	DRS	DRS		
19	121	CRS	DRS	IRS		
20	124	CRS	CRS	DRS		
21	132	CRS	DRS	DRS		
22	137	CRS	CRS	DRS		
23	139	CRS	CRS	IRS		
24	139	DRS	DRS	DRS		
25	149	CRS	DRS	DRS		
26	178	CRS	CRS	CRS		
27	178	CRS	CRS	CRS		
28	185	CRS	DRS	DRS		
29	185	CRS	CRS	DRS		
30	186	CRS	CRS	DRS		
31	193	CRS	CRS	CRS	TE=1	SE=1
32	194	CRS	CRS	CRS		
33	200	CRS	DRS	DRS		
34	205	CRS	CRS	DRS		
35	206	CRS	DRS	CRS		
36	224	CRS	CRS	DRS		
37	225	CRS	CRS	DRS		
38	226	DRS	DRS	DRS		
39	230	CRS	CRS	DRS		
40	249	CRS	CRS	DRS		
41	342	DRS	DRS	DRS	TE=1	
42	377	DRS	DRS	DRS	TE=1	

Table 4 shows that, across the three years, there is a relatively high amount of variation in the return to scale classifications of the hospitals. However, as mentioned in the discussion relating to Figure 6, it was expected that the return to scale classifications of the hospitals would remain fairly stable over the three year period. In 2007, three of the largest hospitals (hospital 38, 41 and 42) operate under DRS. This is as expected because large hospitals when measured in terms of number of beds, would be more likely to operate on the DRS section of the production frontier. However, it would also be expected that small hospitals would operate under IRS which is not the case in 2007. In 2008, two smaller hospitals (hospital 1 and 4) operate under IRS, and the same three large hospitals as 2007 (hospital 38, 41 and 42) operate under DRS. In 2009, 8 out of the 10 smallest hospitals operate under IRS, and 9 of the largest 10 hospitals operate under DRS (including the three large hospitals that operate under DRS in 2007 and 2008). The return to scale profiles of 2008 and 2009 are therefore more in line with expectations.

It is interesting to note that out of the 7 hospitals that exhibit DRS in 2007, 6 of these hospitals also exhibit DRS in 2008 and 2009. This provides some evidence of consistency in return to scale classifications across years. It is also interesting to note that the two smallest and two largest hospitals are technically efficient across all three years. This is perhaps because these hospitals operate on the extreme sections of the production frontier where there are few hospitals of a similar size that they can be compared against for efficiency measurement purposes.

Most hospitals in 2009 and a large number of hospitals in 2008 operate under DRS. A hospital exhibiting DRS implies that the hospital has excess capacity. Excess capacity could be in the form of underutilised inputs, such as beds and nurses. It should be noted that the DRS classification of a particular hospital, and hence the implication of excess capacity, has been determined relative to the other hospitals in the dataset and is not an absolute standard. From the perspective of a South African private hospital, excess capacity may be desirable since paying patients expect to be treated promptly without waiting for treatment capacity to become available. This may be an incentive for private hospitals to err on the side of being too large, thereby operating with excess capacity, rather than being too small – even if this comes at the cost of decreased efficiency. If this is the case then hospitals may be more likely to exhibit DRS than IRS, which is consistent with the results of this investigation across all three years. It could even be expected that most hospitals would operate under DRS as was the case in 2009. The reader is directed to the discussion relating to Table 6 for further details regarding excess capacity within the South African private hospital industry.

The incentive for private hospitals to operate with excess capacity may also imply that hospitals operating under IRS may tend to be more scale efficient than those operating under DRS hospitals. This is because the incentive may lead hospitals operating under IRS to increase the scale of their operations, bringing these hospitals closer to the region of CRS which, by definition, contains the

most scale efficient hospitals. However, hospitals operating under DRS already possess excess capacity and would therefore not have the same incentive to shift their operations closer to the region of CRS. These dynamics would lead to hospitals operating under IRS, on average, being more scale efficient than those operating under DRS. This is consistent with the results in Table 3, where average scale efficiency scores are greater for hospitals operating under IRS than DRS.

The results presented in Table 4 for 2009 are displayed graphically in Figures 7 and 8 below. It was decided to focus on 2009 since this was the most recently available data. However, the results for all years are displayed in Figures 12 and 13. The relationship between hospital size (measured in terms of number of beds), return to scale classification and technical efficiency in 2009 is displayed graphically in Figure 7 below.

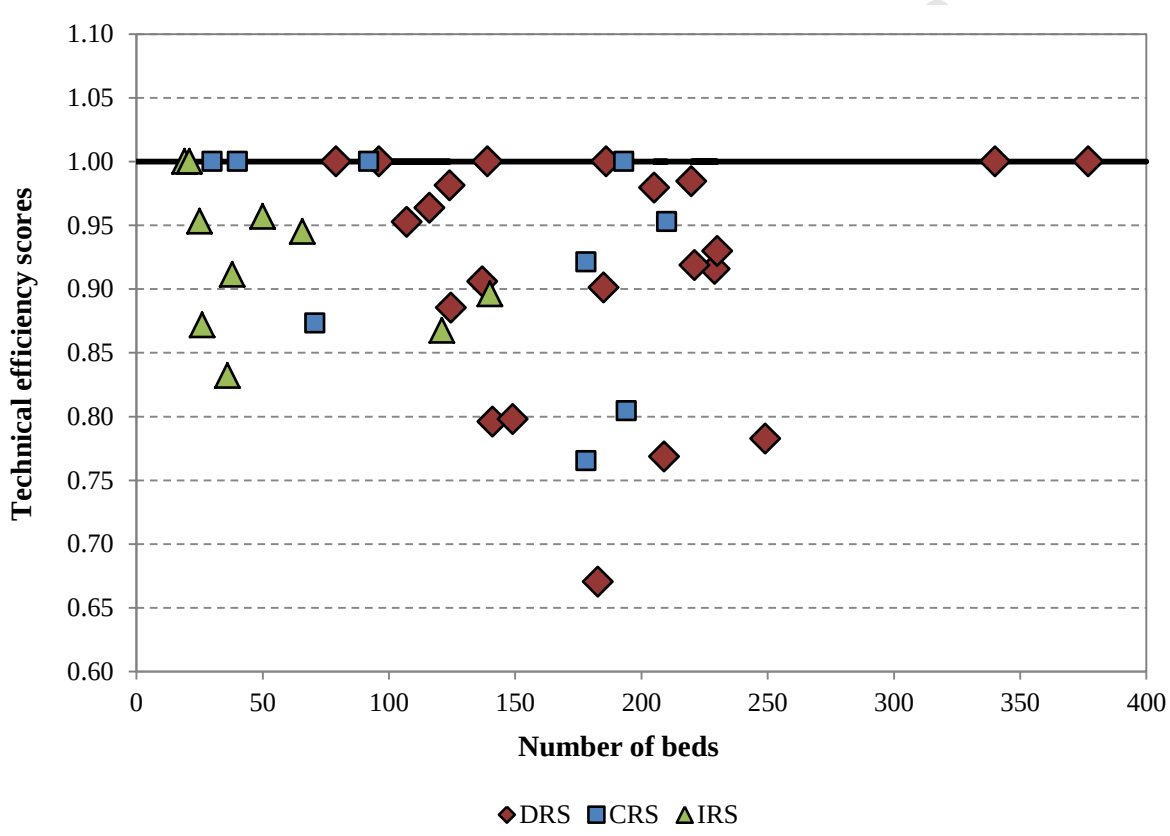

Figure 7: Technical efficiency and number of beds for each hospital in 2009 under the 3x1y model

In Figure 7 it can be seen that a relatively large number of technically efficient hospitals have less than 100 beds. Furthermore, unlike hospitals with more than 100 beds, these hospitals do not have technical efficiency scores of less than 0.8. A possible explanation of these results is that hospitals with fewer beds and other inputs may be easier to manage in an efficient manner. Smaller hospitals may also be different to larger hospitals with regard to specialised treatment units and the complexity of cases. For example, extreme or complex cases may be referred to larger hospitals which have the resources and facilities necessary to treat them. These extreme cases may not be adequately reflected

in the case-mix adjustments, making larger hospitals appear less efficient relative to smaller hospitals in the dataset. Furthermore, smaller hospitals may treat cases that are simpler and more homogeneous than those treated by larger hospitals. This is supported by Figure 5 which shows that case-mix adjustment factors tend to be lower for smaller hospitals and higher for larger hospitals. Management of smaller hospitals may therefore be able to improve their technical efficiency, relative to larger hospitals, by focusing their attention on the efficient treatment of frequently occurring, simpler cases. Consistent with the results displayed in Table 4, Figure 7 also shows that hospitals with a lower number of beds tend to operate under IRS. As the number of beds increases some hospitals exhibit DRS and some CRS. The largest hospitals exhibit DRS. Note that the development over time of each hospital's technical efficiency and return to scale classification is displayed graphically in Figure 12.

For presentational purposes, hospital 13 is not displayed in Figure 7 as it has the lowest technical efficiency score of 0.36 in 2009. This hospital is a specialist hospital that exhibits IRS in 2009. Specialist hospitals are typically resource intensive and may be subject to different production dynamics, which has led to this particular hospital having a much lower technical efficiency score. Interestingly, even though hospital 13 has very low technical efficiency, it is still within the normal range of scale efficiency relative to the other hospitals in the dataset.

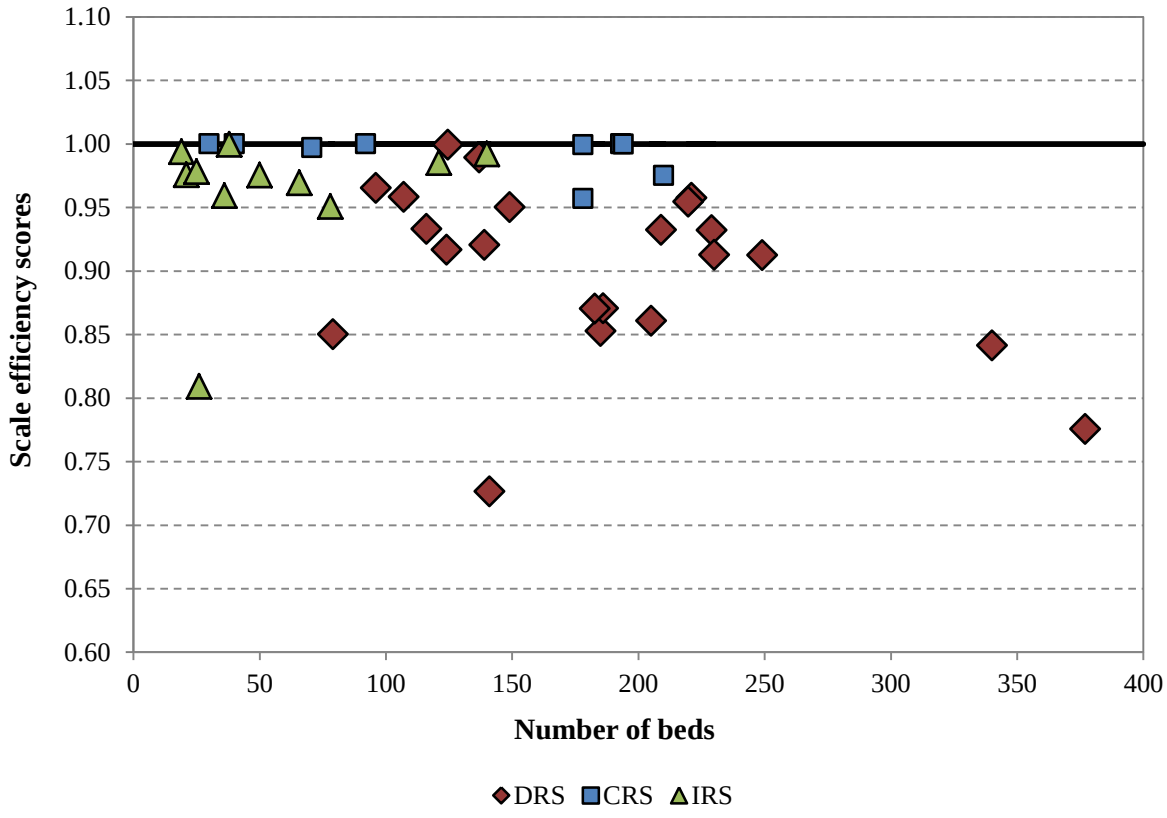

Figure 8: Scale efficiency and number of beds for each hospital in 2009 under the 3x1y model

Figure 8 displays the relationship between hospital size (measured in terms of number of beds), return to scale classification and scale efficiency in 2009. It can be seen that hospitals with a lower number of beds appear to be operating with greater scale efficiency. Consistent with the discussion relating to Table 4, it can also be seen that the scale efficiency scores of IRS hospitals appear, on average, to be higher than the scale efficiency scores of DRS hospitals.

It can also be seen in Figure 8 that hospitals with less than 30 beds experience IRS and hospitals with more than 210 beds experience DRS. As expected, all scale efficient hospitals operate under CRS. Furthermore, these scale efficient hospitals lie within the range of 30 to 194 beds. This implies that the most productive scale size (MPSS), measured using the proxy of beds, is likely to lie within the region of the efficient frontier where hospitals have between 30 and 194 beds. However, it should be noted that the MPSS depends on all inputs into the production process, not only on the number of beds. The development of each hospital's scale efficiency and return to scale classification over time, under the 3x1y model, is displayed graphically in Figure 13.

Table 5: The correlations between number of beds and efficiency scores under the 3x1y model

3x1y	2007	2008	2009
Correlations between number of beds and TE scores	-11.4%	-17.6%	-0.2%
Correlations between number of beds and SE scores	-71.6%	-59.4%	-41.9%

Table 5 shows that the correlations between scale efficiency scores and number of beds are relatively large and negative. This suggests that smaller hospitals may be more scale efficient, which is consistent with the results displayed in Figure 8. The negative correlations between technical efficiency scores and number of beds are much weaker. This may suggest that scale efficiency scores are more sensitive to changes in number of beds than technical efficiency scores. However it should be noted that there is a relatively large amount of variation in both sets of correlations across all three years.

Table 6: The average and maximum occupancy rates for 2007 to 2009

Occupancy rates	2007	2008	2009
Average	63.1%	64.5%	65.6%
Maximum	80.2%	81.5%	83.3%

Table 6 displays the average and maximum annual occupancy rates. The average annual occupancy rates are relatively low, ranging from 63.1% in 2007 to 65.6% in 2009. This is consistent with the findings of the African National Congress (2010) that South African private hospitals have average occupancy rates of around 65%. The maximum occupancy rates are in line with international best

practice, being less than 85% across all three years (Keegan, 2008). The reader is directed to Figures 9 and 10 for further details regarding occupancy rates and return to scale classifications.

The relatively low average annual occupancy rates, together with the result that most hospitals operate under non-increasing returns to scale, reinforces the general criticism that excess capacity exists within the private hospital industry (African National Congress, 2010). However, excess capacity within the private hospital industry may be appropriate given the operational goals of private hospital organisations and the nature of their ownership. Private hospitals provide healthcare services to medical scheme patients and patients that pay directly for their services. These patients expect to be treated promptly without waiting for treatment capacity to become available. From a business perspective, the expectations and satisfaction of these patients are of central concern when setting hospital capacity levels. Reducing excess capacity will increase the scale efficiency of a hospital but may have a negative impact on the demand for its healthcare services. For example, long waiting periods may alienate patients who then seek healthcare from other providers. This creates an incentive for private hospitals to operate with excess capacity. In contrast, public hospitals by definition are not businesses and can, in theory, reduce capacity to optimal levels without impacting the sustainability of their organisation. Therefore the operational goals and the nature of ownership of private hospitals, when considered in relation to the public sector and the current South African healthcare system, may naturally lend itself to excess capacity within the private industry.

The results discussed in this section thus far are consistent with Kibambe & Koch (2007), who found that public hospitals in Gauteng are more likely to operate under DRS than IRS. Kibambe & Koch (2007) proposed that these hospitals may operate under DRS due to the emigration of medical professionals or the need to hold excess capacity in order to cope with potential medical catastrophes. However, these reasons may not be applicable to the private sector. For example, the ability to cope with wide-spread medical catastrophes may be seen as a function that should be fulfilled by the state rather than by private hospitals. With regard to emigration, higher earnings and better working conditions in the private sector may make private hospitals less susceptible to emigration than public hospitals. However, some drivers of emigration are common to both the private and public sectors. For example, a survey conducted by Arnold & Lewinsohn (2010) found that the most common reason for the emigration of South African doctors to Australia from 1990 onwards was a concern over the level of violent crime in South Africa. As skilled medical professionals emigrate, hospital management may struggle to fully staff their hospitals. Consequently, hospitals may not have sufficient medical professionals to operate at their optimal capacity. This could lead to an oversupply of other inputs. For example, beds may become too numerous to be attended to by the remaining nursing staff. It is through this mechanism that emigration may contribute to hospitals operating under DRS.

Zere *et al* (2001) found that approximately 50% of public hospitals in the Northern, Eastern and Western Cape exhibited DRS, 13% exhibited CRS, and 37% exhibited IRS. Their finding that hospitals operating under DRS and CRS are more common than those operating under IRS is consistent with this analysis. However, as mentioned in section 4.7, caution must be applied when comparing results from the public and private sectors as this may not be appropriate.

The relationship between occupancy rates and hospital size (measured in terms of number of beds) in 2009 is displayed graphically in Figure 9 below. The return to scale classifications for Figures 9 and 10 were derived using the 3x1y model.

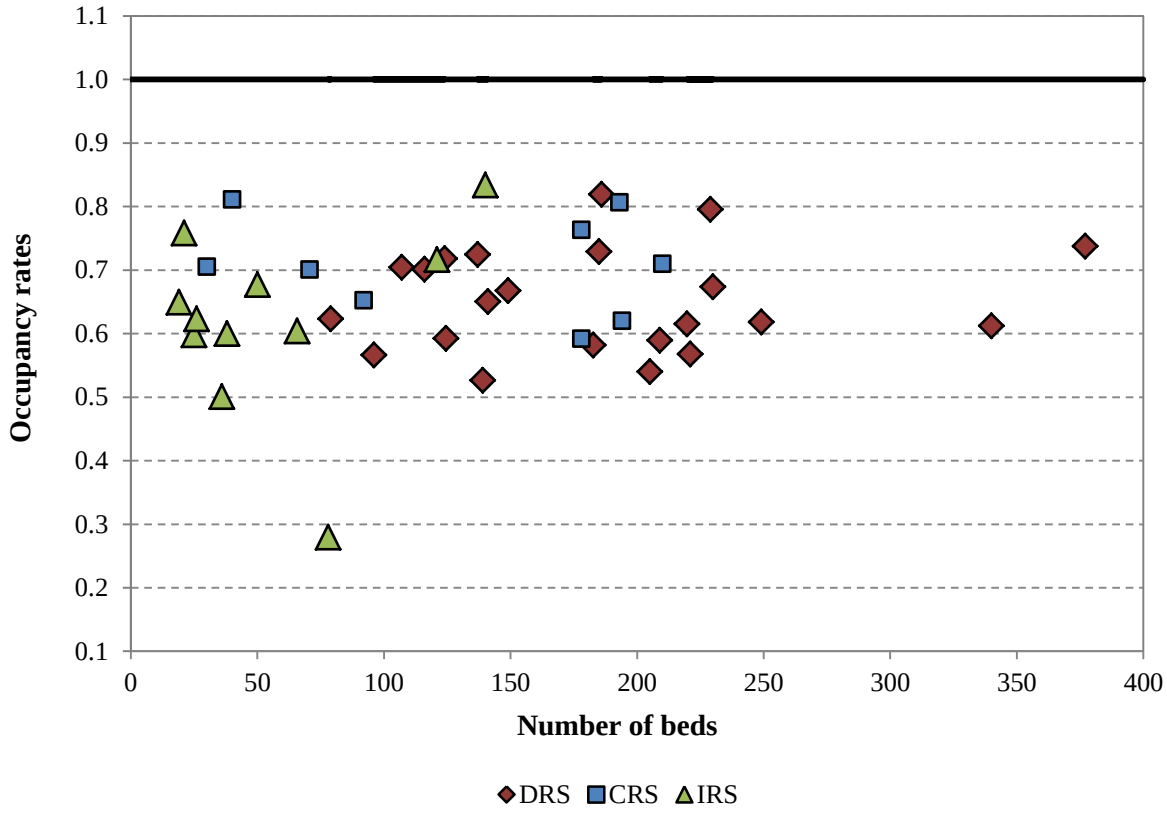

Figure 9: Occupancy rates and number of beds for each hospital in 2009 under the 3x1y model

From Figure 9, there does not appear to be a clear relationship between hospital occupancy rates and size. Figure 10 shows the relationship between occupancy rates and return to scale classification in 2009.

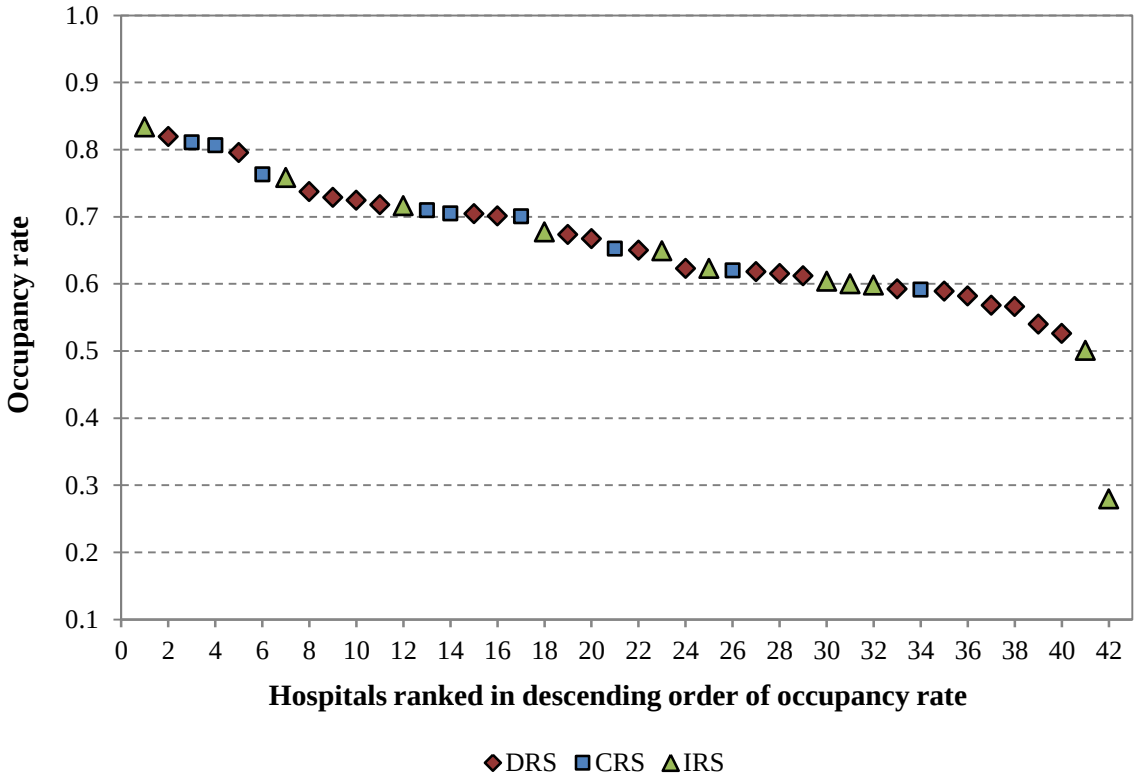

Figure 10: Occupancy rates and return to scale classification in 2009 under the 3x1y model

From Figure 10, there does not appear to be a clear relationship between hospital occupancy rates and return to scale classification in 2009. The hospital with the lowest occupancy rate (hospital 13) is the same specialist hospital that has a very low technical efficiency score. This was discussed in the commentary relating to Figure 7.

Figures 9 and 10 provide good examples of how oversimplified analyses can lead to incorrect conclusions regarding scale. A simple approach to determining scale may be to rank hospitals by number of beds or occupancy rates, as was done in Figures 9 and 10 respectively, and then label the regions of IRS, CRS, and DRS according to these rankings (Sherman & Zhu, 2006). Given the current dataset, these types of analyses would have led to incorrect conclusions regarding scale. This can be seen in Figures 9 and 10 where there does not appear to be a clear relationship between return to scale classification and number of beds or occupancy rates. By only using one factor (such as number of beds or occupancy rates) to analyse scale, much of the richness of the dataset is lost.

Using the 2009 hospital data, the hospitals were divided into three equal groups based of their number of beds. A third of the 2009 hospitals have less than 93 beds while two thirds have less than 185 beds. The maximum number of beds in each year was 377. Table 7 shows the proportion of hospitals falling into each of these groups for 2007 to 2009. The three groups are labelled group A, B, and C.

Table 7: The proportion of hospitals falling into each group (A, B or C) based on their number of beds

3x1y	All hospitals		
	2007	2008	2009
0 to < 93 beds (group A)	35.7%	35.7%	33.3%
93 to < 185 beds (group B)	28.6%	28.6%	33.3%
185 to ≤ 377 beds (group C)	35.7%	35.7%	33.3%

Figure 11 shows the proportion of technically efficient and scale efficient hospitals falling into each of these three groups. Table 8 then shows the average technical efficiency and scale efficiency scores of each of the three groups. All technical and scale efficiency scores are shown for 2007 to 2009 and are calculated using the 3x1y model.

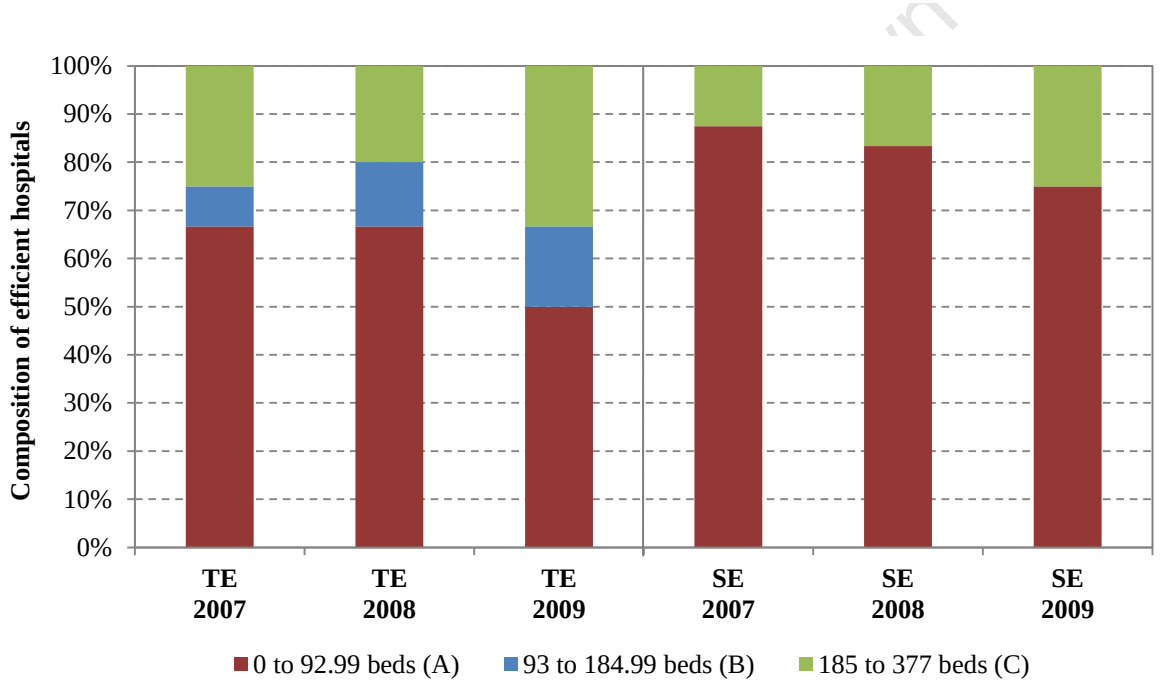

Figure 11: The proportion of technically efficient and scale efficient hospitals grouped by number of beds, for the 3x1y model

It is clear that group A has the highest proportion of technically efficient hospitals and, by far, the highest proportion of scale efficient hospitals. Group B has the lowest proportion of technically efficient hospitals and no scale efficient hospitals. Group C contains the second highest proportion of technically efficient hospitals and scale efficient hospitals. For each year, the proportion of technically efficient hospitals and scale efficient hospitals in group A is at least 1.5 times that of any other group. Figure 11 suggests that smaller hospitals and larger hospitals (measured in terms of number of beds) are more likely to be technically or scale efficient, with smaller hospitals being the most likely. The reader is directed to the discussion relating to Figure 7 for a possible explanation as to why smaller

hospitals exhibit higher technical efficiency. The higher scale efficiency of smaller hospitals is discussed below.

Table 8: The average technical efficiency and scale efficiency scores grouped by number of beds, for the 3x1y model

3x1y	Average technical efficiency (VRS)			Average scale efficiency		
	2007	2008	2009	2007	2008	2009
0 to < 93 beds (group A)	93.7%	94.8%	90.7%	98.8%	97.9%	96.1%
93 to < 185 beds (group B)	85.3%	87.63%	88.6%	94.1%	92.8%	94.0%
185 to ≤ 377 beds (group C)	86.5%	87.64%	90.2%	87.5%	89.7%	92.3%

The average efficiency scores in Table 8 lead to similar conclusions for group A, which has the highest average technical efficiency and scale efficiency scores. Group B, although not containing any scale efficient hospitals, has the second highest average scale efficiency scores, and the lowest average technical efficiency scores across all three years. Group C has the second highest average technical efficiency scores and the lowest average scale efficiency scores across all three years. However, the average technical efficiency scores of group B and group C are relatively similar.

The rankings of the three groups according to their average technical and scale efficiency scores remain constant from 2007 to 2009. Average technical efficiency scores are therefore higher for smaller hospitals (measured in terms of number of beds), but there is no clear relationship for larger hospitals. Average scale efficiency scores are highest in group A, followed by group B, and lowest in group C. This implies that scale efficiency decreases, on average, with increasing number of beds. This result is consistent with the discussion, relating to Table 4, that there may be an incentive for hospitals to have large operations and excess capacity, and that this may lead to smaller hospitals being more scale efficient than larger hospitals. The results are also consistent with Zere *et al* (2001) who found that smaller public hospitals were relatively more scale efficient than larger public hospitals. However, as mentioned in section 4.7, caution must be applied when comparing results from the public and private sectors as this may not be appropriate. The results presented in Table 8 also suggest that the most productive scale size (MPSS), measured using the proxy of beds, is likely to fall within group A (the group of hospitals with less than 93 beds). This is consistent with the discussion relating to Figure 8.

Figures 12 and 13 display, at a hospital level, the changes in efficiency scores and return to scale classifications from 2007 to 2009 for the 3x1y model.

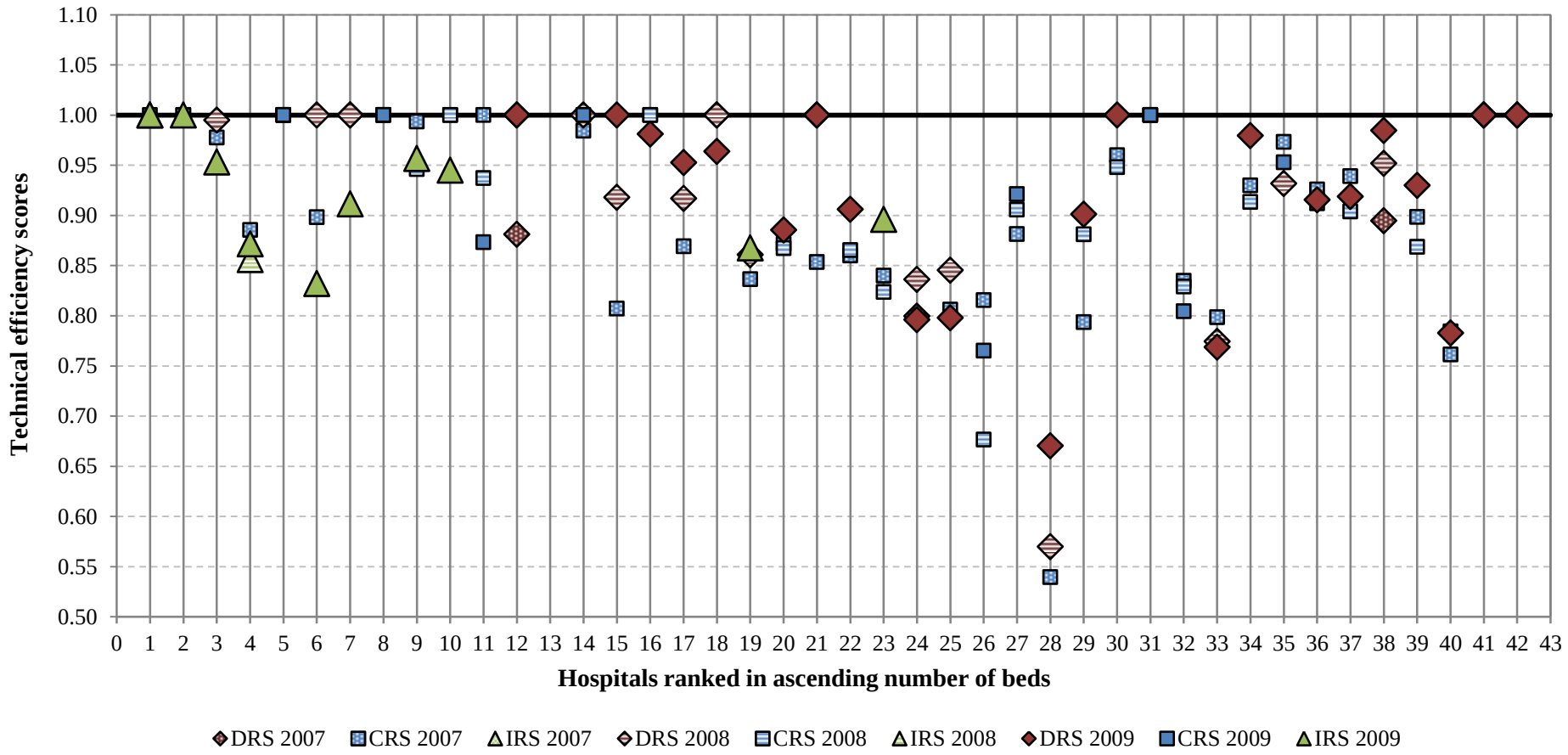

Figure 12: Technical efficiency scores and return to scale classifications for all hospitals from 2007 to 2009 for the 3x1y model

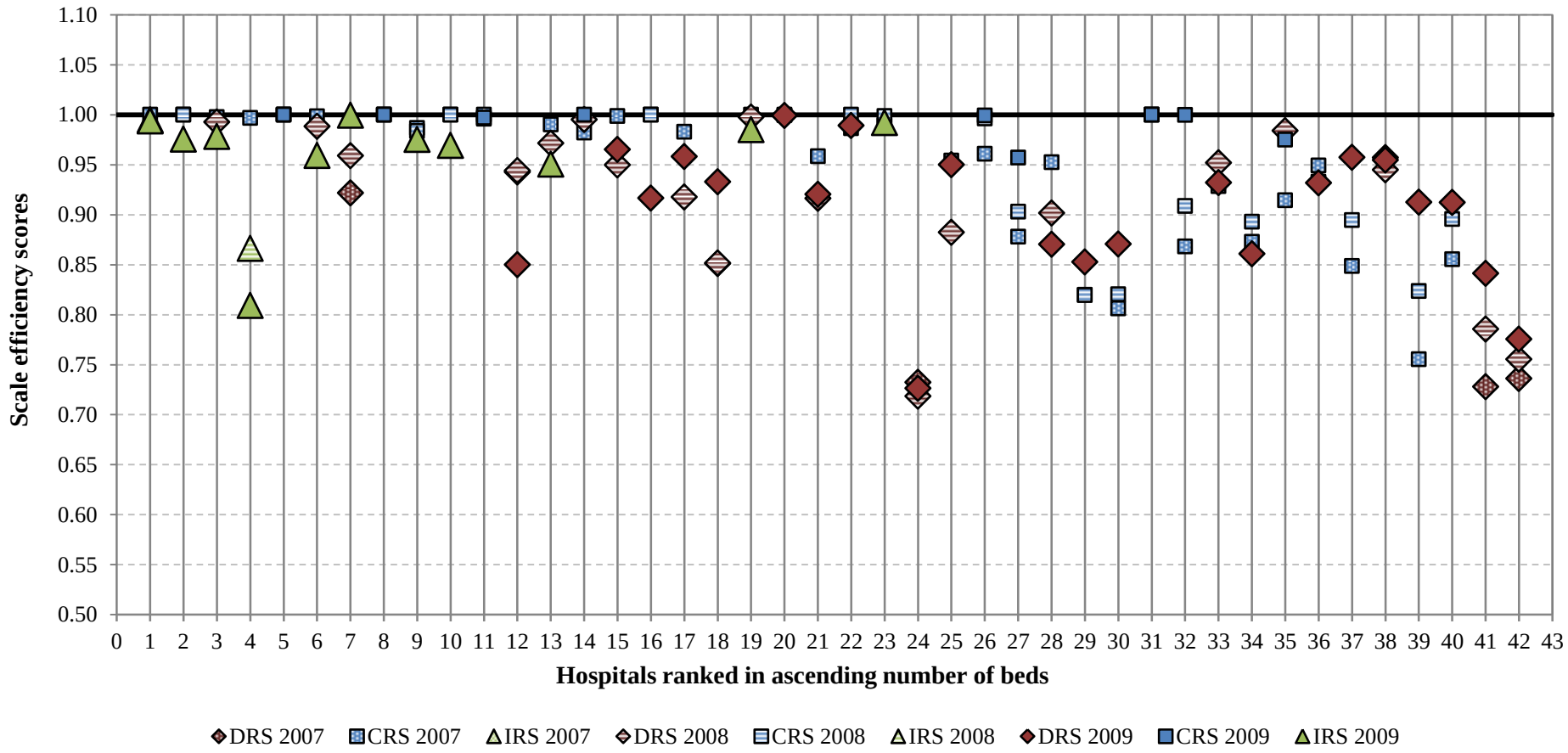

Figure 13: Scale efficiency scores and return to scale classifications for all hospitals from 2007 to 2009 for the 3x1y model

Figures 12 and 13, together with Table 4, assisted with identifying hospitals of interest for further individual analysis. In these figures, each hospital is associated with a vertical line which can be used to trace its change in efficiency and return to scale classification over time. Five hospitals were identified for further analysis, namely hospitals 3, 16, 18, 28 and 41. These hospitals are interesting because:

- Hospital 3 operates under a different return to scale classification in each of the three years.
- Hospital 16 changes from operating under CRS in 2007 and 2008 to operating under DRS in 2009. It is also technically and scale efficient in 2007 and 2008, but not in 2009.
- Hospital 18 exhibits the same return to scale classification across all three years, but loses its technical efficiency in 2009. It is also not scale efficient in any year.
- Hospital 28 operates with low technical efficiency and relatively low scale efficiency across all three years.
- Hospital 41 is technically efficient across all three years, but operates with low scale efficiency.

The reader is directed to section 6.4 for further details of the individual hospital analysis. Note that hospital 13 is not displayed in Figure 12. Again, this is for presentational purposes as hospital 13 is a specialist hospital and has the lowest technical efficiency scores (0.43 in 2007, 0.48 in 2008, and 0.36 in 2009).

Table 9 below provides a breakdown of each hospital's efficiency reference set (ERS) in 2009 under the 3x1y model. Hospitals 1 to 42 are listed in the first column of the table. Only technically efficient hospitals appear in the ERS, and these hospitals, together with their return to scale classifications, are provided in the other columns of the table. The return to scale classification of each of the ERS hospitals is summarised in the last row of the table. A count of the number of times each of these hospitals appears in another hospital's ERS is provided in the second last row of the table. The last column of the table provides the return to scale classification of each hospital, which corresponds to the results presented in Table 4. Table 9 provides a reference as to how the return to scale classification of each hospital is determined. The reader is directed to section 4.4 for details regarding ERSs and their role in return to scale classification. The ERS of a hospital indicates which set of technically efficient hospitals has the most similar operations to that hospital. For a particular hospital, a linear combination of the hospitals in its ERS will project that hospital's operations onto the efficient frontier. The ERS of selected individual hospitals is examined in section 6.4.

As an indication of how the return to scale classifications of the ERS hospitals (technically efficient hospitals) change over time, the reader is directed to Figure 6 which shows the proportion of technically efficient hospitals operating under each return to scale classification for each of the three years. Note that the set of ERS hospitals contains 12 hospitals in 2007, 15 in 2008, and 12 in 2009. A

count of the total number of hospitals appearing in each hospital's ERS for each year is provided in Table 15. This is discussed at the end of section 6.3.

Table 9: Each hospital's ERS and return to scale classification in 2009 under the 3x1y model

	1	2	5	8	12	14	15	21	28	31	41	42	RTS
1	IRS												IRS
2		IRS											IRS
3	IRS		CRS										IRS
4	IRS					CRS							IRS
5			CRS										CRS
6	IRS		CRS										IRS
7	IRS		CRS	CRS									IRS
8				CRS									CRS
9	IRS	IRS				CRS				CRS			IRS
10	IRS					CRS				CRS			IRS
11			CRS			CRS				CRS			CRS
12					DRS								DRS
13	IRS			CRS						CRS			IRS
14						CRS							CRS
15							DRS						DRS
16					DRS		DRS	DRS					DRS
17				CRS			DRS	DRS		CRS			DRS
18					DRS		DRS	DRS					DRS
19	IRS					CRS				CRS			IRS
20				CRS			DRS			CRS			DRS
21								DRS					DRS
22				CRS			DRS			CRS			DRS
23	IRS					CRS				CRS			IRS
24					DRS			DRS					DRS
25				CRS			DRS	DRS		CRS			DRS
26			CRS			CRS				CRS			CRS
27						CRS				CRS			CRS
28				CRS	DRS			DRS					DRS
29								DRS	DRS				DRS
30									DRS				DRS
31										CRS			CRS
32			CRS			CRS				CRS			CRS
33								DRS	DRS	CRS			DRS
34								DRS	DRS	CRS			DRS
35			CRS							CRS			CRS
36									DRS	CRS			DRS
37									DRS	CRS			DRS
38										CRS	DRS		DRS
39									DRS	CRS			DRS
40								DRS	DRS	CRS			DRS
41											DRS		DRS
42												DRS	DRS
Count	10	2	8	8	5	10	7	11	8	22	2	1	
RTS	IRS	IRS	CRS	CRS	DRS	CRS	DRS	DRS	DRS	CRS	DRS	DRS	

6.3. Results comparison across all three models

This section compares the results of the three models specified in section 5.5 (model 3x1y, 3x3y and 1x3y). As discussed in section 5.5, the 3x3y model captures the greatest amount of available output information, but at the cost of double counting some outputs; while the 1x3y examines the relationship between operational beds and returns to scale. The 3x1y model was discussed in the previous section, and is expected to be the most representative of the hospital production process.

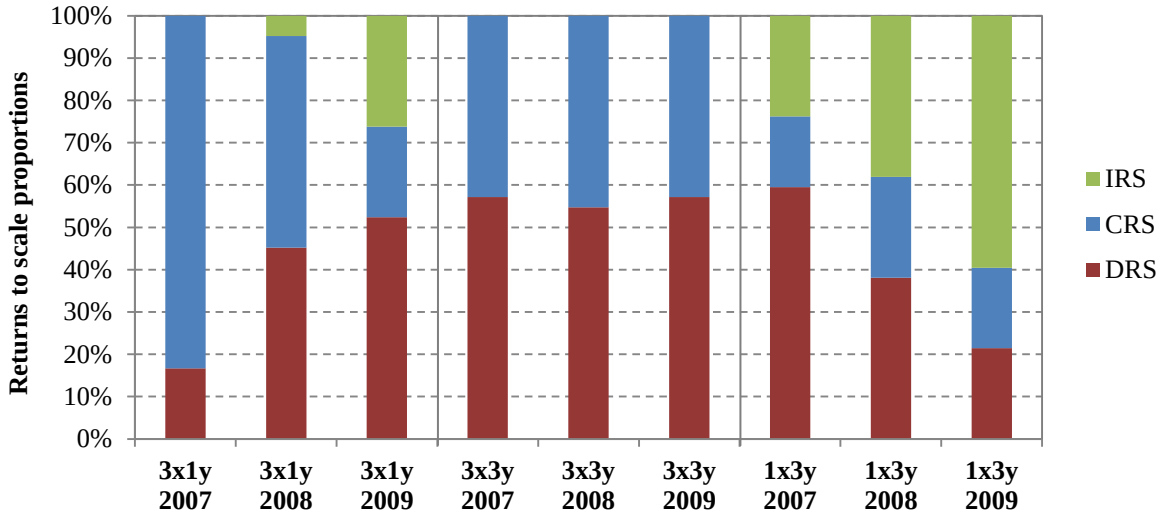

Figure 14: The proportion of the set of all hospitals operating under each return to scale classification for each model and year

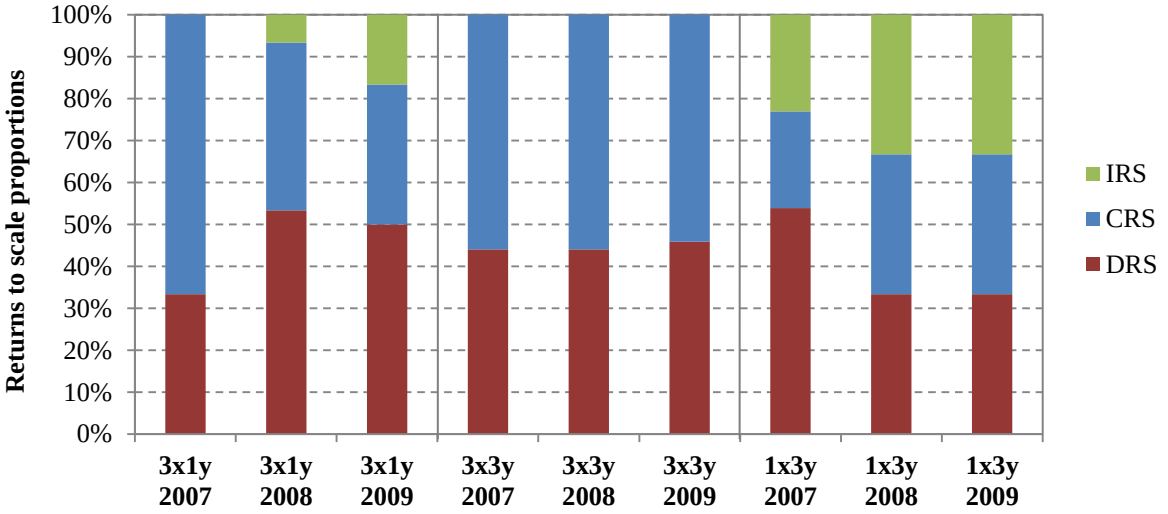

Figure 15: The proportion of the set of technically efficient hospitals operating under each return to scale classification for each model and year

Figures 14 and 15 display the proportion of the set of all hospitals and the set of technically efficient hospitals operating under each return to scale classification for each model and year. It can be seen in

these figures that the results show some inconsistencies across the three models. For example, some hospitals operate under IRS in the 3x1y and 1x3y models; while no hospitals operate under IRS in the 3x3y model. Additionally, for the set of all hospitals, the 3x1y and 1x3y models both show an increasing proportion of hospitals operating under IRS over time. However, the proportion of hospitals operating under DRS increases in the 3x1y model and decreases in the 1x3y model. These results are not consistent, which implies that they are sensitive to model specification.

As mentioned in section 6.2, it was expected that the proportion of hospitals operating under each return to scale classification would be fairly stable across the three years as changes in scale are expected to happen gradually. This is indeed the case for the 3x3y model, but not for the 3x1y and 1x3y models.

In all but one case (1x3y in 2009 for the set of all hospitals), CRS and DRS are the most prevalent return to scale classifications. This result is consistent with the claim that excess capacity exists within private hospitals.

Table 10: The average technical and scale efficiency scores grouped by each return to scale classification for the 3x3y model

3x3y	**Technical efficiency (VRS)**			**Scale efficiency**		
	2007	**2008**	**2009**	**2007**	**2008**	**2009**
IRS	-	-	-	-	-	-
CRS	95.6%	97.9%	94.0%	99.8%	99.2%	99.6%
DRS	95.4%	93.1%	96.1%	91.0%	92.8%	93.9%
All hospitals	95.5%	95.3%	95.2%	94.7%	95.7%	96.3%
Std dev of all hospitals	9.7%	9.4%	11.0%	5.9%	5.0%	4.2%

Table 11: The average technical and scale efficiency scores grouped by each return to scale classification for the 1x3y model

1x3y	**Technical efficiency (VRS)**			**Scale efficiency**		
	2007	**2008**	**2009**	**2007**	**2008**	**2009**
IRS	81.5%	83.9%	83.7%	89.1%	91.5%	95.2%
CRS	86.6%	89.5%	89.4%	100.0%	100.0%	98.8%
DRS	86.5%	88.6%	92.4%	94.7%	93.5%	94.6%
All hospitals	85.3%	87.0%	86.7%	94.2%	94.3%	95.8%
Std dev of all hospitals	13.2%	12.9%	12.7%	6.9%	6.4%	5.3%

The reader is directed to Table 3 for the above results for the 3x1y model. Note that it is not possible to compare changes in efficiency and average efficiency directly across models and years. This is because the efficiency scores are calculated relative to the other hospitals in a particular dataset, using a particular model. When different models or datasets are used (including data from different years),

the differences in the calculated efficiency scores do not necessarily represent absolute differences or changes in efficiency (Sherman & Zhu, 2006).

It is expected that, for a particular model, the average technical efficiency scores and average scale efficiency scores will only change gradually over the three year period. This is indeed the case for the average technical efficiency scores which, for a particular model, remain fairly stable across the three years. The average scale efficiency scores also remain relatively stable, but increase over the three year period.

Across all three models and years, hospitals operating under CRS are, on average, more scale efficient than those operating under IRS or DRS. This is as expected because hospitals operating under IRS or DRS, by definition, cannot be scale efficient. From the results of the 3x1y model and the discussion relating to Table 4, it was expected that average scale efficiency scores for hospitals operating under IRS may be greater than those operating under DRS. This is observed for the 1x3y model in 2009, but is not the case in 2007 and 2008. The 3x3y model has no hospitals operating under IRS and therefore cannot be analysed. The results across models show some inconsistencies, which provide further support that that the results are sensitive to model specification. As an example, consider the average technical efficiency scores for the hospitals operating under each return to scale classification in 2009. In the 3x1y model, hospitals operating under CRS are most efficient; while hospitals operating under DRS are the most efficient in the 3x3y and 1x3y models.

The 3x3y model has the highest average efficiency scores. This is as expected because the 3x3y model has the highest number of input and output variables. This reduces the power of the DEA model to identify inefficient hospitals, thereby increasing the average efficiency scores of the model (Coelli *et al*, 2005). The 1x3y has the lowest average technical efficiency scores. Proceeding with caution regarding the direct comparison of efficiency scores across models, the low average technical efficiency scores of the 1x3y model may be driven by hospitals with more severe case-mixes that utilise a high number of nurses and pharmaceuticals as inputs. For example, consider a hospital that performs many surgeries or has a high number of ICU patients. As inputs into the production process, these hospitals will use a relatively high number of nurses and pharmaceuticals, and a relatively low number of beds. Note that the outputs, DRG adjusted number of cases, theatre minutes and number of billed days spent in hospital, will reflect the more severe case-mix of these hospitals. In the 1x3y model, it is likely that these hospitals will be operating efficiently as they will be producing a large number of outputs using a low number of inputs (a low number of beds). Relative to these hospitals, other hospitals will appear disproportionately inefficient, leading to lower average technical efficiency scores than the 3x1y or 3x3y models. This is also an example of how incorrect conclusions may be drawn by excluding important inputs of the production process – such as total number of nurses and billed pharmacy amount.

For a particular model, the standard deviations of technical efficiency scores are relatively stable across the three years; while the standard deviations of scale efficiency scores decrease across the three years. These decreases may be partly due to increasing average scale efficiency scores across the three years. This limits the extent to which deviations above this average can occur as efficiency scores cannot be greater than one. Similarly, the 3x3y model has the lowest standard deviation of efficiency scores which may be because it has the highest average efficiency scores. The 3x1y model, when compared to the 1x3y model, has a lower standard deviation of technical efficiency scores and a higher standard deviation of scale efficiency scores. Again this may be partly explained by average efficiency scores, as the 3x1y model has higher average technical efficiency scores and lower average scale efficiency scores than the 1x3y model.

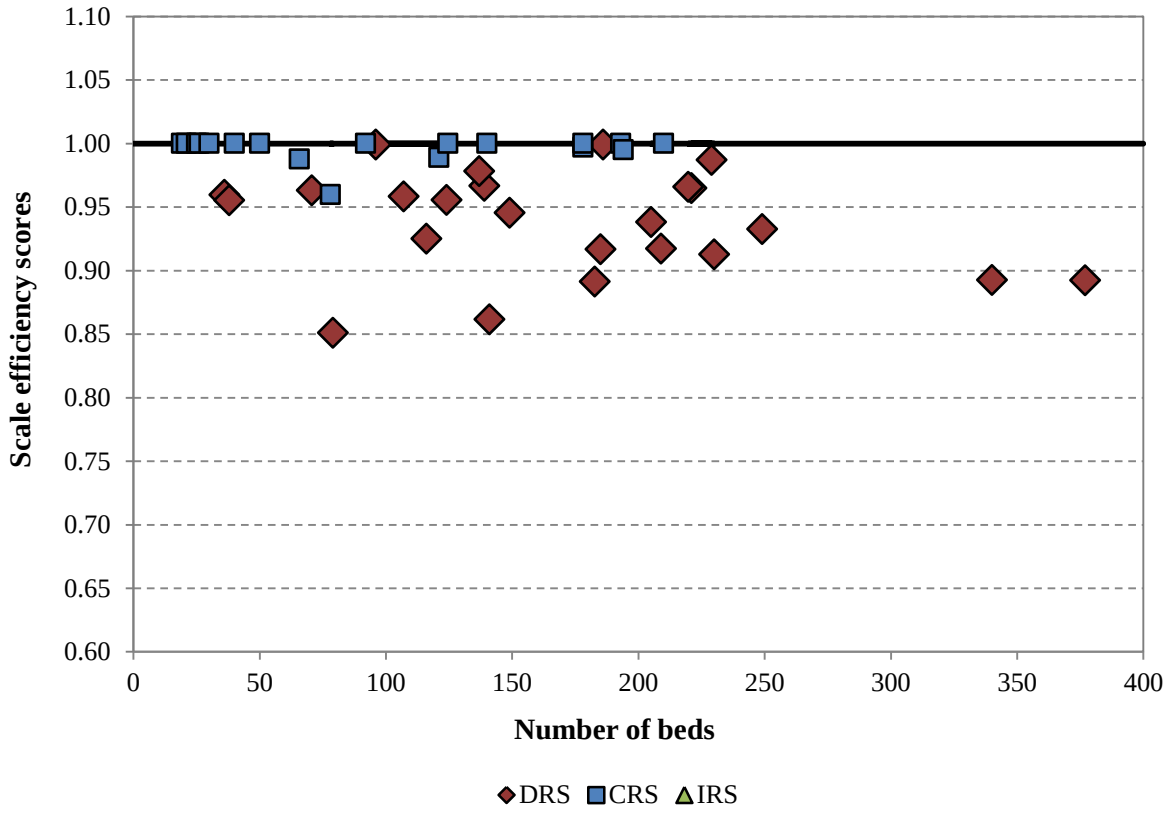

Figure 16: Scale efficiency and number of beds for each hospital in 2009 under the 3x3y model

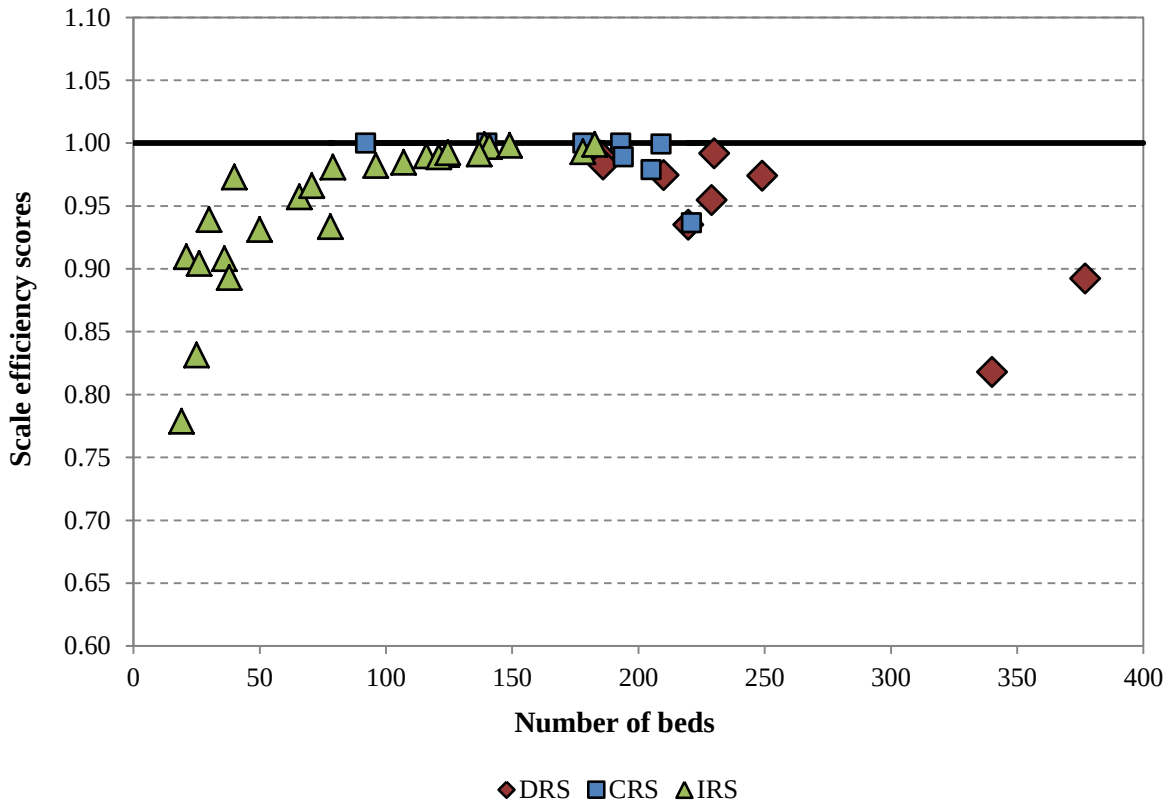

Figure 17: Scale efficiency and number of beds for each hospital in 2009 under the 1x3y model

Figures 16 and 17 display, for the 3x3y and 1x3y models respectively, the relationship between hospital size (measured in terms of number of beds), return to scale classification and scale efficiency in 2009. For comparative purposes the reader is directed to Figure 8, which displays the above information for the 3x1y model. These figures provide further details of the scale efficiency distributions that underlie the results presented in Tables 3, 10 and 11; namely the average scale efficiency scores and the standard deviations of these scores. It can also be seen in these figures that hospitals operating under CRS are indeed the most scale efficient.

The relationship between hospital size and return to scale classification under each of the three models in 2009 is provided in Table 12. Note that average number of beds has been used as a proxy for size.

Table 12: Average number of beds and return to scale classification in 2009 under each of the three models

Hospital number	Average number of beds	3x1y	3x3y	1x3y
1	19	IRS	CRS	IRS
2	20	IRS	CRS	IRS
3	25	IRS	CRS	IRS
4	26	IRS	CRS	IRS
5	30	CRS	CRS	IRS
6	36	IRS	DRS	IRS
7	38	IRS	DRS	IRS
8	40	CRS	CRS	IRS
9	48	IRS	CRS	IRS
10	54	IRS	CRS	IRS
11	65	CRS	DRS	IRS
12	71	DRS	DRS	IRS
13	78	IRS	CRS	IRS
14	91	CRS	CRS	CRS
15	96	DRS	DRS	IRS
16	101	DRS	DRS	IRS
17	107	DRS	DRS	IRS
18	116	DRS	DRS	IRS
19	121	IRS	CRS	IRS
20	124	DRS	CRS	IRS
21	132	DRS	DRS	IRS
22	137	DRS	DRS	IRS
23	139	IRS	CRS	CRS
24	139	DRS	DRS	IRS
25	149	DRS	DRS	IRS
26	178	CRS	CRS	IRS
27	178	CRS	CRS	CRS
28	185	DRS	DRS	IRS
29	185	DRS	DRS	DRS
30	186	DRS	DRS	DRS
31	193	CRS	CRS	CRS
32	194	CRS	CRS	CRS
33	200	DRS	DRS	CRS
34	205	DRS	DRS	CRS
35	206	CRS	CRS	DRS
36	224	DRS	DRS	DRS
37	225	DRS	DRS	CRS
38	226	DRS	DRS	DRS
39	230	DRS	DRS	DRS
40	249	DRS	DRS	DRS
41	342	DRS	DRS	DRS
42	377	DRS	DRS	DRS

Table 12 highlights the extent to which the return to scale classifications are model dependent. However, some consistency across the three models can be seen. For example, the largest hospitals all exhibit DRS across all three models. The consistencies between the return to scale classifications of the three models from 2007 to 2009 are displayed in Table 13.

For comparative purposes, it is easier to compare two models that have either the same input variables or the same output variables. This removes the complexity of comparing models with simultaneous changes in inputs and outputs. It is therefore easiest to compare 3x1y with 3x3y, and 3x3y with 1x3y. This is done in the following two paragraphs.

Compared to the 3x1y model, the 3x3y model has a greater proportion of hospitals operating under DRS, and no hospitals operating under IRS. Figure 14 shows that this relationship is true across all three years. As discussed in section 5.5, the two additional output variables of the 3x3y model, namely billed days spent in hospital and total number of theatre minutes, may result in double counting of some outputs. However, the increase in output due to double counting may not be the same for all hospitals. For example, a hospital may have a case-mix that has a relatively low number of surgical cases resulting in a low number of theatre minutes produced by that hospital. For this hospital, any double counting due to theatre minutes would result in a smaller increase in its output relative to the other hospitals. It is also possible that some efficient hospitals may produce less billed days and theatre minutes than inefficient hospitals because these are, to some extent, under the control of management. Therefore, the increase in outputs due to double counting for efficient hospitals may be less than that of inefficient hospitals. This could result in the scale of operations of efficient hospitals being relatively smaller than inefficient hospitals, leading to a greater number of hospitals operating under DRS. This is one possible explanation of why the 3x3y model has a greater proportion of hospitals operating under DRS than the 3x1y model, as well as no hospitals operating under IRS.

The 1x3y model has a large proportion of hospitals operating under IRS, while the 3x3y model has none. Compared to the 3x3y model, the 1x3y model does not include the two input variables of salary adjusted number of nurses, and total billed pharmacy amount. These two variables provide material insight into the production process of the hospitals, and their exclusion is expected to have a large impact on efficiency scores and return to scale classifications. This is indeed the case – the impact on efficiency scores can be seen in Table 14. Differences in efficiency scores arise, for example, when a hospital uses a comparatively large number of nurses and pharmaceuticals, and a comparatively small number of beds, to produce its outputs. This occurs, say, when hospitals treat a large number of ICU patients. These hospitals will appear more technically efficient in the 1x3y model than the 3x3y model. It is also likely that the number of beds, being the only input variable, has a much greater impact on the scale profile of the 1x3y model. This can be seen in Table 12, where all hospitals with a

small number of beds exhibit IRS. As the number of beds increases the hospitals exhibit a mix of IRS and CRS, then a mix of CRS and DRS, and finally only DRS. This is clearly shown in Figure 17, which displays the relationship between number of beds, return to scale classification, and scale efficiency in 2009 under the 1x3y model. Table 12 shows that there are a large number of hospitals in the 1x3y model that operate under IRS. However, it is unlikely that so many hospitals operate under IRS given the South African private hospital environment. Therefore, by only using number of beds as an input variable, it is likely that the 1x3y model does not sufficiently represent the hospitals' production dynamics.

Table 13 below identifies the hospitals that operate under the same return to scale classification across all three models for a particular year. There is consistency across the three models for some of the larger hospitals, measured in terms of number of beds, that operate under DRS. In particular, hospitals 38, 41 and 42 operate under DRS across all three models and years. There appears to be a lack of consistency across the three models for smaller hospitals. Furthermore, there are no hospitals that operate under IRS across all three models. This is because no hospitals in the 3x3y model operate under IRS in any of the three years. Additional return to scale consistency across the three models appears to be distributed between hospitals of different sizes.

Table 13: Hospitals that have the same return to scale classification across all three models in a particular year

Hospital number	Average number of beds	2007	2008	2009
1	19	-	-	-
2	20	-	-	-
3	25	-	-	-
4	26	-	-	-
5	30	-	-	-
6	36	-	-	-
7	38	-	-	-
8	40	-	-	-
9	48	-	-	-
10	54	CRS	-	-
11	65	CRS	-	-
12	71	-	-	-
13	78	-	-	-
14	91	CRS	-	CRS
15	96	-	-	-
16	101	-	CRS	-
17	107	-	-	-
18	116	DRS	-	-
19	121	-	-	-
20	124	-	-	-
21	132	-	DRS	-
22	137	-	-	-
23	139	-	CRS	-
24	139	-	-	-
25	149	-	-	-
26	178	-	-	-
27	178	-	CRS	CRS
28	185	-	-	-
29	185	-	-	DRS
30	186	CRS	CRS	DRS
31	193	-	-	CRS
32	194	-	-	CRS
33	200	-	DRS	-
34	205	-	-	-
35	206	-	DRS	-
36	224	-	-	DRS
37	225	-	-	-
38	226	DRS	DRS	DRS
39	230	-	-	DRS
40	249	-	-	DRS
41	342	DRS	DRS	DRS
42	377	DRS	DRS	DRS

Table 14: Technical efficiency scores and scale efficiency scores for each of the three models in 2009

2009		Technical efficiency			Scale efficiency		
Hospital number	**Average number of beds**	**3x1y**	**3x3y**	**1x3y**	**3x1y**	**3x3y**	**1x3y**
1	19	1	1	1	0.99	1	0.78
2	20	1	1	1	0.98	1	0.91
3	25	0.95	1	0.86	0.98	1	0.83
4	26	0.87	1	1	0.81	1	0.90
5	30	1	1	0.90	1	1	0.94
6	36	0.83	0.998	0.66	0.96	0.96	0.91
7	38	0.91	1	0.82	0.9995	0.96	0.89
8	40	1	1	1	1	1	0.97
9	48	0.96	1	0.95	0.98	1	0.93
10	54	0.95	0.96	0.88	0.97	0.99	0.96
11	65	0.87	0.98	0.88	1.00	0.96	0.97
12	71	1	1	0.76	0.85	0.85	0.98
13	78	0.36	0.36	0.36	0.95	0.96	0.93
14	91	1	1	1	1	1	1
15	96	1	1	0.82	0.97	0.999	0.98
16	101	0.98	1	0.87	0.92	0.96	0.99
17	107	0.95	0.98	0.87	0.96	0.96	0.98
18	116	0.96	0.99	0.86	0.93	0.93	0.99
19	121	0.87	0.89	0.89	0.99	0.99	0.99
20	124	0.89	1	0.83	0.999	1	0.99
21	132	1	1	0.76	0.92	0.97	0.999
22	137	0.91	0.95	0.90	0.99	0.98	0.99
23	139	0.90	1	1	0.99	1	1
24	139	0.80	1	0.78	0.73	0.86	0.997
25	149	0.80	0.86	0.80	0.95	0.95	0.998
26	178	0.77	0.83	0.77	0.999	0.997	0.99
27	178	0.92	1	1	0.96	1	1
28	185	0.67	0.80	0.70	0.87	0.89	0.999
29	185	0.90	0.99	0.88	0.85	0.92	0.99
30	186	1	1	1	0.87	0.999	0.98
31	193	1	1	1	1	1	1
32	194	0.80	0.88	0.83	0.9997	0.995	0.99
33	200	0.77	0.85	0.72	0.93	0.92	0.999
34	205	0.98	1	0.75	0.86	0.94	0.98
35	206	0.95	1	0.94	0.98	1	0.97
36	224	0.92	1	1	0.93	0.99	0.95
37	225	0.92	0.92	0.85	0.96	0.96	0.94
38	226	0.98	0.99	0.90	0.95	0.97	0.94
39	230	0.93	0.93	0.83	0.91	0.91	0.99
40	249	0.78	0.83	0.76	0.91	0.93	0.97
41	342	1	1	1	0.84	0.89	0.82
42	377	1	1	1	0.78	0.89	0.89

Table 14 shows the technical and scale efficiency scores for each hospital and model for 2009. Note that when rounding to two decimal places resulted in an efficiency score rounding up to one, thereby appearing efficient when it is not, that efficiency score was rounded to more than two decimal places.

In Table 14 it can be seen that all technically efficient hospitals under the 3x1y model are also technically efficient under the 3x3y model; and all scale efficient hospitals under the 3x1y model are also scale efficient under the 3x3y model. Similarly, all technically efficient hospitals under the 1x3y model are also technically efficient under the 3x3y model; and all scale efficient hospitals under the 1x3y model are also scale efficient under the 3x3y model. These results were verified for all three years, and highlight some of the consistencies across the three models. For technical efficiency, this occurs because a firm is only labelled as inefficient after all possible weights have been considered, and no other weights will provide a higher technical efficiency rating (Charnes *et al*, 1978). This means that, if extra variables are added to a model and these variables decrease the technical efficiency rating of a particular hospital, then DEA can assign zero weightings to the additional variables (Coelli *et al*, 2005). This gives the benefit of the doubt to the hospital (Sherman, 1984) and results in the set of technically efficient hospitals of the new model including at least the set of technically efficient hospitals of the original model. This also means that the technical efficiency scores of each of the hospitals in the new model will be at least as large as the technical efficiency scores of the original model. Note that this result does not necessarily apply to scale efficiency scores. For example, a hospital that is scale efficient in the original model will, by definition, also be technically efficient under both VRS and CRS production technologies. Therefore, in the new model, this hospital will remain technically efficient under both production technologies and will hence also remain scale efficient. However, the new scale efficiency score of a hospital that is scale inefficient in the original model is indeterminate, as it depends on the relative change in its technical efficiency scores calculated under both VRS and CRS production technologies. The above discussion explains why there is overlap between efficient hospitals in the 3x3y model and efficient hospitals in the 3x1y or 1x3y models. It also explains, to some extent, why models with a greater number of variables have higher average technical efficiency scores than those with fewer variables.

The 3x3y model has a large number of small hospitals (measured in terms of number of beds) that are scale efficient. However, it is expected that very small hospitals would benefit, at least to some extent, from economies of scale. Therefore it seems unlikely that so many small hospitals are scale efficient. Therefore, the scale efficiency profiles of 3x1y and 1x3y appear to provide a more realistic representation.

Hospitals 6, 12 and 21 have much lower technical efficiency scores under the 1x3y model than under the 3x1y and 3x3y models. This is also true for 2007 and 2008. On further examination, it was found that these hospitals use a relatively low number of inputs in the form of nurses and pharmaceuticals

when compared to other hospitals with a similar number of beds. It is likely that the efficient use of these two inputs (total number of nurses and billed pharmacy amount) drives the relatively high efficiency of these hospitals and once they are excluded, as is the case in the 1x3y model, these hospitals appear to operate with lower efficiency. However, caution must be applied when comparing efficiency scores across different models as they represent relative, not absolute, efficiency estimates. Note that hospital 13 is a specialist hospital and has the lowest technical efficiency score across all three models and years.

Table 15: A count of the total number of efficient hospitals in each hospital's ERS for each model

Number of hospitals with the following ERS:	**2007**			**2008**			**2009**		
	3x1y	**3x3y**	**1x3y**	**3x1y**	**3x3y**	**1x3y**	**3x1y**	**3x3y**	**1x3y**
ERS contains 1 hospital	12	25	13	15	25	12	12	24	12
ERS contains 2 hospitals	16	0	10	13	2	11	11	2	6
ERS contains 3 hospitals	13	6	16	8	4	13	16	3	15
ERS contains 4 hospitals	1	8	3	6	5	6	3	3	9
ERS contains 5 hospitals	0	3	0	0	5	0	0	6	0
ERS contains 6 hospitals	0	0	0	0	1	0	0	4	0
Average number in ERS	2.1	2.1	2.2	2.1	2.2	2.3	2.2	2.5	2.5

It can be seen in Table 15 that, for a particular model, the average number of hospitals in each hospital's ERS (the average number of peer hospitals) is relatively stable across the three years. The distribution of the number of peer hospitals is relatively similar for the 3x1y and 1x3y models, but different for the 3x3y model. Most hospitals in the 3x1y and 1x3y models have three or less peer hospitals, and none have more than four; while the majority of 3x3y hospitals have one peer hospital, and some have more than four.

Note that a hospital must be technically efficient in order to be a peer hospital and that a technically efficient hospital only has one peer hospital (itself). The 3x3y model has the highest number of variables and therefore the lowest power to identify inefficient hospitals. This results in the 3x3y model having the largest number of technically efficient hospitals, which is the reason why the 3x3y model has the largest number of hospitals with one peer. In fact, in Table 15, all hospitals with one peer are technically efficient hospitals; however this need not be the case.

The 3x3y model has some hospitals with more than four peers. This does not occur in the 3x1y or 1x3y models. A possible reason for this is that the greater number of technically efficient hospitals in the 3x3y model may make it easier to find a linear combination of hospitals that projects a particular inefficient hospital onto the frontier. This linear combination will better project the hospital onto the frontier if it combines the production dynamics of as many hospitals as possible that lie on the frontier surrounding the inefficient hospital. Since more hospitals lie on the frontier of the 3x3y model, it is

likely that a greater number of peer hospitals will be found in the ERS of each inefficient hospital. Another possible reason for this is that the double counting of some outputs in the 3x3y model may exacerbate the operational differences between hospitals. The reader is directed to section 5.5 for further information regarding the possible double counting of outputs in the 3x3y model. Note that the ERS of a hospital represents the set of efficient hospitals that have operations most similar to itself, and that a linear combination of these hospitals will project the inefficient hospital onto the production frontier. However, greater operational differences between hospitals mean that there will be fewer hospitals that are closely similar. Since potential peer hospitals are less similar, more of these hospitals may be required in order to project an inefficient hospital onto the frontier.

6.4. Analysis of selected hospitals from the 3x1y model

This section examines the scale dynamics of five individual hospitals, namely hospitals 3, 16, 18, 28 and 41. These hospitals were selected because they exhibit interesting characteristics.

Table 16 shows the number of technically efficient and scale efficient hospitals for each year under the 3x1y model. In this section, all hospitals that are technically efficient or scale efficient have been assigned the rank of one. The inefficient hospital with the highest efficiency score is then assigned the rank equal to the number of efficient hospitals plus one. For example, out of the set of technically inefficient hospitals, the most efficient of these hospitals would be assigned the rank of 13 in 2009. Similarly, the most inefficient hospital would be assigned the rank of 42. Table 16 is provided as a reference to better understand the rankings of the individual hospitals analysed in this section.

Table 16: The number of technically efficient and scale efficient hospitals according to the 3x1y model

	2007	**2008**	**2009**
Number of technically efficient hospitals	12	15	12
Number of scale efficient hospitals	8	6	4

6.4.1. Analysis of hospital 3

Hospital 3 is interesting because it operates under a different return to scale classification in each of the three years. This can be seen in Table 17.

Table 17: Efficiency and return to scale information for hospital 3 under the 3x1y model

Hospital 3	2007	2008	2009
Returns to scale	CRS	DRS	IRS
ERS count	3	3	2
Technical efficiency (VRS)	0.977	0.995	0.953
Scale efficiency	0.998	0.993	0.978
Technical efficiency rank	15	16	18
Scale efficiency rank	13	15	14

The different return to scale classifications of hospital 3 can be better understood by examining its ERS. This is shown in Table 18. The ERS contains three hospitals in 2007 and 2008, and two in 2009.

Table 18: Information regarding the hospitals contained in the ERS of hospital 3 under the 3x1y model

2007			2008			2009		
ERS	**Scale**	**Lambda weights**	**ERS**	**Scale**	**Lambda weights**	**ERS**	**Scale**	**Lambda weights**
Hospital 1	CRS	32%	Hospital 2	CRS	70%	Hospital 1	IRS	62%
Hospital 2	CRS	19%	Hospital 5	CRS	6%	Hospital 5	CRS	38%
Hospital 5	CRS	49%	Hospital 6	DRS	23%			

In 2007, all three hospitals in the ERS operate under CRS, which results in hospital 3 also being classified as operating under CRS. The reader is directed to section 4.4 for details regarding the relationship between a hospital's ERS and its return to scale classification.

Hospital 6 is included in the ERS in 2008, where it exhibits DRS, but is not included in 2007 or 2009. It should be noted that hospital 6 is not technically efficient in either 2007 or 2009 and therefore cannot form part of any hospital's ERS. Hospital 6 drives the change in return to scale classification of hospital 3 from CRS in 2007 to DRS in 2008. Similarly, hospital 1 is included in the ERS of hospital 3 in 2009, where it exhibits IRS. This drives the IRS classification of hospital 3 in 2009. Note that hospital 1 is included in the ERS of hospital 3 in 2007 where it operates under CRS. In 2008, hospital 1 is technically efficient and operates under DRS but does not form part of the ERS of hospital 3.

Table 18 also provides the lambda weights assigned to each hospital in the ERS of hospital 3. A linear combination of the inputs and outputs of the hospitals in the ERS, calculated using the lambda weights, results in a projection of hospital 3 onto the efficient frontier. As such, the weights give an indication of the closeness of each hospital in the ERS to hospital 3. In 2008, the lambda weighting assigned to hospital 6, which exhibits DRS, was only 23%. This implies that the extent to which

hospital 3 exhibits DRS in 2008 may be relatively low. However, in 2009 a 62% weighting was assigned to hospital 1 which exhibits IRS. This implies that the extent to which hospital 3 exhibits IRS in 2009 may be relatively high. Hospital 3 is the third smallest hospital measured in terms of average number of beds, average billed pharmacy amount, or average DRG adjusted number of cases – all averaged across the three years. It is also the fourth smallest hospital in terms of average salary adjusted number of nurses. Therefore, given its small size, hospital 3 could be expected to operate under IRS. However, this is only the case in 2009.

Even though the return to scale classification of hospital 3 is not stable across the three year period, it is still able to operate with relatively high technical and scale efficiency. The efficiency scores and rankings of hospital 3 are shown in Table 17 and should be considered in conjunction with Table 16 which shows the number of efficient hospitals in each year.

6.4.2. Analysis of hospital 16

Hospital 16 is interesting because it changes from operating under CRS in 2007 and 2008 to operating under DRS in 2009. It is also technically and scale efficient in 2007 and 2008, but not in 2009. This can be seen in Table 19.

Table 19: Efficiency and return to scale information for hospital 16 under the 3x1y model

Hospital 16	2007	2008	2009
Returns to scale	CRS	CRS	DRS
ERS count	1	1	3
Technical efficiency (VRS)	1	1	0.981
Scale efficiency	1	1	0.917
Technical efficiency rank	1	1	14
Scale efficiency rank	1	1	31

Table 20: The inputs, output and occupancy rates of hospital 16 under the 3x1y model

Year	Inputs			Output	Occupancy rate
	Number of beds	Nurses	Billed pharmacy	DRG adjusted number of cases	
2007	90	115	22,810,997	10,407	0.77
2008	90	121	25,359,368	10,002	0.81
2009	124	143	30,947,995	9,802	0.72

Since hospital 16 operates under CRS in 2007, it is expected that an increase in inputs would result in a proportional increase in output. Table 20 shows that, in 2008, inputs increased without a

proportional increase in output. In fact output decreased slightly in 2008. However, hospital 16 remained efficient and operating in the CRS region of the frontier. The increase in occupancy rate in 2008 could have contributed to keeping hospital 16 efficient. However, in 2009, there has been a greater increase in inputs and decrease in output than in 2008. This has occurred to such an extent that the hospital is no longer technically or scale efficient and is now operating under DRS, which means that the hospital is now operating with excess capacity. The reduced occupancy rate in 2009 also supports this. Hospital 16 would now benefit by reducing its scale by reducing its number of inputs, which should result in a less than proportional reduction in output. However, the increase in the scale of operations in 2009 could have been a result of actions taken by management to meet business needs. In 2008, hospital 16 possessed the highest occupancy rate out of all 42 hospitals. The occupancy rate was also high relative to international best practice, which recommends that occupancy rates should not exceed 85% as this compromises infection control and the ability to cope with emergencies (Keegan, 2008). It is therefore likely that management took active steps to increase the scale of operations of hospital 16, thereby expanding its capacity and decreasing its occupancy rate. This resulted in the occupancy rate ranking of hospital 16 reducing to eleventh place in 2009. The fact that hospital 16 shifts to operating under DRS in 2009 suggests that the expansion in capacity was beyond what was necessary to meet the current demand. This could have been a strategic decision because, when undertaking an expansion project, it may be appropriate to expand beyond current needs with the intention of being able to meet expected future demand.

Note that, from Table 9, it can be seen that the ERS of hospital 16 in 2009 consists of three hospitals which all operate under DRS, namely hospitals 12, 15 and 21. This results in hospital 16 also operating under DRS.

6.4.3. Analysis of hospital 18

Hospital 18 is interesting because it exhibits the same return to scale classification across all three years, but loses its technical efficiency in 2009. It is also not scale efficient in any year. This can be seen in Table 21.

Table 21: Efficiency and return to scale information for hospital 18 under the 3x1y model

Hospital 18	2007	2008	2009
Returns to scale	DRS	DRS	DRS
ERS count	1	1	3
Technical efficiency (VRS)	1	1	0.964
Scale efficiency	0.851	0.852	0.933
Technical efficiency rank	1	1	16
Scale efficiency rank	35	36	27

In 2007 and 2008, hospital 18 is technically efficient and therefore only has one hospital in its ERS (itself). The return to scale classification of a technically efficient hospital is defined in terms of the region of the frontier in which it operates; while the return to scale classification of an inefficient hospital is defined in terms of the peer hospitals in its ERS. The reader is directed to section 4.4 for a discussion of this. When hospital 18 becomes technically inefficient in 2009, the number of peer hospitals in its ERS increases to three. From Table 9, it can be seen that the ERS in 2009 consists of three hospitals which all operate under DRS, namely hospitals 12, 15 and 21. This means that when hospital 18 moves off the frontier in 2009 and is no longer technically efficient, it still operates under the same return to scale classification as in 2007 and 2008 (namely DRS). This shows consistency between the scale dynamics used to classify returns to scale of technically efficient and technically inefficient hospitals.

6.4.4. Analysis of hospital 28

Hospital 28 is interesting as it operates with low technical efficiency and relatively low scale efficiency across all three years. This can be seen from the technical and scale efficiency rankings, which are shown in Table 22.

Table 22: Efficiency and return to scale information for hospital 28 under the 3x1y model

Hospital 28	2007	2008	2009
Returns to scale	CRS	DRS	DRS
ERS count	3	3	3
Technical efficiency (VRS)	0.540	0.570	0.670
Scale efficiency	0.953	0.902	0.870
Technical efficiency rank	41	41	41
Scale efficiency rank	25	30	35

Hospital 28 provides an opportunity to investigate the potential resource savings that can be achieved through the improvement of technical and scale efficiency. The technical efficiency scores in Table 22 show that, in theory, hospital 28 could have produced the same level of output by using up to 46.0% less inputs in 2007, 43.0% less in 2008, and 33.0% less in 2009. Similarly, the scale efficiency scores show that hospital 28 could have produced the same level of output by using up to 4.7% less inputs in 2007, 9.8% less in 2008, and 13.0% less in 2009. The inputs used by hospital 28 from 2007 to 2009 are displayed in Tables 23, and the potential input savings in 2009 are displayed in Table 24. The discussion below these tables focuses on the potential input savings that could have been achieved through efficiency improvements in 2009.

Table 23: The inputs, output and occupancy rates of hospital 28 under the 3x1y model

Year	Inputs			Output	Occupancy rate
	Number of beds	Nurses	Billed pharmacy	DRG adjusted number of cases	
2007	186	179	39,868,372	9,146	0.60
2008	186	170	39,861,359	8,828	0.62
2009	183	153	40,988,138	9,082	0.58

Table 24: The potential input savings arising from efficiency improvements for hospital 28 in 2009 under the 3x1y model

2009	Potential input savings		
	Number of beds	Nurses	Billed pharmacy
Technical efficiency savings	60	50	13,507,920
Scale efficiency savings	24	20	5,315,996

The potential input savings in 2009 represent a 33.0% saving from technical efficiency improvements and a 13.0% saving from scale efficiency improvements. This means that if the total output produced by hospital 28 is maintained then, under the assumption of a VRS production technology, a 33.0% reduction in inputs would project hospital 28 onto the VRS efficient frontier. Under the assumption of a CRS production technology, a further 13.0% reduction in inputs would project hospital 28 onto the CRS efficient frontier. Such a hospital would be technically efficient, scale efficient, and operate under CRS at the MPSS.

The potential resource savings in 2009 amount to 84 beds, 70 nurses and R18,823,916 in pharmaceutical supplies. This represents a significant quantity of inputs. If output could be kept at the current level and the number of inputs reduced, for hospital 28 and all other hospitals in this investigation, it would greatly increase the efficiency with which these hospitals provide healthcare services. However, there would be practical constraints and other considerations that must be taken into account when implementing scale improvements. The reader is directed to section 7.2 for a discussion of this.

6.4.5. Analysis of hospital 41

Hospital 41 is interesting because it is technically efficient across all three years, but operates with low scale efficiency.

Table 25: Efficiency and return to scale information for hospital 41 under the 3x1y model

Hospital 41	2007	2008	2009
Returns to scale	DRS	DRS	DRS
ERS count	1	1	1
Technical efficiency (VRS)	1	1	1
Scale efficiency	0.728	0.786	0.841
Technical efficiency rank	1	1	1
Scale efficiency rank	42	40	39

From the rankings in Table 25, it can be seen that hospital 41 has the lowest scale efficiency score in 2007, third lowest in 2008, and fourth lowest in 2009. This means that hospital 41 could benefit from large resource savings by improving its scale efficiency. In theory, hospital 41 could have produced the same level of output by using up to 27.2% less inputs in 2007, 21.4% less in 2008, and 15.9% less in 2009. Again, the reader is directed to section 7.2 for a discussion of the practical difficulties of achieving this.

Hospital 41 exhibits DRS across all three years which implies that its low scale efficiency is consistently driven by it operating at a scale that is too large. In terms of size, hospital 41 is the second largest hospital measured in terms of average number of beds, average salary adjusted number of nurses, average billed pharmacy amount, or average DRG adjusted number of cases – all averaged across the three years. Therefore, given its large size, hospital 41 could be expected to operate under DRS, which is indeed the case in each of the three years.

Scale efficiency can be defined as the ratio of two technical efficiency scores. More precisely, it can be defined as the ratio of the technical efficiency score that is calculated under the assumption of a CRS production technology, to the technical efficiency score that is calculated under the assumption of a VRS production technology. The reader is directed to section 3.8 to the equation specified in (6) for a formal definition of scale efficiency. Table 25 shows that the technical efficiency scores for hospital 41, calculated under a VRS production technology, are equal to one across all three years. By the above definition, this means that the scale efficiency scores, also shown in Table 25, are equal to the technical efficiency scores calculated under a CRS production technology. Since hospital 41 has relatively low scale efficiency scores, it would also have relatively low technical efficiency scores when calculated under CRS production technology. The large differences between the technical efficiency scores of hospital 41, when calculated under VRS and CRS production technologies, may be due to its large size. Hospital 41 is the second largest hospital and may therefore be considered an extreme case. Under the assumption of a VRS production technology, hospital 41 could appear to operate on the efficient frontier simply because it lacks other hospitals of a similar size that it can be compared against. Hospital 41 can be interpreted as operating at the edge of the DRS portion of the

VRS efficient frontier. However, under the assumption of a CRS production technology, all hospitals will be compared against technically efficient hospitals that, by definition, operate under CRS. These hospitals will also be scale efficient and operate at the MPSS. Compared to these hospitals, hospital 41 appears to operate with low technical efficiency, which translates into low scale efficiency under the assumption of a VRS production technology. This is a possible explanation of why hospital 41, under the assumption of a VRS production technology, is technically efficient but operates with low scale efficiency across all three years.

7. Conclusions and recommendations for further research

7.1. Conclusions

This paper attempts to contribute to a limited body of research by using DEA to examine the relationship between scale and efficiency within a set of South African private hospitals. An understanding of this relationship is important as it is needed to determine whether scale inefficiencies exist, and whether they are able to be addressed. DEA was used in this investigation as a tool to measure efficiency and classify hospitals as operating under IRS, CRS or DRS. Additionally, DEA provided other useful management information, such as identifying best practice hospitals and quantifying potential resource savings.

The results of three different DEA models were examined in this paper. These results were based on data drawn from a set of South African private hospitals for the three year period from 2007 to 2009. As such, the results of this investigation may not necessarily be relevant for more recent, or future, time periods. Of the three DEA models, the 3x1y model was expected to be the most representative of the hospital production process and was therefore the focus of this investigation.

Under the 3x1y model, the results of the investigation show that scale is an area where efficiency improvements are likely to be possible. The average scale efficiency scores of the 3x1y model indicate that the set of hospitals, on average, could have produced the same level of output by using 6.9% less inputs in 2007, 6.8% less in 2008, and 6.2% less in 2009. However, the extent to which these theoretical savings could be realised in practice is likely to be limited. In addition to scale efficiency savings, the average technical efficiency scores indicate that technical efficiency savings are also possible.

The return to scale classification of each hospital was determined relative to the other hospitals in the dataset and not an absolute standard. Under the 3x1y model, the results of this investigation show that most hospitals operate under CRS in 2007 and 2008, and DRS in 2009. Across all three years the majority of hospitals operate under non-increasing returns to scale. Out of the set of technically efficient hospitals, most hospitals operate under CRS in 2007, and DRS in 2008 and 2009. It was expected that the proportion of hospitals operating under each return to scale classification would remain relatively stable across the three years, unless there have been significant changes to the production process. However, this was not the case. A possible explanation of this variability could be that a large proportion of hospitals are operating close to the point where their return to scale classifications change. This could cause their classifications to oscillate from year to year.

Additionally, the average annual occupancy rates of the set of hospitals are relatively low, ranging from 63.1% in 2007 to 65.6% in 2009. This, together with the result that most hospitals operate under non-increasing returns to scale, reinforces the general criticism that excess capacity exists within the South African private hospital industry. However, from the perspective of a South African private hospital, excess capacity may be appropriate given the operational goals of private hospital organisations and the nature of their ownership. Private hospitals provide healthcare services to medical scheme patients and patients that pay directly for their services. These patients expect to be treated promptly without waiting for treatment capacity to become available. From a business perspective, the expectations and satisfaction of these patients are of central concern when setting hospital capacity levels. Reducing excess capacity will increase the scale efficiency of a hospital but may have a negative impact on the demand for its healthcare services. For example, long waiting periods may alienate patients who then seek healthcare from other providers. This creates an incentive for private hospitals to err on the side of being too large rather than being too small. Therefore, hospitals may be more likely to exhibit DRS than IRS, which is consistent with the results of this investigation. It could even be expected that most hospitals would operate under DRS, as was the case in 2009.

The result that hospitals are more likely to operate under DRS than IRS is consistent with the findings of Kibambe & Koch (2007) and Zere *et al* (2001). Kibambe & Koch (2007) found that public hospitals in Gauteng are more likely to operate under DRS than IRS. They proposed that these hospitals may operate under DRS due to the emigration of medical professionals or the need to hold excess capacity in order to cope with potential medical catastrophes. However, these reasons may not be applicable to the private sector. For example, the ability to cope with wide-spread medical catastrophes may be seen as a function that should be fulfilled by the state rather than by private hospitals. With regard to emigration, higher earnings and better working conditions in the private sector may make private hospitals less susceptible to emigration than public hospitals. However, some drivers of emigration are common to both the private and public sectors. For example, a survey conducted by Arnold & Lewinsohn (2010) found that the most common reason for the emigration of South African doctors to Australia from 1990 onwards was a concern over the level of violent crime in South Africa. As skilled medical professionals emigrate, hospital management may struggle to fully staff their hospitals. Consequently, hospitals may not have sufficient medical professionals to operate at their optimal capacity. This could lead to an oversupply of other inputs. For example, beds may become too numerous to be attended to by the remaining nursing staff. It is through this mechanism that emigration may contribute to hospitals operating under DRS.

Zere *et al* (2001) found that approximately 50% of public hospitals in the Northern, Eastern and Western Cape exhibited DRS, 13% exhibited CRS, and 37% exhibited IRS. Their finding that

hospitals operating under DRS and CRS are more common than those operating under IRS is consistent with the results of this investigation. However caution must be applied when comparing results from the public and private sectors, as this may not be appropriate.

Under the 3x1y model, this investigation found that smaller hospitals, when measured in terms of number of beds, are more likely to operate with higher technical and scale efficiency. Furthermore, it was found that the majority of scale efficient hospitals are smaller hospitals. There was also evidence that scale efficiency tends to decrease with increasing number of beds. One possible explanation of why smaller hospitals exhibit higher technical efficiency than larger hospitals is because smaller hospitals, with fewer inputs, may be easier to manage in an efficient manner. Another possible explanation is that smaller hospitals may have fewer specialised units and may treat cases that are simpler and more homogeneous than those treated by larger hospitals. This is supported by the observation that smaller hospitals tend to have lower DRG case-mix adjustment factors than larger hospitals. Management of smaller hospitals may therefore be able to improve their technical efficiency, relative to larger hospitals, by focusing their attention on the efficient treatment of frequently occurring, simpler cases. A possible explanation of why smaller hospitals have higher scale efficiency is because of the incentive for private hospitals to operate with excess capacity. This may lead smaller hospitals, which are the most likely to operate under IRS, to increase the scale of their operations bringing these hospitals closer to the region of CRS which, by definition, contains the most scale efficient hospitals. However, hospitals operating under DRS already possess excess capacity and would therefore not have the same incentive to shift their operations closer to the region of CRS. This would result in smaller hospitals, operating under IRS, being on average more scale efficient than larger hospitals operating under DRS.

These results are consistent with Zere *et al* (2001) who found that smaller public hospitals were relatively more scale efficient than larger public hospitals. Again, caution must be applied when comparing results from the public and private sectors.

In order to gain further insight, the results of the 3x1y model were compared with the results of two other models, namely the 3x3y and 1x3y models. This comparison identified various inconsistencies across the three models leading to the conclusion that the investigation is sensitive to the selection of input and output variables. Model specification is therefore important as it has a significant impact on the results and therefore the conclusions of this investigation. There is a trend in the DEA literature to focus on model specification in terms of technical efficiency rather than scale efficiency and return to scale classification. This investigation highlighted that model specification also has a significant impact on scale efficiency and return to scale classification. However, it should be noted that the results were not entirely inconsistent across the three models. For example, except for the 1x3y model in 2009, the majority of hospitals exhibited non-increasing returns to scale across all models and

years. This provides further support for the claim that excess capacity exists within South African private hospitals.

Sherman & Zhu (2006) claim that scale issues can be easily oversimplified or misunderstood. It was shown in this investigation that simplified approaches to determining scale can lead to incorrect conclusions. For example, a simple approach may be to rank hospitals by number of beds or occupancy rates, and then label the regions of IRS, CRS, and DRS according to these rankings. Given the current dataset, these types of approaches would have led to incorrect conclusions regarding scale. It was noted that a relatively large number of hospitals in the 1x3y model operate under IRS in 2009. However, it is unlikely that so many hospitals operate under IRS given the South African private hospital environment. Therefore, by only using number of beds as an input variable, it is likely that the 1x3y model does not sufficiently represent the dynamics of the hospital production process. The 3x1y model, which is the focus of this investigation, improves on these types of simplified analyses by using a greater number of factors to analyse scale, thereby capturing more aspects of the production process.

The hospitals identified in this investigation as operating under DRS would benefit from reducing the scale of their operations. In these hospitals, a reduction in inputs should result in a less than proportional reduction in outputs. Under the 3x1y model, a reduction in inputs could be achieved by reducing one or more of the following: the number of beds, the number of nurses, or the quantity of pharmaceuticals used to treat patients. However, in addition to these inputs there are other non-modelled inputs that would also need to be reduced, such as the number of doctors or the amount of equipment used in the hospital production process. Furthermore, some inputs would not easily be reduced or may be subject to operational constraints. For example, it may not be possible in the short term to reduce the flow of capital services due to, say, the indivisible nature of hospital buildings. However, it should be noted that all inputs, including capital inputs, are variable in the long run. Expenses that are proportional to these inputs, such as building maintenance expenses, will also be difficult to reduce in the short term. Since the hospitals that operate under DRS possess excess capacity, the current levels of demand for their healthcare services would still be able to be met if these hospitals reduced their scale of operations. This would reduce the supply of healthcare services and the excess capacity of these hospitals. However, as discussed above, excess capacity may be an operating requirement of South African private hospitals.

Similarly, the hospitals identified in this investigation as operating under IRS would benefit from increasing the scale of their operations. In these hospitals, an increase in inputs should result in a more than proportional increase in outputs. However, the current levels of demand may not be able to support operations at a larger scale. Furthermore, management typically cannot influence the demand for healthcare services, which may limit the extent that hospitals operating under IRS can benefit from

scale improvements. It is arguably easier to exploit scale efficiencies present in hospitals operating under DRS than IRS, since supply is more controllable than demand. There are also additional considerations that must be taken into account when implementing scale adjustments. These are discussed briefly in the next section.

Hospitals can be examined at an individual level in order to identify drivers of scale inefficiency and determine how this could be improved upon. In this investigation, individual analysis was performed for five hospitals with interesting characteristics. This provided insight into their operations and scale inefficiencies. Even if it is not possible for management to improve scale efficiency in the short term, or if this is restricted by current levels of demand, it is still important that they understand the relationship between scale and efficiency and how this impacts their operations. This would allow management to identify best practice hospitals, quantify potential resource savings and set goals for long term scale efficiency improvements.

From a public policy perspective, DEA can help policymakers make more informed decisions regarding the allocation of healthcare resources. The Department of Health (2011b) has indicated that reducing the high costs of private healthcare is critical to the success of NHI. Addressing scale inefficiencies could assist with reducing the relatively high costs of private healthcare. DEA could provide a method whereby the Department of Health could measure and assess scale efficiency within the private sector. If this is implemented through consultation and partnership with the private industry, it could lead to better public-private collaboration and ultimately better health outcomes.

However, further research into the relationship between scale and efficiency is needed in order to better understand its impact on South Africa's healthcare system. In particular, research into the implementation of scale improvements is required. This, and other possible areas of further research, are discussed in the next section.

7.2. Recommendations for further research

As discussed in section 4.7.4, the measurement of hospital efficiency, particularly using frontier models, is an under-researched area of investigation within Sub-Saharan Africa (Zere *et al*, 2001). Therefore, any investigation into hospital efficiency and scale within a South African context is likely to contribute to the understanding of these issues. This paper attempts to examine the relationship between scale and efficiency within a set of South African private hospitals. However, there is still large scope to contribute to the understanding of this relationship, and the relationship between hospital scale and efficiency within South Africa more generally.

Further research could be conducted by extending the investigation outlined in this paper. For example, the investigation could be extended to incorporate more recent data, if these are made

available for research purposes. Another possible extension could involve the application of Malmquist productivity indices to analyse changes in efficiency over time, with a particular focus on scale efficiency. It would also be possible to use the current dataset to perform Stochastic Frontier Analysis, and compare the results with the results of this investigation.

This paper accounts for case-mix differences by using DRG case-mix adjustments, and assumes that quality of care is consistent across hospitals. Further research could involve an investigation into the different methods of adjusting for case-mix differences between hospitals; as well as an investigation into the different methods of adjusting for quality. This research would aid efficiency measurement by providing further insight into the hospital production process. Again, these investigations would be dependent on the availability of appropriate data.

This paper conducts sensitivity testing by examining the results of three DEA models, each specified by a different combination of input and output variables. Given the importance of sensitivity testing, and the relatively large impact of adding or removing variables when the number of variables specified in the model is small, additional sensitivity testing may be warranted. For example, further research could include running a particular DEA model multiple times, each time removing one of the efficient hospitals and calculating rank correlation coefficients. These coefficients could then be used to determine to what extent the rankings are consistent. This would facilitate the identification of outliers and provide further credibility to the results of the investigation.

With regard to hospital occupancy rates, international best practice is to keep occupancy rates below 85% in order to facilitate infection control and cope with emergencies (Keegan, 2008). It would be reasonable to assume that South African private hospitals, catering for paying patients, would adhere to at least this minimum standard. Indeed, as discussed in section 6.2, excess capacity is expected to be larger than this to ensure that paying patients receive prompt treatment without waiting for treatment capacity to become available. An interesting piece of further research could involve quantifying the extent of the excess capacity that is required in order to meet surges in hospital admissions. This could be achieved by examining the distribution of occupancy rates over time. Additionally, spikes in admissions may need to be analysed as a separate stochastic process. This could provide support, or otherwise, for the excess capacity within the private sector. It could also assist in determining private hospital best practice with regard to occupancy rates. From a policy perspective, the results of such an investigation could be used to inform NHI debate.

The results and conclusions of this investigation are based on a dataset sourced from a single private hospital provider and, as such, may not be representative of the South African private hospital industry as a whole. Further research could be conducted by extending this investigation to include a dataset sourced from multiple private hospital providers that covers a greater proportion of the

industry. This type of investigation could be used for benchmarking and informing policy debate. Benchmarking would allow the industry to identify best practice and to better understand the drivers of scale efficiency. It could also provide a quantitative means of ranking the efficiency of each private hospital, which could be used as the basis for contracting with the NHI. However, the data from each provider would need to be consistent and sufficiently homogeneous in order to perform a meaningful comparison.

This investigation cannot be used to draw conclusions regarding public sector hospitals, as these hospitals differ significantly from private hospitals in terms of operating constraints and objectives. Ideally, the methodology adopted in this paper could be applied to a sample of public hospitals. The results could be used by the Department of Health and hospital management for benchmarking and improving public hospital performance, particularly within an NHI environment. Importantly, these results could also be used to inform policy decisions. For example, an increased understanding of public hospital scale dynamics would be useful when determining the size of a new hospital, or when deciding how many hospitals should be located within a particular geography. Benchmarking could also be used, say, to design performance incentives for public hospitals' management. However, as documented by Kibambe & Koch (2007), the dataset necessary to conduct this type of investigation is unlikely to be available due to a lack of appropriate information systems and human resources within public sector hospitals.

This paper examines the relationship between hospital scale and efficiency. However, it does not examine the practical implementation of scale improvements. This is a complex issue that could form the basis for further research. Such an investigation could examine the feasibility of implementing scale changes, which is likely to be subject to long implementation periods and operational constraints. These constraints could include, *inter alia*, the indivisible nature of some capital inputs, or a dependency on management cooperation. Furthermore, the potential benefits of scale improvements will be limited by the supply and demand of healthcare services. If a hospital is private, market research into the impact of scale changes on its competitive position would be necessary. Within the South African private hospital environment, this may require a greater understanding of excess capacity within the industry and whether this excess capacity is necessary. Additionally, Hollingsworth (2008) suggests that scale improvements should be accompanied by an investigation into whether the freed resources could to be reallocated to other more efficient activities.

Hospital efficiency, and the relationship between scale and efficiency, will become increasingly important as South African moves to an NHI environment. Further research and understanding is needed in order to better shape the future of South African healthcare.

References

African National Congress. 2010. ANC National General Council 2010 Additional Discussion Documents (Health). [Online]. Available: http://www.anc.org.za/docs/discus/2010/aditionalo.pdf [8 February 2013].

Anderson, P. & Peterson, N. 1993. A procedure for ranking efficient units in Data Envelopment Analysis. *Management Science* 39: 1261-1264.

Arnold, P.C. & Lewinsohn, D.E. 2010. Motives for migration of South African doctors to Australia since 1948. *The Medical Journal of Australia* 192 (5): 288-290. [Online]. Available: https://www.mja.com.au/journal/2010/192/5/motives-migration-south-african-doctors-australia-1948 [8 February 2013].

Banker, R.D. 1984. Estimating most productive scale size using Data Envelopment Analysis. *European Journal of Operational Research* 17: 35-44.

Banker R.D., Bardhan I., & Cooper W.W. 1996a. A note on returns to scale in DEA. *European Journal of Operational Research* 88: 583-585.

Banker, R.D., Chang, H. & Cooper, W.W. 1996b, Equivalence and Implementation of Alternative Methods for Determining Returns to Scale in Data Envelopment Analysis. *European Journal of Operational Research* 89: 473-481.

Banker, R.D., Charnes, A. & Cooper, W.W. 1984. Models for the estimation of technical and scale inefficiencies in Data Envelopment Analysis. *Management Science* 30: 1078-1092.

Banker, R.D., Cooper, W.W., Seiford, R.M. Thrall, R.M., & Zhu, J. 2004. Returns to scale in different DEA models. *European Journal of Operational Research* 154: 345-362.

Banker, R.D. & Thrall, R.M. 1992. Estimation of returns to scale using Data Envelopment Analysis. *European Journal of Operational Research* 62: 74-84.

Barro, R. 2008. *Macroeconomics, A Modern Approach*. United States of America: Thomson South-Western.

Bogetoft, P. & Otto, L. 2011. *Benchmark and frontier analysis using DEA and SFA* (version 0.20). [online]. Available: http://cran.r-project.org/web/packages/Benchmarking/ [8 February 2013].

Carrin, G., Mathauer, I., Ke Xu, K. & Evans, D.B. 2008. Universal coverage of health services: tailoring its implementation. *Bulletin of the World Health Organization* 86(11): 857-863.

Centre for Development and Enterprise. 2011. Reforming healthcare in South Africa: What role for the private sector?. *CDE Research no 18*. [Online]. Available: http://www.cde.org.za/article.php?a_id=413 [8 February 2013].

Charnes, A. & Cooper, W.W. 1989. Data Envelopment Analysis. *Center for Cybernetic Studies Research Report* 626.

Charnes, A., Cooper, W.W. & Rhodes, E. 1978. Measuring the efficiency of decision making units. *European Journal of Operational Research* 2: 429-444.

Clement, J.P., Vivian, G., Vladmanis, V., Bazzoli, G.J., Zhao, M. & Chukmaitov, A. 2008. Is more better? An analysis of hospital outcomes and efficiency with a DEA model of output congestion. *Health Care Management Science* 11: 67-77.

Cleverley, W.O., Stanko, B.B. & Zeller, T.L. 1997. A new perspective on hospital financial ratio analysis. *Healthcare Financial Management* 51(11).

Coelli, T.J. 1996. A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program. *Centre for Efficiency and Productivity Analysis Working Paper Series* (WP08/1996).

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

Cook, W.D. & Zhu, J. 2007. Classifying inputs and outputs in data envelopment analysis. *European Journal of Operational Research* 180: 692-699.

Competition Tribunal of South Africa. 2006. *Phodiclinics (Pty) Ltd & four others and Protector Group Medical Services (Pty) Ltd & five others. Case No. 122/LM/Dec05*.

Competition Tribunal of South Africa. 2008. *Netcare Hospital Group (Pty) Ltd and Community Hospital Group (Pty) Ltd. Case No. 27/CR/Mar07*.

Competition Tribunal of South Africa. 2010. *Life Healthcare Group and Amabubesi Hospitals (Pty) Ltd & Bayview Private Hospital (Pty) Ltd. Case No. 11/LM/Mar10*.

Coovadia, H., Jewkes, R., Barron, P. Sanders, D. & McIntyre, D. 2009. The health and health challenges of South Africa: Historical roots of current public health challenges. *The Lancet* 374: 817-834.

Council for Medical Schemes (CMS). 2011. *Annual Report 2010 – 2011*. Pretoria. [online]. Available: http://www.medicalschemes.com/Publications.aspx?id=7&category=Annual%20Reports [8 February 2013].

Department of Health. 2011a. National Health Act, 2003. Policy on National Health Insurance. *Government Gazette* 34523. Government Notice 554, Republic of South Africa, Pretoria.

Department of Health. 2011b. Media Statement – Release of Green Paper on National Health Insurance. Released on 11 August 2011. [Online]. Available: http://www.doh.gov.za/show.php?id=2895 [8 February 2013].

Department of Health. 2011c. Vision and Mission. [online]. Available: http://www.doh.gov.za/show.php?id=2870 [8 February 2013].

Department of Health. 2011d. National Health Act, 2003. Policy on the Management of Hospitals. *Government Gazette* 34522, Government Notice 656, Republic of South Africa, Pretoria.

Development Bank of Southern Africa. 2008. Health Roadmap. [online]. Available: http://www.dbsa.org/Research/Documents/Health%20Roadmap.pdf [8 February 2013].

Eckermann, S. & Coelli, T. 2008. Including quality attributes in a model of health care efficiency: A net benefit approach. *Centre for Efficiency and Productivity Analysis.* Working Paper Series (WP03/2008).

Econex. 2010a. Integration of the public and private sectors under a National Health Insurance (NHI) system in SA. *Health reform note 4*. [online]. Available: http://www.econex.co.za/index.php?option=com_docman&task=doc_download&gid=58&Itemid=60 [8 February 2013].

Econex. 2010b. Updated GP and Specialist numbers for SA. *Health reform note 7*. [online]. Available: http://www.mediclinic.co.za/about/Documents/ECONEX_Health%20reform%20note_7.pdf [8 February 2013].

Emrouznejada, A., Parkerb, B.R. & Tavaresc, G. 2008. Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Socio-Economic Planning Sciences* 42: 151-157.

Färe, R.S., Grosskopf, S. & Lovell C.A.K. 1985. *The Measurement of Efficiency of Production*. Boston: Kluwer Nijhoff. Quoted by Tone, K. 1996. A simple characterization of returns to scale in DEA. *Journal of the Operations Research* 39(4): 604-613.

Farrell, M.J. 1957. The measurement of productive efficiency. *Journal of the Royal Statistical Society*, Series A (General) 120(3): 253-290.

Gordhan, P. (Minister of Finance). 2011. 2011 Budget Speech, presented to the Parliament of South Africa on 23 February 2011. [online]. Available: http://www.sars.gov.za/home.asp?pid=66066 [8 February 2013].

Gordhan, P. (Minister of Finance). 2012. 2012 Budget Speech, presented to the Parliament of South Africa on 22 February 2012. [online]. Available: http://www.sars.gov.za/home.asp?pid=75305 [8 February 2013].

Grosskopf, S. & Valdmanis, V. 1987. Measuring Hospital Performance, A Non-Parametric Approach. *Journal of Health Economics* 6 89-107.

Hollingsworth, B. 2008. The Measurement of Efficiency and Productivity of Health Care Delivery. *Health Economics* 17: 1107-1128.

Hollingsworth, B., Dawson, P. & Maniadakis, N. 1999. Efficiency measurement of health care: a review of non-parametric methods and applications. *Health Care Management Science* 2: 161-172.

Jacobs, R. 2001. Alternative Methods to Examine Hospital Efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis. *Health Care Management Science* 4: 103-115.

Keegan, A. 2008. Hospital Bed Occupancy. Australian Doctors Fund. [online]. Available: http://www.adf.com.au/archive.php?doc_id=168 [8 February 2013].

Kibambe, J.N. & Koch, S.F. 2007. DEA applied to a Gauteng sample of public hospitals. *South African Journal of Economics* 75: 351-368.

Lovell, C.K. 1996. Applying efficiency measurement techniques to the measurement of productivity change. *Journal of Productivity Analysis* 7: 329-340.

Linna, M. 1998. Measuring hospital cost efficiency with panel data models. *Journal of Health Economics* 7: 415-427.

McCallion, G., Glass, J.C., Jackson, R., Kerr, C.A., & McKillops, D.G. 2000. Investigating productivity change and hospital size: a nonparametric frontier approach. *Applied Economics* 32: 161-174.

Matsebula, T. & Willie, M. 2007. Private hospitals. *South African Health Review 2007*: 159-174.

McIntyre, D. 2010. Private sector involvement in funding and providing health services in South Africa: Implications for equity and access to health care. *EQUINET Discussion Paper Series* 84, Health Economics Unit (UCT), ISER Rhodes University, EQUINET: Harare.

McIntyre, D., Thiede, M., Nkosi, M., Mutyambizi, V., Castillo-Riquelme, M., Gilson, L., Erasmus, E. & Goudge, J. 2007. A critical analysis of the South African health system. *Shield Work Package 1 Report*.

Mehrabian, S., Jahanshahloo, G.R., Alirezaee, M.R. & Amin, G.R. 2000. An Assurance Interval for the Non-Archimedean Epsilon in DEA Models. *Operations Research* 48(2): 344-347.

Nguyen, K. & Coelli, T.J. 2009. Quantifying the effects of modelling choices on hospital efficiency measures: A meta-regression analysis. *Centre for Efficiency and Productivity Analysis*. Working Paper Series (WP07/2009).

Norman, C., & Weber, A. 2009. *Social Health Insurance, a guidebook for planning.* 2nd Ed. [online]. Available: http://www2.gtz.de/wbf/4tDx9kw63gma/Guidebook_SHI_WHO-GTZ-ILO-ADB.pdf [8 February 2013].

O'Neill, L., Raunerb, M., Heidenbergerb, K. & Krausc, M. 2008. A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences* 42: 158-189.

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.

R Foundation for Statistical Computing. 2012. *R* (version 2.15.2). [online]. Available: http://cran.r-project.org/ [8 February 2013].

Ramanathan, R. 2003. *An Introduction to Data Envelopment Analysis: A Tool for Performance Measurement*. New Delhi: Sage Publications.

Ramjee, S. & McLeod, H. 2007. Medical Schemes. *South African Health Review 2007*: 47-70.

Ramjee, S. & McLeod, H. 2010. Private sector perspectives on national health insurance. *South African Health Review 2010*: 179-194.

Republic of South Africa. 1996. *The Constitution of the Republic of South Africa* (Act no. 108 of 1996). [online]. Available: http://www.info.gov.za/documents/constitution/93cons.htm [8 February 2013].

Republic of South Africa. 1998. *Medical Schemes Act* (Act no. 131 of 1998). [online]. Available: http://www.info.gov.za/view/DownloadFileAction?id=70667 [8 February 2013].

Scheller-Kreinsen, D., Geissler, A. & Busse, R. 2009. The ABC of DRGs. *Euro Observer, The Health Policy Bulletin of the European Observatory on Health Systems and Policies* 11(4): 1-12.

Schieber, G., Baeza, C., Kress, D. & Maier, M. 2006. Financing health systems in the 21st Century. *Disease control priorities in developing countries*. 2nd ed. New York: Oxford University Press.

Seiford, L.M. 1997. A bibliography for Data Envelopment Analysis (1978–1996). *Annals of Operations Research* 73: 393-438.

Seiford, L.M. & Thrall, R.M. 1990. Recent developments in DEA: The mathematical programming approach to frontier analysis. *Journal of Econometrics* 46: 7-38.

Sherman, H.D. 1984. Hospital Efficiency Measurement and Evaluation: Empirical Test of a New Technique Author. *Medical Care* 22(10): 922-938.

Sherman, H.D. & Zhu, J. 2006. *Service Productivity Management: Improving Service Performance Using Data Envelopment Analysis (DEA)*. Boston: Springer.

Smith, P. 1997. Model misspecification in Data Envelopment Analysis. *Annals of Operations Research* 73: 233-252.

Smith, P.C. & Street, A. 2005. Measuring the efficiency of public services: the limits of analysis. *Journal of Royal Statistical Society* 168: 401-417.

Tone, K. 1996. A simple characterization of returns to scale in DEA. *Journal of the Operations Research* 39(4): 604-613.

Valdmanis, V.G., Rosko, M.D. & Mutter, R.L. 2008. Hospital Quality, Efficiency, and Input Slack Differentials. *Health Services Research* 43: 1830-1848.

World Health Organisation (WHO). 2012. Global Health Observatory Data Repository. [online]. Available: http://apps.who.int/gho/data/ [8 February 2013].

Worthington, A. C. 2004. Frontier Efficiency Measurement in Health Care: A Review of Empirical Techniques and Selected Applications. *Medical Care Research and Review* 61: 135-170.

Zere, E., McIntyre, D. & Addison, T. 2001. Technical efficiency and productivity of public sector hospitals in three South African provinces. *South African Journal of Economics* 69(2): 336-358.

Zuckerman, S., Hadley, J. & Iezzoni, L. 1994. Measuring hospital efficiency with frontier cost functions. *Journal of Health* Economics 13: 255-280.