UNIVERSITY OF CAPE TOWN

Using DEA to profile in-hospital surgeon services: A South African funder perspective

Matan Abraham

ABRMAT003

Dissertation submitted in full fulfilment of the requirements for the degree of Master of Philosophy in Actuarial Science

Faculty of Commerce

November 2014

Abstract

The comparative assessment of physician performance, also known as 'physician profiling' is frequently used by healthcare funders. It aims to identify and improve the resource efficiency and quality of physician care. South African private healthcare funders use a wide range of profiling techniques; however, currently the use of frontier analysis is absent. This study explores the use of the non-parametric frontier analysis technique called Data Envelopment Analysis (DEA) for the profiling of physicians in South Africa. This is investigated by following a DEA profiling approach to evaluate the performance of 403 general/ paediatric surgeons in providing in-hospital services in 2012. A 7-input 1-output VRS DEA model is used to determine the efficiency of the surgeons. The profiling results are then analysed to determine their usefulness. It results reveal that 58 surgeons are efficient, representing only 14.4% of surgeons profiled. Therefore, the DEA approach reveals a large potential for efficiency improvements. The average efficiency score of inefficient surgeons is found to be 0.68. This means that, on average, inefficient surgeons have to decrease resource utilisation by 32% to achieve efficiency. The DEA approach is also found to be proficient at identifying the physicians presenting the most severe levels of inefficiency. 37 surgeons are found to be significantly inefficient. The approach also allows for the identification of peers against which inefficient surgeons are able to directly compare their practices. These results are determined to be of significant potential use to South African private healthcare funders. It is, however, noted that the analysis and results obtained was solely of a statistical nature. Closer consideration of the clinical appropriateness of the results is essential. In any case, this study concludes that a DEA profiling approach can be considered a useful technique in the comparison of physician performance in South Africa.

Declaration

I hereby declare that:

1. this is my own unaided work, and that each significant contribution to, and quotation in, this dissertation from the work of other people has been cited and referenced.

2. neither the substance nor any part of the thesis has been submitted in the past, or is being, or is to be submitted for a degree at this University or any other University.

I grant the University of Cape Town free license to reproduce for the purpose of research either the whole or any portion of the contents

M Abraham

November 2014

Acknowledgements

My friend and supervisor, Shivani Ramjee, has been the most important guiding force through the completion of my masters. I would like to thank her for the unwavering motivation, expertise and support. The knowledge and skills she has imparted over the past two years have proved invaluable to my academic and personal growth.

A special thanks to my co-supervisor, Kathryn Dreyer. Her generosity of time and knowledge has been essential throughout. I would also like to thank my parents, Haim and Yafa Abraham. Everything that I have achieved up to this point I owe to them.

A massive thank you to Medscheme for providing the data used in this masters. The investigations carried out in this Masters would not have been possible without them.

I would like to acknowledge the University of Cape Town for the financial assistance and academic resources provided toward the completion of my masters.

Finally, I would like to thank the examiners for their useful comments and insights during the review process of this Masters.

Table of Contents

List of Figures

List of Tables

1 Introduction

The cost of healthcare services worldwide has experienced rapid growth over the past decade and this trend is expected to continue (Deloitte, 2013). This issue is of major concern in South Africa where healthcare in the private sector is only affordable to a minority of the population (McLeod & Grobler, 2010; Mills et al., 2012; van den Heever, 2012). Escalating healthcare service prices result in increased expenditure faced by medical schemes[1] in order to provide indemnity cover for their beneficiaries' healthcare needs. The increasing healthcare service costs in South Africa represent one of the major forces resulting in annual medical scheme contribution rate increases exceeding CPI-inflation by at least 4% for each of the 12 years preceding 2013 (du Preez, 2013; McIntyre, 2010; Ramjee, Kooverjee, & Dreyer, 2013).

In an attempt to curb the rising cost of healthcare, medical schemes have prioritised what has come to be known as 'managed care'. Managed healthcare in is defined in South African regulation as "an arrangement through which utilisation of healthcare is monitored though the use of mechanisms which are designed to monitor appropriateness, promote efficacy, quality and cost effectiveness of delivery of relevant health services" (Medical Schemes Act No. 131 of 1998). Managed care, therefore, comprises the entire range of cost-saving and quality-enhancement techniques in healthcare delivery (Edmunds, 1997).

The focus of this study is the managed-care technique called 'physician profiling'. Lasker, Shapiro, and Tucker (1992) describe physician profiling as the use of

[1] Medical schemes are non-for-profit entities that play the dominant private healthcare-funding role in South Africa. Regulation limits the financing of comprehensive health insurance to medical schemes. They are distinct from South African health insurance companies as they are regulated separately based on social-solidarity principles (McLeod, 2005). A detailed description of medical schemes and the role they play in the South African healthcare environment is provided in Section 2.1.1.

epidemiological methods to compare practice patterns of physicians[2] on the basis of cost, service use, or quality of care. Lasker et al. (1992, p. 288) elaborate that physician profiling is used to "identify overutilisation of services, to uncover problems with the efficiency and quality of care, and thereby assess physician performance". Expenditure on physician healthcare services in South Africa comprised more than 30% of total private healthcare expenditure in 2012 (Council for Medical Schemes, 2013). This percentage rises substantially if one includes all related healthcare services that physicians utilise in the treatment of their patients (Hodge, Fiandeiro, Lynch, & Mohamed, 2012). The significant proportion of healthcare services for which physicians are responsible makes monitoring physician efficiency essential in order to curb rising healthcare costs. Profiling is a method of achieving this through the evaluation of physician performance as well as providing physicians with education and incentives aimed at increasing their level of efficiency (Garnick et al., 1994). Physician profiling therefore takes into consideration the full range of managed-care mechanisms listed above.

Over the past 25 years, frontier analysis techniques have increasingly been employed by international healthcare funders to evaluate the performance of physicians (Hollingsworth, 2008). While medical schemes do extensively employ physician profiling, the use of frontier analysis techniques is absent. The purpose of this study is to investigate the use of the non-parametric frontier analysis technique known as 'Data Envelopment Analysis' (DEA) for profiling applications in South Africa. This is explored using a DEA approach to profile the efficiency of 403 general/ paediatric surgeons when providing in-hospital services in the private sector. The results obtained are then analysed to determine what profiling using DEA may reveal about the efficiency of these surgeons as well as what efficiency improvement interventions these results may be able to inform. It

[2] Throughout, the term 'physician' refers to a professional who practices medicine, and who is involved in promoting, maintaining or restoring human health through the study, diagnosis, and treatment of disease, injury, and other physical and mental impairments (WHO, 2010).

is important to note that this study is limited to the assessment of DEA as a profiling methodology. Profiling outcomes obtained from the DEA approach are not directly compared with other profiling techniques; frontier analysis or otherwise. The next section provides a description of the South African healthcare industry in order to provide insight into the environment in which physician profiling is performed. Section 3 provides a detailed discussion of physician profiling in order to ensure understanding of the problem being addressed and the role profiling plays in its solution. Section 4 & 5 discuss the nature of efficiency and the use of DEA in order to evaluate performance. Section 6 outlines in detail the physician profiling methodology undertaken and Section 7 describes the principal results obtained from the profiling process. Finally, Section 8 concludes this study with a discussion on what has been revealed by the investigation of physician efficiency using DEA.

2 South African healthcare environment

The South African healthcare environment is composed of both private and public healthcare sectors, catering to different segments of the population (McLeod, 2005). The private and public sectors are differentiated according to the manner in which healthcare services are both funded and provided (Hodge et al., 2012). Accordingly, South Africa operates a dual healthcare system with regard to both the funding and provision of healthcare. 'Private healthcare' refers to healthcare services that are both privately funded by the patient and accessed through private for-profit service providers (Hodge et al., 2012). Conversely, 'public healthcare' is funded by the government through tax contributions and accessed through public healthcare facilities staffed and run by government employees (McLeod, 2005).

While the above describes the current state of the South African healthcare environment, significant healthcare policy reforms are planned. These involve a transition to a National Health Insurance (NHI) environment. The current funding and provision systems are, therefore, expected to change markedly. The details of NHI are, however, still in their early stages (Matsoso & Fryatt, 2013). Physician profiling in this study is performed with consideration of the current state of the South African healthcare environment. Future consideration of the profiling process undertaken and the results obtained must reflect any changes that have occurred.

2.1 Private healthcare funding

Private healthcare funding in South Africa comprises purchasing health insurance and/or self-funding (Hodge et al., 2012). Each is discussed in turn below.

2.1.1 Medical schemes

Healthcare financing regulations in South Africa require that indemnity health insurance take the form of 'medical schemes', which are distinguished from conventional health insurers in a number of ways (Medical Schemes Act No. 131 of 1998). Medical schemes are not-for-profit entities owned by their members (McLeod & Grobler, 2010; McLeod & Ramjee, 2007). Furthermore, they are regulated according to three core social-solidarity principles, namely: 'open enrolment', 'community-rating', and all medical scheme products must provide cover for a package of 'prescribed minimum benefits' (PMBs) (McLeod, 2005).

Open enrolment requires that all schemes accept anyone who wants to become a member, provided they are able to afford the monthly contributions (McLeod, 2005). The only exception is 'restricted membership' schemes that limit membership of individuals based on their particular employer or industry (McLeod & Ramjee, 2007). Community rating requires that all medical schemes charge their members the same standard contribution rate, regardless of age, gender and/or health status (past and present) (van den Heever, 2012). Therefore, medical schemes are prohibited from underwriting and setting contribution rates according to the risk of the individual member (McLeod, 2005). Schemes may, however, impose waiting periods and age-related late-joiner penalties, as defined in regulation, in order to mitigate anti-selection and moral hazard. Every product offered by a medical scheme must provide cover for a defined minimum package of benefits (McLeod, Mubangizi, Rothberg, & Fish, 2003). These PMBs comprise 272 diagnosis-treatment pairs, 26 chronic diseases and all emergency care; the reimbursement of which must be covered in full by the scheme, without financial limits or co-payments (McLeod & Ramjee, 2007; Taylor, Taylor, Burns, Rust, & Grobler, 2007). There were 90 medical schemes providing near-indemnity cover to 8.8 million individuals in 2012 (Council for Medical Schemes, 2013). Medical schemes are the primary private healthcare funders in South Africa covering 16.6% of the population (Council for Medical Schemes, 2014).

Medical schemes are required to perform certain administrative functions. "Some schemes elect to perform the administration tasks themselves ('self-administered' schemes), whilst others contract with a third-party administrator undertaking the administration role on behalf of the scheme" (Hodge et al., 2012, p. 18). Third-party administrators may provide their services to multiple schemes. Furthermore, unlike medical schemes, they are for-profit entities remunerated by schemes on an agreed-upon basis (Hodge et al., 2012; van den Heever, 2012). Examples of services performed by the administrator on behalf of the scheme include: collecting and reconciling contributions, validation and payment of all claims, financial reporting, information management and data control as well as customer services and acquisitions (Council for Medical Schemes, 2010).

As stated in Section 1, the term 'managed care' in South Africa encompasses all techniques designed to promote the efficient provision of high-quality care, including physician profiling (Medical Schemes Act No. 131 of 1998). Examples of managed care techniques utilised in medical schemes (apart from physician profiling) include: selective contracting, pre-authorisation, and capitation as well as case-, disease-, and chronic medication management (National Library of Medicine, 2014). Managed-care techniques are of particular importance in South Africa since medical schemes are subject to community rating and open enrolment. Medical schemes, therefore, have very limited ability to control contribution rates using underwriting and risk rating. As a result, they are heavily dependent on managed-care techniques to control healthcare costs (McLeod, 2005).

Managed care services in South Africa are primarily provided by managed care operators (MCOs) as well as the medical scheme administrators themselves (van den Heever, 2012). MCOs are also for-profit entities but the services that they provide to medical schemes are limited to managed-care functions. In addition, the medical scheme may itself perform some of the managed care functions and outsource the more specialised functions to a MCO or administrator (Hodge et al.,

2012). There were 27 administrators and 41 MCOs accredited to provide services to medical schemes in 2012 (Council for Medical Schemes, 2013; Econex, 2013).

2.1.2 Other forms of healthcare funding

There has been growth in other healthcare financing products provided by conventional for-profit insurers in South Africa. Such products include hospital cash plans, critical illness cover, disability cover and gap cover (Econex, 2013). However, healthcare funding by conventional health insurers is still limited to providing cover to a very small percentage of the population (Econex, 2013).

There is also a significant amount of self-funding by individuals in South Africa (McLeod & Grobler, 2010). Self-funding involves the financing of private healthcare services through out-of-pocket payments made directly to the private provider (van den Heever, 2012). Self-funding is required when individuals do not have any (or sufficient) healthcare insurance but still access private healthcare services (Hodge et al., 2012). The combination of insurance- and self-funding results in approximately 35% of the population being served to some degree by the private sector (Centre for Development and Enterprise, 2011; Hodge et al., 2012).

The remainder of the South African population rely solely on public sector for the provision and funding of their healthcare needs. Healthcare in the public sector is provided for free to those individuals earning below the means test of R6 000 per month (Reference). Individuals earning salaries higher than this are required to make payments out-of-pocket for their public healthcare needs. These payments are based on the Uniform Patient Fee Schedule (UPFS) (Western Cape Government Hospital Tariffs, 2014).

It is important to reiterate at this point, however, that this study investigates the use of DEA as a profiling methodology in the South African private sector. The funding mechanism of focus, therefore, is through medical schemes.

2.2 Private healthcare provision

Private healthcare provision in South Africa refers to medical professionals treating patients utilising a wide range of services provided in privately-owned healthcare facilities (Hodge et al., 2012). The main types of private healthcare service provider are: primary-care providers, specialists, and private hospitals (Hodge et al., 2012; McLeod, 2005; van den Heever, 2012).

Primary-care providers include general practitioners (GPs), dentists and pharmacies (van den Heever, 2012). These providers are often the patient's first contact with the health system and they assume responsibility for the provision of continual and comprehensive medical care to individuals, families, and communities (Hodge et al., 2012; WHO, 2010). Services provided by primary-care providers include diagnosis and performing medical procedures as well as referrals for hospital-based treatment and for further examination by specialists (van den Heever, 2012).

Specialists provide many of the same services as primary care providers, however, they focus their practices on certain disease categories, types of patients, or methods of treatment (Hodge et al., 2012; WHO, 2010). In order to become a specialist, doctors further their medical education in a specific field of medicine. Examples of specialists are oncologists, surgeons and psychiatrists.

Private hospitals are medical facilities that provide medical, surgical and psychiatric care as well as treatment of the sick or injured (Hodge et al., 2012). Hospitals provide patients with a bundle of discipline-related specialist services (surgery, oncology, paediatrics etc.), and general-support services (nursing, recovery, radiology, pathology etc.) (Competition Commission, 2013; van den Heever, 2012). However, the mix and/or quality of services provided may vary substantially between hospitals (Hodge et al., 2012).

According to South African regulation, private hospitals may not employ physicians; neither primary-care providers nor specialists (Hodge et al., 2012; van den Heever, 2012). Private hospitals are thus limited to providing physicians with (1) the facility in which their patients are treated and subsequently recover (consultation rooms, operating theatres, various types of wards etc.), and (2) the supporting resources needed to effectively treat patients (Hodge et al., 2012). These supporting resources include: medical equipment, pharmaceuticals, surgical items, nursing services and 'hotel services' such as beds, catering and associated administration services (Hodge et al., 2012). An important consequence of hospitals not employing physicians is that hospital managers are predominantly concerned by the quality of physician care and are largely indifferent to physician resource efficiency (Chilingerian & Sherman, 1990). As a result, the profiling of resource efficiency in South Africa falls squarely on funders and/or third-party managed care providers.

These healthcare service providers are thus separate entities and are reimbursed by the medical scheme and/or the patient; depending on the nature of private healthcare funding (McLeod & Grobler, 2010). Nevertheless, they are required to work effectively together in order to provide integrated healthcare solutions to patients (Hodge et al., 2012).

3 Physician profiling[3]

3.1 Definition, objectives, applications and outcomes

Physician profiling is a managed care technique, which uses epidemiological methods in order to compare the practice patterns of physicians according to cost, resource utilisation and quality of care (Garnick et al., 1994; Lasker et al., 1992; Shapiro et al., 1993). From a funder perspective, profiling is the comparative performance assessment of the physicians providing healthcare services to their beneficiaries; both GPs and specialists (Eijkenaar & van Vliet, 2013; Hollingsworth, 2008). The funder's objective is to evaluate physicians' relative ability to provide resource efficient treatment to patients while still maintaining the highest achievable levels of quality.

While the above definition represents the overarching objectives of physician profiling, particular methods may focus individually on resource utilisation, cost effectiveness or quality of care. Physician profiling has become common practice by funders in various healthcare industries around the world especially for use in incentive-based remuneration of physician practices (Garnick et al., 1994). Rosenthal, Landon, Normand, Frank, and Epstein (2006) state that more than half of the 242 health maintenance organisations (HMOs) that they surveyed in the United States were engaging in some form of physician profiling.

A key methodological feature needs to be noted regarding the choice of norm against which physicians are compared. Comparison is made either by relating a particular physician's practice pattern to a norm determined by other similar physicians ("practice-based norm") or by relating a physician's outcome to an

[3] This section is largely based on the work done by Lasker et al. (1992) and Shapiro, Lasker, Bindman, and Lee (1993) concerning issues of attempting to achieve the full potential of physician profiling.

accepted practice guideline ("standards-based norm") (Lasker et al., 1992, p. 288; Normand, Glickman, & Gatsonis, 1997). An advantage of practice-based norms is that physicians may be more inclined to cooperate with efficiency improvement interventions if profiles are based on the efficiency levels of their peers; as opposed to some empirically determined practice guideline. If, however, the determined practice-based norm is based upon comparison with inefficient physicians then it will not reflect the optimal level of efficiency that can be achieved (Garnick et al., 1994). The standards-based norm can avoid this issue if appropriate practice guidelines can be determined upon which to base the norm (Shapiro et al., 1993). This is, however, a very complicated task and may lead to physicians contesting the level and construction of the standard practice pattern upon which the standard-based norm is determined.

Lasker et al. (1992) state three main applications of physician profiling, namely: quality improvement, utilisation review and the assessment of physician performance. Physician profiling can be used in a number of ways to improve the quality of healthcare. Many medical conditions and procedures exhibit large variations in patient outcomes. Profiling can be used to analyse whether certain physicians are achieving higher rates of negative outcomes than others (Garnick et al., 1994). Furthermore, profiling is able to inform whether these negative outcomes are the result of external factors, such as differences in case-mix, or due to poor quality of care by certain physicians (Findlay, 1993; Garnick et al., 1994). Funders can then distribute the results of the profiling process to physicians in order to provide them with detail on the parties and other elements impacting the quality of services they provide. Physicians are thereby provided with information allowing them to identify how and by whom quality may be improved (Shapiro et al., 1993).

Profiling also allows for the analysis of the utilisation rate of healthcare resources used by physicians; known as 'utilisation review'. This is achieved by identifying outlier physicians among those profiled who have used significantly more

resources to treat their patients than the norm (Lasker et al., 1992). The focus solely on outlier physicians has the benefit over alternative methods of utilisation review since it exempts most of the physicians from detailed case-by-case review of their practices (Findlay, 1993). In addition, funders can provide the outlier physicians with standard- and/or practice-based norms, providing them with information on how to achieve greater resource utilisation efficiency.

Potentially the most versatile use of physician profiling is in the assessment of physician performance. The assessment of physician performance is often based on a combination of the above quality and resource utilisation analyses. Lasker et al. (1992) provide the following potential applications for physician performance assessment. First, physician assessments can form the basis for the accreditation of physicians as well as advising necessary levels of continued education. Second, physician assessments can routinely be performed to monitor continued compliance with standard practice guidelines or quality improvement measures. Third, performance assessments can be used to provide funders with evidence on the efficiency of physicians that will be prescribed to their members. In particular, the assessment of physicians based on cost-efficient performance has become common practice in the United States and often forms the basis of incentive-based remuneration of physicians (Charvet, 2009). As mentioned above, assessment of provider performance and remuneration is achieved by comparing provider's performance to some norm; either standards- or practice-based.

Parente (2002) summarises the various outcomes produced by a physician profiling process allowing funders to achieve the objectives stated above. Profiles help funders to monitor the performance of physicians in order to ensure that they are maintaining expected levels of resource utilisation and quality of care. Funders are able to present patients with feedback on the level of efficiency and quality of care with which they are being treated. This allows patients to make informed decisions when choosing a particular physician. Funders are also able to provide physicians with detailed results of the profiling process, which can be used to

provide them with focused education of how to improve their resource efficiency and quality of care. This feedback can be used to convey financial incentives to physicians for achieving the required levels of efficiency as well as providing justification for punitive action against physicians who do not achieve the required efficiency standards (Garnick et al., 1994). An example of such punitive action may be the exclusion of a physician from a funder's 'provider network'[4].

It is important to recognise that a funder undertaking the profiling of physician performance has no direct ability to impact on how the physician practices. The funder can only use incentives, education and relationships to encourage behaviour change. This may also involve working with the professional societies to which physicians' belong in order to help induce changes in practice behaviour.

3.2 Features of the healthcare service market that necessitate profiling

The price charged for the provision of healthcare services is not controlled by the competitive market forces affecting other commodities (Arrow, 1963; Mushkin, 1958; Zweifel, Breyer, & Kifmann, 2009) and as a result physicians need to be monitored in order to ensure they are operating efficiently. There are three main reasons for this, namely: 'information asymmetry', 'supplier-induced demand' and the 'third-party payer problem'. There exists substantial information asymmetry between patients and physicians (Arrow, 1963; Zweifel et al., 2009). This is because physicians have superior knowledge as to the consequences and possibilities of a particular treatment. Medical knowledge is complicated and

[4] A 'provider network' is a set of designated physicians and/or health care facilities that the funder has chosen will deliver specific healthcare services to their beneficiaries. The physician must agree to specified reimbursement and/or practice parameters to remain a member of the network (Twiss, Yamamoto, Pyenson, Allen, & Fredericks, 1998).

patients do not have the relevant education to question the method of treatment advised by their physician. This is further exacerbated by the fact that patients are often in a physically and mentally vulnerable state when needing treatment. Even patients with medical knowledge will struggle to make informed decisions in such a state (Zweifel et al., 2009). As a result, patients have to trust that their physicians are acting in their best interest and not in an attempt to maximise profits (Arrow, 1963). As patients have insufficient information with which to question the price they are being charged, information asymmetry gives physicians substantial power over patients regarding what they charge for their services. Furthermore, patients often are not provided with the price of treatment in advance because it depends on factors that are initially unknown, such as the diagnosis and the patient's recovery rate (Arrow, 1963). This largely constrains the patient from shopping for the cheapest price of their treatment needs.

Physicians play a dual role in relation to their patients. They are both the provider of healthcare services and the advisor of the services demanded by patients (Zweifel et al., 2009). The consequence of information asymmetry is that this advisory role changes into one where the physician potentially acts as the 'decider' of the level and nature of services to be demanded. This leads to 'supplier-induced demand'; a phenomenon where the purchase of healthcare services by patients is higher than what would have been the case if there was no information asymmetry. The physician is granted substantial power to control the quantity of services demanded by their patients. An incentive is subsequently created for physicians to stray away from efficient performance and provide unnecessary excess treatment for their own financial gain (Zweifel et al., 2009).

In addition, having health insurance removes the incentive on the part of patients to shop around for the best price of their healthcare needs (Mataconis, 2011). Patients have no financial incentive to take into consideration the price they are being charged when they are not paying the bill. This is known as the 'third-party payer problem' (Mataconis, 2011). The result is that physicians are able to charge

inflated prices without competitive market forces pushing them down (Mushkin, 1958).

For these reasons funders are required to profile physician practices to ensure that they are not abusing their price-setting power and not wasting healthcare resources.

3.3 Methods

Regardless of the profiling methodology that may be chosen, it is imperative that it exhibits two key features in order to yield physician profiles that can provide effective and useful performance improvements. First, the model must incorporate "adequate risk-adjustment to prevent systematic misclassification of providers due to differences in case mix" (Eijkenaar & van Vliet, 2013, p. 731). Case-mix adjustment attempts to remove the effect of differences in performance attributable solely to differences in the population treated by a specific physician (Christiansen & Morris, 1997; Normand et al., 1997). Examples of such differences include patient characteristics (such as age, gender and health status), as well as the type and severity of condition for which the patient is being treated. (Further details on case-mix adjustment are discussed in Section 5.2). The second feature is that the model must exhibit "adequate reliability to prevent random misclassification of physicians because of chance. When profiles have low reliability, they are driven by random chance instead of true performance, and interventions based on them may arbitrarily and unfairly penalize or reward physicians" (Eijkenaar & van Vliet, 2013, p. 731). In other words, reliability ensures that the variation between the efficiency profiles of physicians can more confidently be interpreted as resulting from true variation in efficiency and not from random chance. For profiles to be reliable, a large sample of physicians need to be profiled and a large volume of clinical data is required to ensure that each physician is being profiled based on their performance in treating a sufficiently large number of patients. This achieves statistical reliability as it ensures that the

variation between physicians' performance is sufficiently large relative to the variation of performance within physicians' practices (Eijkenaar & van Vliet, 2013).

It should also be noted that the profiling models used by funders are often built in-house and represent a method for funders to gain a competitive advantage over their peers. This is because a superior profiling method provides a funder with superior ability to assess the performance of physicians. As a result, the funder can better affect physician performance improvements, more accurately set performance remuneration levels, and more effectively select physicians for inclusion in their provider networks. Consequently, the details of the profiling methods used by funders are often intellectual property (IP) and are thus not available to the public. As a result, this Section does not describe the details of the methods used by specific funders but instead describes the main categories of profiling methods used.

One of the prominent methods of physician profiling in the United States are risk assessment models. Risk assessment in the context of physician profiling considers whether the expenditure of a particular physician in treating its patients is greater than that which was expected (Lodh, Raleigh, Uccello, & Winkelman, 2010). This involves the funder using one of a number of available models to estimate the expected cost of physician services to be reimbursed by the funder for treating the physicians particular population of patients and then comparing this with the actual cost of services provided (Pope & Kautter, 2007). Inefficient physicians are those whose actual costs exceed the expected amount; given quality of care provided, demographic characteristics and the health status of the physician's patient population (i.e. adjusting for case-mix and quality) (Thomas, Grazier, & Ward, 2004). The type of risk assessment model used to estimate the expected cost of treating the patients' defines how the results should be interpreted. For example, if the predicted costs are based on average patient expenditures then the efficiency measure will be relative to an average and not the

most efficient physicians (Pope & Kautter, 2007). The various methods in which risk assessment models estimate patients' expected treatment costs differ by the information that they use to adjust for case-mix and the factors that are seen to affect resource utilisation and quality of care (Lodh et al., 2010). Examples of proprietary risk assessment models available for purchase by funders from software vendors include: Diagnostic Cost Groups (DCGs), Episode Risk Groups (ERGs), Medstat Episode Groups (MEGs) and Adjusted Clinical Groups (ACGs) (Cumming, Knutson, Cameron, & Derrick, 2002; Winkelman & Mehmud, 2007).

Another profiling method that is widely implemented is the use of statistical techniques. There are a number of techniques that are used to detect quality of care, resource cost and usage outliers. Smith (1994) uses an analysis of variance (ANOVA) approach to analyse patient mortality rates of particular healthcare providers and thereby allocate the total variance according to severity of illness, quality of care and random fluctuation. Christiansen and Morris (1997) address the use of significance testing to determine if patient mortality rates of providers are significantly higher than those expected. A Poisson distribution is assigned to mortality rates. The null hypothesis of the test is that a provider's true mortality rate equals the average rate for all providers. Gillis and Hixson (1991) regress mortality rates using Monte Carlo simulation to determine the predicted change in the rate when specific factors are varied. Hierarchal regression modelling has also been extensively used to analyse variations in healthcare utilisation and outcomes (Normand et al., 1997). Parente (2002) discusses the use of multivariate regression to profile physicians by using logistic regression analysis of patient outcomes treated by a specific physician. Dependent variables are chosen which are believed to significantly affect resource utilisation and quality of care. The dependent variables are also chosen to account for case-mix and patient attributes.

The final method to consider is the frontier-analysis approach to profiling. This method is the focus of this study and as yet is not used in South Africa. Frontier analysis has, however, been shown to be an effective profiling technique in

academic literature (Andes, Metzger, Kralewski, & Gans, 2002; Wagner, Shimshak, & Novak, 2003). Frontier analysis defines physician efficiency relative to the set of 'best practice' physicians who exhibit the highest level of efficiency out of all the physicians being profiled (Lovell, 2006). Accordingly, the most efficient physicians make up what is known as a "best practice frontier", and the rest of the physicians lie at lower levels of efficiency within the frontier. Efficiency is determined relative to the most efficient physicians within the particular sample. Frontier analysis, therefore, profiles physicians against a practice-based norm and does not provide an absolute measure of efficiency. The two most prominent frontier techniques are Stochastic Frontier Analysis (SFA) and DEA. SFA is a parametric technique, which assumes two sources of fluctuation around a minimum total-cost function (Coelli, Rao, O'Donnell, & Battese, 2005). The first source of fluctuation is the inefficiency residual, which arises from the assumption that all firms are inefficient. The second source is a normally distributed error term, which accounts for noise (Ozcan, 2008). SFA can be used for hypothesis testing as well as measuring various types of efficiency. Profiling using SFA (or any of the other methods above) is not considered in this study. Profiling is performed exclusively using DEA. The details regarding the evaluation of efficiency using DEA are described in detail in Section 5.

3.4 Practical considerations

There are a number of issues that may potentially complicate the physician profiling methods discussed above. These can be divided into a number of different categories. Only those relating to the presentation of the profiling results to respective parties and their subsequent behaviour are discussed in this section. The issues relating to the choice and measurement of inputs and outputs as well as those relating to the profiling model construction and interpretation are discussed throughout Section 5. Those relating to the data used in the profiling methodology are discussed in Section 6.

Physicians and patients often do not understand or appreciate the results of the profiling process (Charvet, 2009) as well as the process itself. This may lead to the unfounded mistrust of physicians by patients. Consequently, deterioration in the relationship between the physicians and funder may ensue (Shapiro et al., 1993). Therefore, particular care needs to be taken to ensure that the profiling of physicians is a constructive process and not one that undermines the relationships between the parties involved in the financing and delivery of healthcare services (Lasker et al., 1992). One way this could possibly be achieved is by ensuring that the profiling methodology is as objective and accurate as possible, as well as exhibiting particular sensitivity regarding the manner in which the results are presented to the relevant parties.

The profiling methodology can result in physician behaviour that is adverse to the intended objectives of the process. An example of this is a profiling methodology that is too focused on resource utilisation and cost reduction. The result of such a methodology runs the risk of adverse behaviour on the part of physicians. Physicians may refuse to treat potentially high resource patients in order to increase their perceived efficiency (Charvet, 2009). Alternatively, physicians will accept to treat all patients but engage in widespread under-servicing of these patients to artificially augment efficiency levels.

In addition, profiling identifies those physicians that are considered to be resource, quality or cost inefficient compared to their peers or some standard norm. Those physicians may then be notified of their inefficiency. The results of the profiling process may then be conveyed in order to provide them with an understanding of how they can increase their efficiency in the future. The problem with focusing only on improvement interventions of inefficient physicians is that the perceived efficient physicians have no motivation to further strive for increased efficiency (Lasker et al., 1992).

Finally, the efficiency improvements informed by the profiling process need to be aligned with what is operationally achievable by physicians practicing in their specific environment (Chilingerian & Sherman, 1997). For example, increasing efficiency by incentivising physicians to substitute admissions to hospital with procedures performed in their own offices. An example of an operationally unrealistic strategy would be to advise physicians to improve efficiency by substituting the use of anaesthetist services with pharmaceutical painkillers. Such a strategy may theoretically improve efficiency but is clinically unfeasible. The above example illustrates that physician profiles need to be interpreted while keeping in mind feasible and acceptable physician practice styles (Chilingerian & Sherman, 1997).

The description of the methodology in Section 8 attempts to address as many of these profiling issues, in order to achieve as effective a profiling methodology as possible. It is pertinent to note, however, that even if these issues persist some experts still argue "that physicians will prefer profiling over other review techniques because annual review profiles focus on the use of resources and quality outcomes rather than on clinical decision-making" (Kassirer, 1994, p. 634). In other words, physician profiling does not require telling physician how to do their job, it rather requires of them to achieve stated resource utilisation and quality outcomes.

4 The production transformation process and the definition of efficiency[5]

From the above discussion on physician profiling it can be seen that the focus is on improving physician's efficiency and providing an understanding of how to rectify inefficiencies where they occur. In section 5, DEA is discussed as a method of achieving the objectives of physician profiling. A starting requirement is to frame profiling as the comparison of physicians' relative ability to efficiently transform their inputs into outputs of production. Thereafter, it is required to use the production transformation process to define efficiency in its various forms.

4.1 The production transformation process

A physician can be thought of as a firm like any other with a production process transforming inputs into outputs. The term 'outputs' refers to the products or services that the firm produces, while 'inputs' refer to the resources used by the firm to produce the outputs (Figure 4.1). Coelli et al. (2005) provide a simple example of a production transformation process where a shirt factory uses materials, labour and capital (inputs) to make shirts (output). It is important to make explicit that for a particular profiling process a single production transformation process is conceptualised applicable to all the physicians. Therefore, it is implicitly assumed that all physicians use the same inputs, albeit in different quantities, to produce the same outputs; and, furthermore, that all physicians have equal access to these inputs (Cooper, Seiford, & Zhu, 2011b). This assumption is indispensible as the profiling of different processes would provide limited information of value.

[5] The definitions of the different types of efficiency described in this section derive from those in Coelli et al. (2005) who in turn derived the definitions from the work of Färe, Grosskopf, and Lovell (1985), Färe, Grosskopf, and Lovell (1994), Lovell (1993), Farrell (1957), Debreu (1951) and Koopmans (1951).

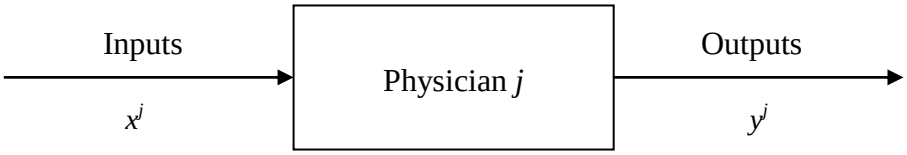

Figure 4.1 Basic illustration of physician production transformation process

In the profiling methodology used in this study, the physicians' inputs of production are the cost of healthcare services used to treat patients. The physicians' output of production is the case-mix adjusted number of patients that they treat within a given period. Therefore, the production transformation process is the physicians' practice of 'producing' treated patients through the utilisation of healthcare services (for example, surgery, anaesthetist and hospital services). The conceptualisation of physicians' production transformation process in this study, with the inputs and outputs of production, is an important part of the profiling methodology discussed in detail in Section 6.1.

Considering the characteristics of the production process is important as it ensures that profiling begins with understanding the responsibilities of the physician in the treatment of their patients. In particular, this involves understanding the decisions made by the physicians regarding the requisite healthcare services they utilised to produce a treated population of patients. Thanassoulis (2001) explains that by not considering a physicians production process the profiler runs the risk of comparing efficiency based on production variables outside the control of the physician. A proper understanding of the production process ensures, first, that the outputs used in the comparison of physician efficiency are actually affected by the inputs. Second, that profiling evaluates a holistic picture of the physicians' practices by including all the inputs responsible for affecting the outputs (Thanassoulis, 2001). Expressed in an example, conceptualising the production process of the shirt factory ensures that the quantity of materials used actually

affects the number of shirts made and that all the inputs affecting the number of shirts made are included in the process (Coelli et al., 2005).

Considering the activities of physicians as production processes is also an essential first step in using a DEA profiling approach. This is because it frames the way in which efficiency is defined and interpreted (Agrell & Bogetoft, 2001; Thanassoulis, 2001).

4.2 Defining efficiency

Utilising a DEA approach to profile performance comprises an efficiency analysis within physicians' production transformation process. 'Efficiency' goes a step further than conceptualising the production process; it requires consideration of the optimal way a firm (in this case a physician) uses its inputs to produce the intended outputs. Thus, 'optimality' dictates efficiency (Balk, 2001). Different measures of efficiency are described in the subsections that follow. It should be noted that throughout the descriptions of the various forms of efficiency, a given technology is assumed. The 'technology' is the exogenously determined environment in which the firm operates that defines the set of feasible combinations of input and output quantities (this will be discussed in detail in Section 5.3) (Agrell & Bogetoft, 2001; Balk, 2001).

4.2.1 Technical efficiency

Koopmans (1951, p. 60) provides the seminal definition of technical efficiency; "an input-output vector is technically efficient if, and only if, increasing any output or decreasing any input is possible only by decreasing some other output or increasing some other input". Technical efficiency thus reflects the firm's ability to obtain the maximum level of outputs from a given level of inputs or obtaining a given level of output from the minimum level of inputs (Coelli et al., 2005).

Koopman's (1951) definition was proposed for the purpose of analysing the technical efficiency of production for entire economies. Farrell (1957) extended the above Koopmans (1951) definition of technical efficiency by applying it at the individual firm level and "used the performances of other firms to evaluate the behaviour of each firm relative to the outputs and the inputs each of the other firms used" (Cooper, Seiford, & Zhu, 2011a, p. 5). This made it possible to evaluate firms' *relative* technical efficiencies.

The Farrell (1957) definition of relative technical efficiency is further refined by Charnes, Cooper, and Rhodes (1978) in the seminal work on DEA, and is the definition used in this investigation: "A firm is to be rated as fully (100%) efficient on the basis of available evidence if and only if the performances of other firms do not show that some of its inputs or outputs can be improved without worsening some of its other inputs or outputs". This efficiency definition better emphasises its 'relative' nature. It does not assume that theoretically possible levels of efficiency are known but rather emphasises that efficiency is defined using only the information that is empirically available (Cooper et al., 2011b). The above definition of relative technical efficiency is best illustrated in an example. Assume that the only input needed by a physician to treat a patient is hospital services. Then a physician is relatively technical efficient compared to other physicians profiled if it is able to treat a given number of patients using the lowest level of hospital services, or if it is able to treat the most patients using a given level of hospital services.

4.2.2 Allocative efficiency

Farrell (1957) also formulated a definition of allocative efficiency by extracting information from the price of the inputs and outputs. Allocative efficiency entails selecting the optimal mix of inputs that produces a given level of output at the minimum cost, or by selecting the optimal mix of outputs that will maximise revenue using a given level of inputs (Coelli et al., 2005). Thus, allocative

efficiency reflects the firm's ability to utilise available inputs and/or produce intended outputs in the optimal proportions. Farrell (1957) went on to define total economic (or overall productive) efficiency as the product of technical and allocative efficiency (Färe et al., 1994).

4.2.3 Scale efficiency

Färe et al. (1994) decomposed technical efficiency into three components, one of which being scale efficiency. A firm may be technically and allocatively efficient but may benefit from changing the scale of its operations (Coelli et al., 2005). A firm is said to be scale efficient when its size of operations is optimal, so that any modifications to its size will reduce the firm's efficiency. For example, if a physician is able to increase its efficiency by increasing the size of its practice and treating more patients, then it is not scale efficient. The relationship between technical and scale efficiency is considered in the discussion of the return-to-scale properties of DEA models in Section 5.3.

4.2.4 Price efficiency

The final type of efficiency important to discuss is price efficiency. A firm is price efficient when all production inputs are purchased at the lowest possible price (Sherman & Zhu, 2006). The measurement of price efficiency thus requires data in order to determine the price of inputs into production. A firm can increase its price efficiency if it is able to purchase its inputs at a lower price, without sacrificing the quality of those inputs (Sherman & Zhu, 2006). For example, if a physician is able to purchase hospital services (such operating theatre time) to treat patients at a lower price in a hospital of equivalent quality then it is able to increase its price efficiency. The measurement of price efficiency is, however, complicated as there are many factors that influence the price efficiency of a firm (Coelli et al., 2005). The price efficiency of a physician practice is determined by the competition of the healthcare environment in which the physician practices (Zweifel et al., 2009). For example, the price efficiency of a surgeon relying

heavily on hospital services to treat its patients is determined by the price competition in the hospital industry. Price efficiency is also a function of the physicians bargaining power when purchasing services used to treat patients (Zweifel et al., 2009).

5 Data Envelopment Analysis[6]

5.1 Introduction

As previously mentioned, this study investigates the use of DEA as a method to profile physicians. Throughout the description of DEA in this section, reference is made to Figure 5.1 in order to graphically depict the pertinent issues discussed. DEA is a data-orientated non-parametric mathematical optimisation technique conceived by Farrell (1957) and later developed and disseminated by Charnes et al. (1978). DEA is a form of frontier analysis that uses linear programming to measure the relative efficiency of firms (Ozcan, 2008). By comparing firm's production process (discussed in Section 4.1), DEA provides a relative measure of a firm's ability to efficiently transform their inputs into outputs (Bogetoft & Otto, 2010).

DEA assumes that the production transformation process and the chosen inputs and outputs of production are the same for all the firms being analysed (Cooper et al., 2011b). The evaluation of efficiency is relative to other firms' practices ('practice-based' norm), rather than a theoretical notion of efficiency ('standards-based' norm). DEA, therefore, requires that an inefficient firm achieve the performance attained by efficient firms in order to be deemed efficient. Some profiling experts believe that recommendations made to inefficient physicians in this form may receive less resistance because they are based on what better practicing physicians actually accomplished using the same technological path taken by the inefficient physician (Chilingerian, 1995). Thus DEA says to inefficient firms: "These evaluations are not based on some pure notion of

[6] Throughout this discussion on the theory of DEA, emphasis is made concerning assumptions underlying the DEA approach; since violation of these underlying assumptions will jeopardise the validity of any results obtained from the model. These assumptions are highlighted in *italics*.

efficiency, but on what your peers with better practices actually accomplished. How come your peers can do it better than you?" (Chilingerian & Sherman, 1990, p. 11).

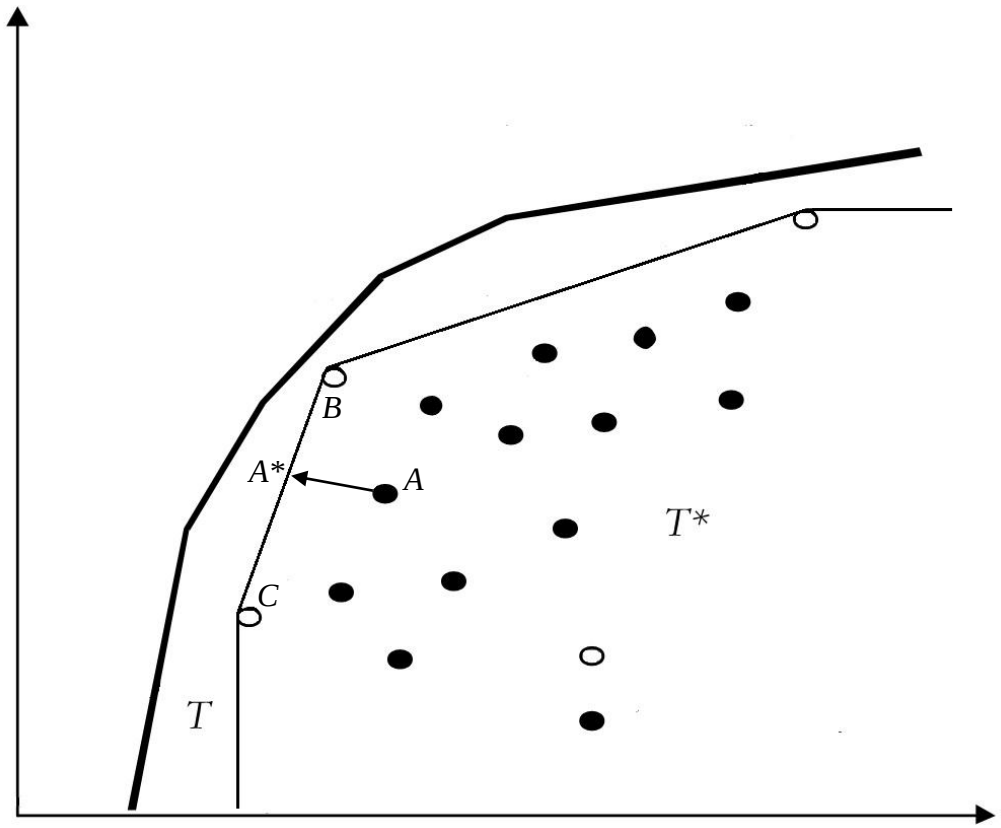

Figure 5.1 Graphical representation of a 1-input-1-output DEA best practice frontier of efficient DMUs along with the 'true' best practice frontier, the defined technology sets as well as the inefficient DMUs

Source: Agrell and Bogetoft (2001)

DEA is by no means a tool reserved for use in the application of physician profiling. It is a method that has been used to measure the relative efficiency of firms in numerous and varied production environments. Charnes et al. (1978) termed the sample of observed firms being analysed with a DEA model as

‘decision-making units’ (DMUs). Agrell and Bogetoft (2001) explain that one can characterise a DMU as any entity that transforms resources (inputs) into products and services (outputs). In other words, a DMU is any firm that has a production transformation process. Understandably, the choice of DMU is dependent on the practical area in which DEA is being applied. In addition, the choice of the DMU determines the types of inputs and outputs of production used in the measurement of efficiency (Thanassoulis, 2001). Therefore, in the application of DEA to physician profiling the DMUs are the physicians being profiled, the inputs are the healthcare services used by the physician to treat patients, and the output is the physician’s treated population of patients. Further examples of the DMUs applicable to the healthcare environment are hospitals, nursing homes and day clinics. In keeping with the above, ‘DMUs’ is used throughout the theoretical description of DEA to refer to the sample of firms whose efficiency is being analysed.

Only ‘classic’ DEA models are considered to form part of the profiling methodology followed in this investigation. Classic DEA models are limited to those maintaining the assumption that production activities can be characterised as a deterministic process of transforming quantifiable and homogenous inputs into quantifiable and homogenous outputs (Kuosmanen, 2001a). “Therefore, it is assumed that any stochastic variations in the process as well as any quality differences and non-measurable factors are assumed non-existing, negligible for the purposes of the analysis, or ‘correctable’ by means of some kind of data pre-processing” (Kuosmanen, 2001b, p. 9). DEA models allowing for stochastic variation of variables are known as ‘stochastic’ DEA models; however, these models are not considered in this study. Furthermore, classic DEA models are ‘proportional’, in the sense that all the inputs or outputs of DMUs need to be reduced or augmented in the same proportion in order for efficiency to be increased (Ozcan, 2008). There exist alternative DEA models that simultaneously aim to achieve both input reduction and output augmentation. In such models the

input reduction and output augmentation need not be proportional (Ozcan, 2008). These models are known as 'additive' or 'non-oriented' models, however, these models are also not considered.

Many profiling studies have been limited to the analysis of efficiency based on the use of a specific resource, such as pathology or radiology services (InterStudy, 1989). Chilingerian and Sherman (1997) stress, however, that if profiles are to approximate a physician's performance, they should include the multiple healthcare services that physicians utilise to treat their populations of patients. For ease of illustration, Figure 5.1 portrays a single-input single-output model, however, DEA explicitly allows for multiple inputs and multiple outputs to be incorporated into the model (Coelli et al., 2005). In addition, DEA requires very few *a priori* assumptions to be made regarding the nature of the chosen inputs and outputs, unlike standard forms of statistical regression analysis (Ozcan, 2008). As a result, "DEA has opened up possibilities for use in cases that have been resistant to other approaches because of the complex (often unknown) nature of the relations between the multiple inputs and multiple outputs involved in firms" (Cooper et al., 2011a, pp. 1, 2).

DEA analyses DMUs observed input-output combinations and uses these to develop a frontier of the most efficient DMUs against which all inefficient firms are compared (Ozcan, 2008). All DMUs in the sample that fall within the frontier are relatively inefficient compared to those that make up the efficient frontier (Figure 5.1). This is in contrast to other efficiency evaluation techniques that compare each firm relative to a central tendency by trying, for example, to fit a regression plane through the centre of the data; as is common in statistical regression (Chilingerian, 1995; Cooper et al., 2011a). This contrast is represented in Figure 5.1 by noting the difference between determining a firm's efficiency relative to the 'DEA frontier' as opposed to the 'regression line'. This is valuable in profiling applications as it encourages inefficient physicians to strive to emulate the practices of the observed 'best' physicians rather than just aiming to be above

average (Bogetoft & Otto, 2010). Chilingerian and Sherman (1997, p. 36) elaborate that "efficiency comparison should not be made based on prevailing practice standards or norms reflecting the average behaviour of physicians over time" and "averaging profiles will not pinpoint how an individual physician achieved a best practice". Furthermore, only large improvements in physician efficiency, likes those achieved by comparison with best-practice, can contain escalating health care costs (Chilingerian & Sherman, 1997).

Banker, Charnes, and Cooper (1984), Kuosmanen (2001b) and Bogetoft and Otto (2010) consider the measurement of relative efficiency using DEA comprising three distinct parts. The first is determining the purpose that DEA is to be applied within the chosen production environment. This involves the conceptualisation of the production process and the corresponding definition of efficiency upon which performance is based (discussed in Section 4). This part also includes the choice of the particular DMUs whose efficiency is to be evaluated and the variable selection process of the inputs and outputs. The second is defining the production possibility set (which Bogetoft and Otto (2010) term the 'technology set' or just the 'technology'), and the third is considering the method used to measure efficiency. The remainder of this section provides a detailed explanation of these three parts as well as the classic DEA models arising from their combination. In so doing, a detailed description is provided of the theory and rationale behind the DEA profiling methodology used in this study.

5.2 The choice of DMUs, inputs and outputs

As stated previously, the purpose to which DEA is being applied is the profiling of physicians in order for a funder to compare their relative efficiency. The DMUs are the individual physicians being profiled and thus the process is performed within the healthcare environment in which physicians practice. Furthermore, the

conceptualization of the physicians' production transformation process and the various definitions of efficiency are discussed in Section 4.

5.2.1 Choice of DMUs

With the above in mind, further consideration is needed regarding the choice of physician; comprising the DMUs under investigation. Profiling must be performed separately on different types of physicians. For example, the profiling of psychologists is performed and interpreted separately from that of gynaecologists (Chilingerian & Sherman, 1990). Comparing the efficiency of a psychologist to that of a gynaecologist provides no meaningful information to the funder or the respective physicians profiled. The physician types profiled are chosen based on the funder's interest in analysing their efficiency. The funder may profile all or only a subset of the physicians of a particular type. As discuss in Section 3, whether a specific subset of a funder's physician can be effectively profiled will depend largely on the volume and accuracy of data available as well as whether there is sufficient detailed information to perform adequate case-mix adjustment. The choice of physician type, therefore, depends on whether the above data requirements are achievable (discussed further in Section 6.4).

5.2.2 Choice of inputs and outputs

Regardless of the type of physician being profiled, there are no prescribed inputs and outputs that have to be used in a profiling process (Thanassoulis, 2001). DEA does provide some guidance by demanding that the choice of input-output variables included in a DEA profiling process depend on the conceptualisation of the production transformation process, the type of efficiency being assessed, the environment in which profiling is being applied, the type of physician being profiled as well as the factors that are and are not under the control of the physician (Thanassoulis, 2001). This, however, still results in the problem encountered in all profiling techniques, of choosing the most appropriate inputs and outputs. Thanassoulis (2001, p. 89) cautions that the "identification of the

input-output variables to be used in an assessment of comparative performance is the first and arguably the most important stage in carrying out the assessment. The results obtained depend crucially on the choice made".

As a result of the above, the starting point of identifying appropriate input and output variables requires professional and pragmatic judgment by individuals that are knowledgeable regarding the type of efficiency being assessed and the transformation of inputs to outputs within the given environment (Golany & Roll, 1989). There are, however, certain important principles governing the choice of the input-output variables. The six core requirements that inputs and outputs must exhibit when profiling using any method are that they be distinct, measurable, quantifiable, homogenous, accurate and that there be as few inputs and outputs as possible (Ozcan, 2008). When profiling physicians using DEA, inputs and outputs must exhibit the additional requirement of being exhaustive (Thanassoulis, 2001). These requirements are discussed in turn including how they affect the choice of physicians included in the profiling process.

Distinct inputs and outputs are ones that do not overlap. This applies to profiling methods that have multiple input and/or output variables (Thanassoulis, 2001). Distinct inputs and outputs are important as they allow the overall level of efficiency to be decomposed in order to identify the affect that a particular input or output has on the overall efficiency level. If the variables overlap it is uncertain as to what extent a specific input or output is contributing to the inefficiency when interpreting the results (Kittelsen, 1993). In addition, ensuring that each variable is distinct ensures that all additional inputs and outputs pooled are adding 'value' to the model in explaining the production transformation process (Wagner & Shimshak, 2007). In other words, this helps ensure that redundant variables are not included in the analysis; they are just linear combinations of other variables. Furthermore, this helps meet the requirement (discussed below) of including as few inputs and outputs as possible (Thanassoulis, 2001). Also, distinct inputs and outputs avoid overweighting the impact of a particular variable; this is done by

preventing it from being represented in multiple inputs or outputs in the model (Kittelsen, 1993). Ensuring that the inter-correlation is acceptably low between the inputs and between the outputs chosen is necessary to ensure that this requirement is met (Wagner & Shimshak, 2007; Wagner et al., 2003).

It seems obvious that inputs and outputs can only be included if their values are *measurable* and *quantifiable* (Coelli et al., 2005). Physician profiling ideally requires the measurement of variables that describe the true nature of service production (Ozcan, 2008). However, in practice, profiling often involves the use a related variable as proxy to an input or output that is not measurable and/or quantifiable (Ozcan, 2008). For example, if the utilisation of radiology services is unknown then the number of x-rays performed may be used as a proxy (Ozcan, 1998; Ozcan, Jiang, & Pai, 2000). It is important to be sure that the proxy appropriately reflects the nature of the underlying variable; otherwise physician performance will be determined by an undesired measure of efficiency.

A related requirement is that the definition and measurement of the inputs and outputs be *homogenous* (Chilingerian & Sherman, 1990). The healthcare environment poses three main sources of heterogeneity affecting the level of inputs and outputs used in the profiling of physicians. First, inputs and outputs can vary significantly depending on the volume and scope of services provided by a particular physician as well as by the severity of the patients they treat (Chilingerian & Sherman, 1990; Ozcan, 2008). This is the reason necessitating adequate case-mix adjustment in all profiling methods (Chilingerian & Sherman, 1990). The importance of case-mix adjustment is revisited later in this section.

Second, the inputs and outputs of physicians can vary considerably based on the quality of services provided. Possible methods of allowing for quality differences in the model are by including quality measures as inputs, outputs or exogenous factors of production (Eckermann & Coelli, 2008). Quality differences are, however, notoriously difficult to measure and adjust-for in the healthcare context

(Ozcan, 2008). Alternatively, the physicians profiled can be chosen provided they exhibit similar characteristics regarding their quality of care; thereby avoiding the need for quality adjustments (Eckermann & Coelli, 2008; Ozcan, 2008). Due to the challenges posed, quality of care is often assumed (implicitly or explicitly) to be constant across all physicians (Hollingsworth, 2008). This is, in fact, the assumption made in this study. The consequences of this on the interpretation of the profiling results obtained are discussed in detail in Section 8.

Third, input measurement may also exhibit heterogeneity due to differences in the "pricing of input units, supply and materials or labour costs across healthcare facilities depending upon region" (Ozcan, 2008, p. 13). The result of this is that technical efficiency cannot be differentiated from price efficiency.

A method to mitigate all the above sources of heterogeneity, and to ensure that the profiling process is comparing 'apples with apples', is to ensure that the physicians chosen to be profiled belong to the same 'peer-group' (Ozcan, 2008). In other words, physicians are chosen that have similar characteristics and are seen to be similar in the types of inputs and outputs used as well as in the way these inputs are used to produce outputs. This is a further reason why different types of physicians are profiled separately.

The measurement of inputs and outputs needs to be *accurate* (Coelli et al., 2005). Any errors or omitted data resulting from, for example, bad reporting practices can potentially have a significant distorting effect on profiling results. Bogetoft and Otto (2010) and Kuosmanen (2001b) explain that this is a particular problem in DEA models due to the fact that it is a non-parametric approach and does thus does not include a stochastic error term. The data checks used to ensure accuracy in this study are discussed in Section 6.4.2.

As stated above, the linear programming models used in carrying out classic DEA are non-parametric, and thus do not make initial assumptions regarding the functional form of the technology set or efficiency frontier (Bogetoft & Otto,

2010). It is, therefore, assumed that only the observed inputs and outputs determine the level of efficiency. Deviation from the frontier is a result solely of inefficient operations and in no-part from a chosen stochastic error term (Ozcan, 2008). *Exhaustiveness*, thus, requires that the included inputs in a DEA approach (and they alone) fully represent the level of the outputs. Thanassoulis (2001, p. 90) elaborates that "the input variables need to capture all the resources and the output variables all the outcomes". Conversely, only inputs and outputs that are necessary to explain the production transformation process should be included in the model (Kittelsen, 1993; Wagner et al., 2003). Furthermore, any environmental factors that affect the transformation of inputs into outputs should be included as part of the input or output set, depending on the direction of impact of the environmental factor (Thanassoulis, 2001). The exhaustiveness requirement can, however, be relaxed if it is assumed that any inputs or outputs omitted will not affect the results of the DEA analysis (Thanassoulis, 2001). Considering the correlations between the chosen inputs and outputs can assess the exhaustiveness of inputs and outputs (Thanassoulis, 2001; Wagner & Shimshak, 2007).

Finally, the challenge of the variable selection process conducted when profiling physicians using DEA is to find a parsimonious model, "using as many input and output variables as needed but *as few as possible"* (Wagner & Shimshak, 2007, p. 58); all of which exhibit the requirements discussed above. Jenkins and Anderson (2003) and Golany and Roll (1989) explain that the greater the number of variables the less power the model has at discerning true variations in physician efficiency from random fluctuations.

5.2.3 Case-mix adjustment

The necessity of adequate case-mix adjustment when profiling physician performance has been expressed in prior sections. However, because case-mix adjustment is such an important methodological requirement of effective

physician profiling, it is important to consolidate the reasons why it is needed and how it solves the problem of differences in case mix.

The purpose of case-mix adjustment is to allow for the effective comparison of efficiency across physician profiles. Case-mix adjustment achieves this by acting as a corrective tool used to homogenise the characteristics of the patient populations treated by the profiled physicians (Salem-Schatz, Moore, Rucker, & Pearson, 1994). The result of case-mix adjustment is the comparison of physicians with patient populations that are similar regarding their health and resource consumption requirements (Jencks & Dobson, 1987). Not adjusting for case-mix unfairly penalises physicians with higher risk patients; for example, those with older and sicker patients, as well as those treating more severe conditions. Case-mix adjustment allows for a comparable analysis where the focus lies on the practice pattern variation of physicians, instead of the differences in the patient population and their unique risk profile (Chilingerian & Sherman, 1990).

Case-mix adjustment is applied to the inputs and/or outputs used. The considerations involved are the same as those explained above in the discussion on achieving homogenous inputs and outputs. The case-mix adjustment process followed in this investigation is discussed in detail in Section 6.5.1.

5.2.4 Practice- vs. procedural-level analyses

This study has explicitly chosen to undertake physician profiling by analysing the performance of each physician on a ‘practice level’. In other words, the study considers the relative efficiency of physicians’ ability to treat all their patients during a particular period. Examples of this approach are the investigations performed by Chillingerian and Sherman (1995; 1997) as well as Wagner et al. (2003). Alternatively, profiling can be undertaken on a ‘procedural level’. In this case, the analysis would involve determining a physician’s efficiency in treating patients suffering from a particular health condition. Examples of this approach are Ozcan (2000) investigation comparing GPs and specialists efficiencies in

treating sinusitis patients as well as Ozcan (1998) investigation into the efficiency of GPs in treatment of otitis media.

The above distinction is important to note as each approach has its advantages as well as imparting its own set of complications and limitations on the study. Analysis on a practice level has the major advantage that DEA is used to provide an indication of a physician's relative efficiency of treating all their patients (Ozcan, 2008). Analysis on a procedural level requires that physicians' efficiency to treat individual types of conditions be analysed separately (Ozcan, 2008). To get a sense of the overall efficiency of a physician, therefore, requires multiple analyses of all the distinct conditions treated by the physician. Apart from the extra work involved, this has the disadvantage of creating multiple criteria upon which to base the efficiency of a particular physician. The result of this is that the DEA analysis becomes little better than ratio analysis at determining the overall relative efficiency level of a physician. This idea is simplified by considering an example. Assume that there are three otorhinolaryngologists each exclusively treating three conditions, namely: asthma, sinusitis and otitis media. It is plausible that none of these physicians strictly dominate the others in their ability to efficiently treat patients with these conditions. This is in fact the case if each physician is found to be most efficient in treating one of the conditions. In this study hundreds of physicians are being analysed treating hundreds of conditions making it highly unlikely for a procedural-level analysis to provide a sense of a particular physicians overall level of efficiency relative to its peers. It can only be inferred from a procedural-level analysis that a physician is more efficient at treating a particular condition over its peers. This limits the usefulness of such an analysis to healthcare funders performing profiling activities.

Furthermore, in order to adhere to the reliability requirement essential to all profiling methodologies, a procedural-level analysis can only be effectively performed for conditions treated by a significant number of physicians. In addition, these physicians will have to have treated a large enough volume of

patients with the condition. It is likely that few conditions will meet these criteria. Therefore, the procedural-level analysis can only give a very partial view of a physicians overall level of efficiency.

The major advantage of the procedural-level analysis is its increased ability to homogenise the DEA process – a key DEA requirement discussed in Section 5.2.2 above. The practice-level analysis considers the efficiency of physicians' ability to treat patients requiring treatment for conditions that potentially vary considerably. Revisiting the previous example, one otorhinolaryngologists may exclusively treat asthma patients while the other two may treat an equal mix of all three conditions. As a result, it is imperative when undertaking a practice-level analysis to adjust for case-mix differences according to the resource intensity of the conditions treated (as discussed above in Section 5.2.3). A significant consequence of this is that the practice-level analysis suffers from the limitation that the accuracy of the results reflects the effectiveness of the case-mix adjustment technique used. Any heterogeneity not accounted for may jeopardise the level of accuracy of the results obtained (Coelli et al., 2005).

In addition, when interpreting a particular physician's efficiency score obtained from a practice-level analysis, the case-mix of the physician needs to be compared to that of its peers. This is necessary in order to ascertain whether it is clinically appropriate to compare the efficiency of particular physician to that of another. For example, a surgeon exclusively performing organ transplants should never be compared to one that performs predominantly colonoscopies. Even though the case-mix adjustment homogenises the two surgeons regarding resource utilisation, the nature, level and extent of any efficiency improvements achievable will be vastly different for the two surgeons given the varied clinical nature of their practices. This will require the results of a practice-level analysis to be interpreted with the aid of an individual with expert clinical knowledge of the physicians' practices in order to make sound judgements in this regard.

Even though a procedural-level analysis will not require as extensive case-mix adjustment, it will still require case-mix adjustment regarding the severity of conditions treated (Ozcan, 1998; 2000). In addition, interpretation of the results needs to consider physician characteristics. For example, is it appropriate to compare two otorhinolaryngologists against each other on their ability to treat asthma where the one is an allergist possessing far great experience in treating such a condition? Or where one physician treats three times as many asthma patients as the other; thus benefiting from the scale and experience efficiencies this affords? Consequently, the results of a procedural-level analysis need to be interpreted in light of the characteristics of the physician; both contributing and impeding their ability to efficiently treat patients with a specific condition (Chilingerian & Sherman, 1990). Such characteristics may be specific to a particular condition, further adding to the complexity in determining an overall view of a physician's efficiency using a procedural-level analysis.

A further complication of a procedural-level analysis is the inability to utilise classic DEA models. As is discussed in Section 5.4 below, classic DEA models make an assumption that the DMUs either achieve optimal efficiency through reduction of input levels (input-orientated) or through augmentation of output levels (output-orientated) (Ozcan, 2008). When undertaking a practice-level analysis, the input-orientated approach rests on the assumption that the physician does not actively attempt to augment the number of patients seen (Ozcan, 2008). Therefore, it assumes that the physician treats as many patients as demand the services for which the physician has the clinical skills to provide. This is a 'weak' assumption to make as it is highly plausible to occur in reality. When performing a procedural-level analysis this assumption is a lot more difficult to make. Illustrated using an example, an otorhinolaryngologist is likely to favour the treatment a particular condition (e.g. asthma), especially if he/she specialises in treating such a condition. As a result, using an input-orientated DEA approach when performing a procedural-level analysis is not appropriate, as the physician

may actively seek to augment their outputs (the number of patients treated with a particular condition). On the other hand, an output-orientated approach is also not appropriate since it cannot be assumed that the physician does not have control of the resources involved in treating their patients. As a result, a non-orientated DEA model is likely to be most appropriate (Ozcan, 2008). Such non-classic models add considerable complexity when carrying out and analysing DEA results.

A final limitation of the procedural-level analysis is the dependence on the ability to accurately categorise patients into homogenous groups based on their conditions. Complications may arise due to poor billing practices by the physician as well as where patients have multiple conditions and/or comorbidities. Any deficiencies in the grouper system used to categorise patients become direct deficiencies in the DEA analysis.

Upon consideration of the above issues, this study undertakes to perform a practice-level DEA analysis of physician efficiency. It is felt that such an analysis is best suited to answering the research question, which considers whether DEA is a useful profiling tool to provide an indication of physicians' overall relative efficiency. The methodology followed in this study is not the only option when undertaking a DEA profiling analysis. The decision to perform a practice-level analysis reflects the view that it is most useful DEA profiling approach at this early stage of investigations of this kind in South African. This decision, therefore, in no way detracts from the value (or otherwise) to be gained from undertaking a procedural-level analysis. Finally, it is essential that the limitations arising when performing a practice-level analysis (discussed above) be kept in mind when analysing the results of this study. These limitations are considered in Sections 7 & 8 when the results are interpreted and conclusions are drawn.

5.3 Technology

Once the DMUs, inputs and outputs have been chosen, and performance efficiency has been defined; the next step in the evaluation of efficiency using DEA is to define an appropriate technology. The technology specifies the set of combinations of input and output levels that are possible in the environment in which the DMUs production is taking place (Agrell & Bogetoft, 2001). In other words, it defines the input levels that can actually produce respective levels of output. Bogetoft and Otto (2010, p. 57) explain that the "technology shows how inputs can be turned into outputs, how inputs can be substituted for each other, how outputs depend on inputs, and whether outputs are the result of a joint or a united process". The social, technical, mechanical, chemical, and biological environment in which the production process takes place determines the technology (Bogetoft & Otto, 2010). Defining the technology is fundamental to the DEA process because it determines the set of possible performance outcomes against which the actual performance of a given firm can be evaluated (Bogetoft & Otto, 2010). In Figure 5.1, the technology is T^*; it represents all the input-output combinations possible under the 'DEA frontier' curve.

The problem that often arises in practice is that there is insufficient *a priori* information about the true underlying technology (i.e. the true underlying technology is unknown) (Agrell & Bogetoft, 2001). It is therefore necessary to estimate the technology set based on observed data points and then to evaluate the observed production of a firm relative to the estimated technology. Figure 5.1 explains this idea graphically. If there is full *a priori* information regarding the level of efficiency that is technologically possible for firms to achieve, then the best practice frontier will be the 'true best practice' frontier. The corresponding technology with be all the input-output combinations under this curve, represented by T. However, in practice there is insufficient *a priori* information to determine the 'true best practice' frontier. So instead, DEA uses the observed input and

output levels of the DMUs to determine an estimate for this frontier and the corresponding technology; as are represented by the 'DEA frontier' and T^* in Figure 5.1. As a result, DEA does not determine the level of efficiency based on what is actually technologically possible in the environment. Instead, it is based on the optimal level of efficiency achieved by the analysed set of DMUs.

5.3.1 Basic determinism postulate and minimum extrapolation principle

The first requirement when estimating the DEA technology is known as the *basic determinism postulate*, "which states that the technology set should contain all observed DMUs" chosen to be analysed (Kuosmanen, 2001a, p. 3). It is this requirement, which results in efficiency being measured relative to other DMUs and not relative to an absolute norm of what is technologically achievable (Kuosmanen, 2001b). The basic determinism postulate, thus, requires that the technology be estimated by using the observations as the starting point. The technology set is then enlarged by adding assumptions that portray plausible characteristics of the production environment (Kuosmanen, 2001b). These assumptions provide the framework as to how the observations can be interpolated and extrapolated (Bogetoft & Otto, 2010), and thus determine the shape of the efficiency frontier generated from the efficient DMUs. From here on, the assumptions applied to the technology are referred to as 'production assumptions' in order to differentiate them from the other assumptions made in DEA. It is critically important to understand which production assumptions one can reasonably make, explicitly or implicitly, so that the resulting estimation of the true underling technology is congruous with the actual observations of input and output levels upon which it is based (Bogetoft & Otto, 2010). Therefore, another requirement of the inputs and outputs chosen can be added to those discussed in Section 5.2.2. It is imperative that the inputs and outputs in the DEA profiling process adhere to the production assumptions constraining the technology set.

Before the individual production assumptions are discussed, it is important to note that DEA is not the only process that substitutes an underlying unknown technology set with an estimated one. This is commonly done in efficiency evaluations using traditional statistical methods and accounting approaches, among others (Bogetoft & Otto, 2010). However, the manner in which the DEA process estimates the technology is different from that of other methods. The DEA process estimates the technology using the *minimal extrapolation principle* (Agrell & Bogetoft, 2001; Banker et al., 1984; Kuosmanen, 2001b). This means that during the DEA process the smallest possible technology set is constructed to contain all the observed DMU's input-output combinations; as well as to satisfy the set of chosen production assumptions (Kuosmanen, 2001b). DEA thus makes a conservative estimate of the technology set, which results in a conservative estimate of a DMUs level of efficiency, as well as any loss due to inefficiency (Banker et al., 1984).

The combination of the basic determinism postulate and the minimal extrapolation principal is that (1) no observed DMUs efficiency is determined relative to the absolute level of what is technologically achievable, and (2) DEA provides conservative estimates of efficiency. This combination results in a relative "best practice" approximation of the efficiency that is cautious (Kuosmanen, 2001b). A popular understanding of the above is that DEA estimates the underlying technology so as to present the DMUs 'in the best possible light' (Bogetoft & Otto, 2010). In other words, the DEA profiling approach will provide each physician with the highest efficiency level possible given the efficiency levels achieved by other physicians profiled.

The minimal extrapolation principle in DEA is not, however, a given. It depends on the production assumptions imposed when estimating the underlying technology. It is, therefore, essential to show that the minimal extrapolation principle holds when applying DEA to physician profiling. Banker et al. (1984) and Bogetoft and Otto (2010) rigorously prove that the minimum extrapolation

principle always holds if it is appropriate to assume that the technology exhibits free disposability, convexity and standard return-to-scale properties (these are discussed in detail below). The appropriateness of these production assumptions (and, therefore, the minimum extrapolation principle) is discussed in the Section 6.6.

5.3.2 Production assumptions

Different DEA models are distinguished by the set of production assumptions imposed on the technology set. The essential assumptions applicable to classic DEA models are: non-negativity, weak essentiality, free disposability, convexity, and return-to-scale properties (Agrell & Bogetoft, 2001; Bogetoft & Otto, 2010; Coelli et al., 2005). It should be noted that all these assumptions are considered 'weak' in the sense that they are most often fulfilled in practice and they contain limited power in extending the technology set (Bogetoft & Otto, 2010).

Non-negativity states that the level of the inputs and outputs are finite, non-negative, real numbers (Coelli et al., 2005). This assumption does allow for the level of inputs and outputs to be zero. This means that efficiency is assessed in the positive quadrant of the input-output plane illustrated in Figure 5.1. There are DEA models that allow for negative values of inputs and outputs, but these are not considered in this study. *Weak essentiality* ensures that the production of a positive output is impossible without the use of at least one input (Coelli et al., 2005). Put another way, if all the input quantities are zero then the output is zero (represented as the point of origin in Figure 2).

Free disposability of inputs states that if a certain quantity of outputs can be produced with a given quantity of inputs, then the same quantity of outputs can be produced with more inputs (Bogetoft & Otto, 2010). In other words, surplus inputs can be freely disposed of. *Free disposability of outputs* states that if a given quantity of inputs can produce a given quantity of outputs, then the same input level can also be used to produce less outputs (Bogetoft & Otto, 2010). In other

words, surplus output can be freely disposed of. Essentially, free disposability ensures that an increase in inputs will yield the same or higher level of outputs (Coelli et al., 2005).

Convexity states that if two input-output combinations are feasible then all weighted averages (convex combinations) of the two are also feasible levels of production (Bogetoft & Otto, 2010). This allows for the interpolation of the efficiency scores of observed DMUs. In so doing, efficiency scores can be determined for hypothetical DMUs lying between the observed DMUs. Convexity thereby extends the technology, which in turn enables us to rely on fewer observations and still attain credible results (Bogetoft & Otto, 2010). Illustrated graphically using Figure 5.1, convexity allows the comparison of, for example, the inefficient DMU *A* to the efficient hypothetical DMU *A**; the convex combination of efficient DMUs *B* and *C*. The convexity assumption therefore allows for the completion of the 'DEA frontier' by filling in the gaps between the observed efficient DMUs.

Free disposability and convexity hold only if inputs and outputs are divisible and not subject to congestion (Kuosmanen, 2001a). Inputs and outputs are *divisible* if they need not be integers. For example, commodities and monetary amounts are easily divisible while inputs or outputs like the number of employees are not (Kuosmanen, 2001a). Without divisibility the interpolation and extrapolation of inputs and outputs required for the convexity assumption to hold, is not possible. Cooper et al. (2011b, p. 174) explain that "*congestion* is said to occur when the output that is maximally possible can be increased by reducing one or more inputs without increasing any other input or decreasing any other output. Conversely, congestion is said to occur when some of the outputs that are maximally possible are reduced by increasing one or more inputs without reducing any other input or increasing any other output". Cooper et al. (2011b) give an example of a coalmine that analyses its efficiency using a 1-input 1-output DEA model. The input is the number of miners and the output is the quantity of coal produced. Therefore,

producing more coal with fewer miners increases efficiency. The coalmine is subject to congestion if increasing the number of miners makes it possible to form 'teams' to perform tasks at a level of efficiency that is impossible to achieve with a smaller number of miners. Therefore, the highest possible rate of coal production per miner is achieved by increasing the number of miners. The above example illuminates how congestion violates the assumption of free disposability. In addition, economies of scale and scope as well as reduced prices from buying in bulk violate the convexity assumption (Kuosmanen, 2001b).

The *return to scale* assumptions refers to whether rescaling is possible in order to increase efficiency (Ozcan, 2008). Returns to scale properties determine the link between technical and scale efficiency discussed in Section 3.2. As previously discussed, a DMU may be technically efficient but potentially able to achieve increased efficiency by augmenting the scale of its operations. The return to scale property chosen determines whether it is assumed that rescaling is actually achievable by the DMU in order to increase efficiency (Banker et al., 1984). In other words, a DMU should be deemed inefficient if it is *a priori* assumed that it could operate at the optimal scale but it is not currently doing so.

Different assumptions can be made regarding the extent and nature of possible rescaling. The weakest assumption is *variable returns to scale (VRS)*, which assumes that no rescaling is possible. In other words, the model considers only the level of efficiency or inefficiency at the given level of operations for each DMU. This means that the scale of a DMU's production has no bearing on its determined efficiency score. The strongest assumption is *constant returns to scale (CRS)* which allows any possible production combination to be arbitrarily scaled up or down to allow a DMU to operate at its optimal scale (Bogetoft & Otto, 2010; Ozcan, 2008). This means that the scale of a DMU's production does affect its efficiency score since its efficiency is being measured relative to a 'best-practice' frontier of DMUs operating at their optimal scale. In between, there exist *decreasing returns to scale (DRS)* and *increasing returns to scale (IRS)*. DRS

assumes decreasing the scale of operations may achieve increased efficiency but increasing the scale will not and may result in a reduction in the level of efficiency (Bogetoft & Otto, 2010; Ozcan, 2008). Conversely, IRS assumes increasing the scale of operations may achieve increased efficiency but decreasing the scale will not and may result in a reduction in the level of efficiency (Bogetoft & Otto, 2010; Ozcan, 2008). Practical reasons for this are that a larger scale of operations may imply more experience, more efficient processes and a better ability to utilize specialisation opportunities (Bogetoft & Otto, 2010).

To conclude this subsection, Kuosmanen (2001b, p. 10) advises "thinking in terms of the efficiency frontier when considering the production assumptions. Free disposability can be viewed as a first-order curvature condition, convexity can be seen as a second-order curvature condition, while returns to scale properties can be thought of as homogeneity conditions". In addition, it is important to reiterate that the production assumptions refer solely to the assumptions concerning the properties of the underlying technology set. There will be additional (explicit and implicit) assumptions made regarding the other two parts of the DEA evaluation process (Kuosmanen, 2001b). For example, as mentioned in the introduction to this section, it is assumed that the production process does not involve stochastic variations, the data are error-free, as well as the various assumptions relating to properties of the inputs and outputs, discussed in Section 5.2.2.

5.4 Measuring efficiency

Once the production assumptions are made and the resulting technology is defined, the final consideration is the method by which the relative efficiency of the DMUs is to be measured. There are two possible methods. The first is known as the 'CCR ratio efficiency measure' which was proposed in the seminal work on DEA by Charnes et al. (1978). This method stems from the generalisation of simple ratio analysis where a single output to input ratio is used to compare the productivity of DMUs (Cook & Zhu, 2006; Sherman & Zhu, 2006). Charnes et al.

(1978) also developed the second method known as the ‘dual efficiency measure’. This was done by, first, transforming the CCR ratio measure into a linear programming (LP) problem (Cooper et al., 2011b). Second, the associated dual of the LP problem was determined; providing the dual efficiency measure (Coelli et al., 2005). The dual efficiency measure arises due to the duality theorem; which states that every LP problem can be converted into a dual problem providing an upper bound to its optimal value (Boyd & Vandenberghe, 2009). The dual efficiency measure is often also referred to as the ‘Farrell efficiency measure’(Cooper et al., 2011b). It should be noted that the dual theorem ensures that both the above methods reach the same efficiency scores of the DMUs (Coelli et al., 2005; Cooper et al., 2011b). More rigorous explanations of the above DEA efficiency measures are discussed in the sub-sections that follow. The reasons for a particular method over the other are also highlighted.

A further important consideration when measuring efficiency, using either of the methods discussed above, is whether efficiency is measured from an input- or output-orientated perspective. In an input-orientation model, one improves efficiency through proportional reduction of inputs, whereas an output orientation model requires proportional augmentation of outputs (Cooper et al., 2011b; Ozcan, 2008). The choice of whether to use an input-orientated model as opposed to an output-orientated model depends on whether the input quantities are the primary decision variables. If this is the case, the DMU has control over the inputs used in production, as it attempts to minimise the combination of inputs used to produce the observed level of outputs. As such, an input-orientated model is used. If the DMUs are given a fixed quantity of resources (inputs) and with these attempt to maximise the level of outputs, then an output-orientated model will be used (Coelli et al., 2005). Essentially, the orientation is chosen according to which quantities (input or outputs) the DMUs have most control over. Coelli et al. (2005, p. 181) make the important point that “the output- and input- orientated DEA models estimate exactly the same frontier and therefore, by definition, identify the

same set of firms as being efficient. It is only the efficiency measures associated with inefficient firms that may differ between the two methods".

The remainder of Section 5 describes the classic models that emerge from combining the three parts of the DEA evaluation process. The descriptions of the classic DEA model that follows comprise both the mathematical development of the model as well as an explanation of the rationale behind the mathematical formulations. All the DEA models below are described from an input-orientated perspective. The mathematics and rationale of the output-orientated DEA models are, however, largely similar. It is important to reiterate that these DEA models are differentiated by the production assumptions used to estimate the unknown technology set as well as the method of measuring efficiency. Where it is not explicit in the description of the DEA models below, the production assumptions made are stated in parentheses.

5.5 CCR model

The Charnes, Cooper and Rhodes (CCR) model is the initial DEA model developed on the work done by Farrell (1957). Both the 'CCR multiplier model' and the 'CCR dual model' are discussed below.

For explanatory purposes, it is assumed that there are n DMUs being analysed. It is also assumed that each DMU consumes varying amounts of m different inputs to produce s different outputs, and all DMUs have equal access to all the m inputs. An individual DMU_j consumes amount x_{ij} of input i and produces amount y_{rj} of output r (Cooper et al., 2011b). Each DMU_j , therefore, exhibits the production process illustrated in Figure 4.1, where $j=1,2,\ldots,n$.

Furthermore, it is assumed that $x_{ij} \geq 0$ and $y_{rj} \geq 0$ (non-negativity) and that each DMU has at least one positive input and one positive output value (weak essentiality). Throughout the descriptions of the DEA models below, the DMU under consideration is referred to as DMU_o (Cooper et al., 2011b).

5.5.1 CCR multiplier model

The derivation of the CCR multiplier model begins by considering a version of the model referred to as the 'CCR ratio model'. The ratio form evaluates efficiency using the CCR ratio measure discussed in Section 5.4. This definition measures the efficiency of a DMU "as a maximum of a ratio of weighted outputs to weighted inputs subject to the condition that the similar ratios for every DMU be less than or equal to unity" (Charnes et al., 1978, p. 430) . Therefore, the CCR ratio measure of efficiency maximises the ratio of the outputs to inputs of DMU_o. This is then evaluated against the input-output ratios of all the other DMUs being analysed (Cooper et al., 2011b). In so doing, the CCR ratio model determines the optimal relative efficiency between DMU_o and DMU_j where $j=1,2,\ldots,n$ (basic determinism postulate). This model assumes CRS and thus inputs and outputs can arbitrarily be rescaled up or down to achieve optimal scale of production (Cook & Zhu, 2006). The above can be expressed formally as follows:

$$\max h_o(a,b) = \frac{\sum_{r=1}^{s} b_r y_{r0}}{\sum_{i=1}^{m} a_i x_{i0}} \qquad (1)$$

subject to:

$$\frac{\sum_{r=1}^{s} b_r y_{rj}}{\sum_{i=1}^{m} a_i x_{ij}} \leq 1 \quad \text{for j} = 1, \ldots, \text{n} \qquad (2)$$

$$a_i\,, b_r \geq 0 \ \text{ for all } i \text{ and } r \qquad (3)$$

where the variables to be determined are the b_r's and a_i's which are the weights assigned to the y_{ro}'s and x_{io}'s; the observed output and input levels respectively of DMU_o. These weights can be thought of as an expression of the relative importance of the specific inputs and outputs (Cook & Zhu, 2006). The solution

of the above set of equations in a set of optimal weights *(a*, b*)* that maximise $h_o(a, b)$, and thereby allows the calculation of efficiency score $h_o(a^*, b^*)$ for DMU_o. This is done separately for each DMU_j to provide a set *n* of optimal weights and *n* respective efficiency scores.

In equation (1) the numerator is essentially reducing the multiple outputs into a single "virtual output" that is calculated as the weighted linear combination of the *s* outputs. The denominator does the same to the *m* inputs (Sherman & Zhu, 2006). Each time an efficiency ratio is determined, the weights are recalculated so as to maximise the CCR efficiency score of that DMU_o and thus give that DMU the highest efficiency score possible (Coelli et al., 2005). As such, each DMU_j is considered in the "best" light relative to its peers. An optimal efficiency score is, thus, determined for each observed DMU, and each is assigned a set of weights that is most favourable to them (i.e. the combination of the basic determinism postulate and minimal extrapolation principal) (Coelli et al., 2005). As a result, DEA provides a conservative measure of efficiency (discussed in Section 5.3.1). It can also be observed from equation (1) that DEA determines the weights from the data and does not require an *a priori* set of weights for each input and output. Therefore, DEA reduces complexity by attempting to keep *a priori* assumptions to a minimum (Cook & Zhu, 2006).

Inequalities (2) and (3) apply a set of normalising constraints (three for each DMU) to the efficiency ratio being determined. This set of constraints provides a reference as to what will be the highest and lowest efficiency scores achievable. By constraining the ratio of all the DMUs to be less than one, the highest attainable efficiency score is one (Cooper et al., 2011b). Therefore, if a DMU attains an efficiency score of one it cannot improve its efficiency relative to its peers. The DMUs with an efficiency score of one thus make up the efficiency frontier. Constraining the weights to be greater than zero ensures that the minimum relative efficiency score possible is zero (Cooper et al., 2011b). All DMUs with an efficiency score between zero and one are relatively inefficient and

sit in the technology set within the efficiency frontier (illustrated graphically in Figure 5.1).

The problem with the CCR ratio model above is that there are an infinite number of possible sets of solutions for the optimal weights that can be used to calculate the efficiency score $h_0(a^*, b^*)$ (Coelli et al., 2005; Cooper et al., 2011b). This is because if (a^*, b^*) is optimal then so is $(\phi a^*, \phi b^*)$ for all $\phi > 0$. In order to resolve this problem Charnes and Cooper (1962) developed a transformation converting the ratio form into a linear program that selects a solution for (a, b) such that $\sum_i^m a_i x_{i0} = 1$ (Cooper et al., 2011b). This transformation ensures that a single set of optimal weights (a^*, b^*) is determined for $h_o(a^*, b^*)$. Formally, the "Charnes-Cooper" transformation applied to the weights are:

$$\beta_r = \frac{b_r}{\sum_i^m a_i x_{ij}} \quad \text{for j} = 1, \dots, \text{n}$$

$$\alpha_i = \frac{a_r}{\sum_i^m a_i x_{ij}} \quad \text{for j} = 1, \dots, \text{n}$$

This transformation turns the CCR ratio model into an LP problem, deriving the CCR multiplier model, expressed formally as:

$$\max z = \sum_{r=1}^{s} \beta_r y_{r0}$$

subject to:

$$\sum_{r=1}^{s} \beta_r y_{rj} - \sum_{i=1}^{m} \alpha_i x_{ij} \leq 0$$

$$\sum_{i=1}^{m} \alpha_i x_{i0} = 1$$

$$\alpha_i \, , \beta_r \geq 0$$

where the β_r's and α_i's are now the variables of interest and due to them being transformed, are now referred to as 'multipliers' as opposed to weights. The model aims to determine the single optimal set of multipliers *(α^*, β^*)* with their respective efficiency score *z**. This is repeated for each of the *n* DMUs. As in the CCR ratio model, DMUs obtaining an efficiency score of one make up the frontier and the DMUs between zero and one fall in the technology set within the frontier (Figure 5.1).

5.5.2 CCR dual model

The LP problem above can now be transformed into its dual form (as discussed in Section 5.4). The CCR dual model is also commonly known as the 'envelopment model' or the 'Farrell model' (Cooper et al., 2011b). Formally, the dual model is expressed as follows:

$$\theta^* = \min \theta$$

subject to:

$$\sum_{j=1}^{n} x_{ij} \lambda_j \leq \theta x_{i0} \quad i = 1,2, \dots, m \quad (5)$$

$$\sum_{j=1}^{n} y_{rj} \lambda_j \geq y_{r0} \quad r = 1,2, \dots, s \quad (6)$$

$$\lambda_j \geq 0 \qquad j = 1,2, \dots, n \quad (7)$$

where the variables of interest are the weights (λ_j) and (θ). An advantage of the dual model is that it involves fewer constraints than the multiplier model. This means less computing time, and thus the dual model is generally preferred (Coelli et al., 2005). This model, like the multiplier model, assumes CRS. It is important to reiterate that by virtue of the dual theorem of LP problems, the efficiency score determined under the dual model θ^* equals z^*; the efficiency score determined under the multiplier model. Hence, the model may be used interchangeably (Cooper et al., 2011b).

Another advantage of the dual model is that it has a more intuitive interpretation (Coelli et al., 2005). The right hand side of inequality (5) attempts to proportionally reduce the level of inputs of DMU_o by the amount θ, while the left hand side of inequalities (5) and (6) ensure that the resultant reduced input level is still within the feasible input and output sets defined by the technology (Bogetoft & Otto, 2010). In other words, with reference to the input and output levels of the other DMUs, by what proportion θ can the inputs of DMU_o be reduced in order to increase its efficiency. If that proportion is equal to one then it is not possible for the DMU to operate more efficiently than it currently is (Bogetoft & Otto, 2010).

Here θ is minimised in order to cast each DMU in the 'best' possible light, as in the multiplier model. Again in keeping with the multiplier model, all relatively efficient DMUs have a θ^* equal to one and these DMUs make up the efficiency frontier. All relatively inefficient firms will have $0 < \theta^* < 1$ and fall in the technology set within the efficiency frontier (Figure 5.1). For all inefficient DMUs, θ^* represents the proportional decrease in inputs they need to achieve in order to be considered relatively efficient (Coelli et al., 2005).

The key advantage of the dual model is that for all relatively inefficient DMUs, the left hand side of inequality (5) produces a weighted average of efficient DMU's input levels that are used to determine the efficiency score, θ^* (convexity)

(Bogetoft & Otto, 2010). Through this process the dual model not only determines a measure of relative efficiency for each DMU, but also determines the DMU's 'efficiency reference set' (ERS) also known as 'peer units' or just 'peers' (Coelli et al., 2005). Bogetoft and Otto (2010) explain that the left-hand side of inequalities (5) and (6) define the 'reference unit' against which DMU_o is compared. The reference unit is a convex combination of efficient DMUs with weights λ_j representing the projection onto the efficient frontier. The set of efficient DMUs that combine to make up the 'reference unit' are DMU_o's peers. Intuitively, the peers are the perceived efficient DMUs against which an inefficient DMU is most clearly determined to be inefficient (Coelli et al., 2005). They can also be thought of as the set of efficient DMUs from which an inefficient DMU's level of inefficiency has been determined. The individual weights λ_j can then be thought of as ranking the individual peers themselves (Cooper et al., 2011b). The bigger the value of λ_j associated with a particular peer, the greater its individual influence in determining the inefficient DMU's score θ^* (Bogetoft & Otto, 2010).

5.5.3 CCR dual model with slacks

It is necessary to consider the concept of slacks and the resultant difference between 'DEA efficiency' and 'weakly DEA efficient' (Cooper et al., 2011b, p. 10). As mentioned earlier, the models considered here are input-orientated and thus are projected onto the efficiency frontier through a proportional reduction in inputs. However, it is common in many DEA models that a DMU is projected onto the vertical or horizontal parts of the efficiency frontier (see the 'DEA frontier' in Figure 5.1) (Bogetoft & Otto, 2010). In these cases, even after the proportional reduction in inputs, efficiency can still be improved by a further deterministic reduction in one or more inputs, or a deterministic augmentation of one or more outputs. Ozcan (2008, p. 29) explains that "slacks exist only for those DMUs identified as inefficient. However, slacks represent only the leftover portions of inefficiencies; after proportional reductions in inputs or outputs, if a

DMU cannot reach the optimal point on the efficiency frontier, slacks are needed to push the DMU to the this point on the frontier".

In order to determine the slacks in a DEA analysis, a second-stage LP problem is required to be solved subsequent to determining the efficiency scores using the dual model (Cook & Zhu, 2006). Thus, if an input-orientated dual model is used as the first-stage linear program, the corresponding second-stage linear program used to determine the slacks is expressed formally as:

$$max \sum_{i=1}^{m} s_i^- + \sum_{r=1}^{s} s_r^+$$

subject to:

$$\sum_{j=1}^{n} x_{ij}\lambda_j + s_i^- = \theta^* x_{i0} \quad i = 1,2,\dots,m;$$

$$\sum_{j=1}^{n} y_{rj}\lambda_j + s_r^+ = y_{r0} \quad r = 1,2,\dots,s;$$

$$\lambda_j, s_i^-, s_r^+ \geq 0 \; \forall \; i,j,r$$

where the variables of interest are s_i^- and s_r^+, representing the input and output slacks respectively.

This leads to the differentiation of DEA efficiency and weakly DEA efficient. Cooper et al. (2011b, p. 10) state that the performance of DMU_o is "DEA efficient if and only if (1) $\theta^* = 1$, and (2) all slacks $s_i^{-*} = s_r^{+*} = 0$". The performance of DMU_o is "weakly DEA efficient if and only if both (1) $\theta^* = 1$, and (2) $s_i^{-*} \neq 0$ and/or $s_r^{+*} \neq 0$ for some i or r in some alternate optima".

5.6 BCC model

The BCC model was first proposed by Banker et al. (1984) and added an additional constraint to the sum of the weights (λ_j) in the CCR dual model in order to alter the return to scale properties from CRS to one of VRS, DRS or IRS. Formally, the options of the constraints that can be added to the dual model weights are:

$$Add \quad \sum_{j=1}^{n} \lambda_j = 1 \quad j = 1,2,\dots,n \quad for\ VRS$$

$$Add \quad \sum_{j=1}^{n} \lambda_j \leq 1 \quad j = 1,2,\dots,n \quad for\ DRS$$

$$Add \quad \sum_{j=1}^{n} \lambda_j \geq 1 \quad j = 1,2,\dots,n \quad for\ IRS$$

The result of added constraint is that the CRS assumption is deleted and instead the DEA model assumes VRS, DRS or IRS. The returns-to-scale property of interest in this study is VRS; the weakest returns-to-scale assumption. If the model exhibits VRS then it no longer allows any rescaling of inputs and outputs. Banker et al. (1984, p. 1084) elaborate that the CRS assumption "enables extrapolation of the performance of the most efficient DMUs (i.e. allows rescaling with efficient scale sizes for their given input and output mixes) and identify any scale inefficiencies that may be reflected in the level of operations of other DMUs". The VRS assumption, on the other hand, "restricts attention strictly to production inefficiencies at the given level of operations for each DMU (i.e. no rescaling allowed), and thus develop an efficiency measurement procedure that

assigns an efficiency rating of one to a DMU if and only if the DMU lies on the efficient production surface, even when it may not be operating at the most efficient scale size".

6 Methodology

This section describes the method followed in adapting DEA for the purpose of profiling physician performance. This is done by addressing, in turn, the three parts involved in evaluating efficiency using DEA (discussed in Section 5). Table 6.1 provides a summary of five prominent international DEA profiling studies. Decisions regarding the nature of inputs and outputs included in the DEA profiling methodology below are justified with reference to these studies.

6.1 Setting the profiling objectives and defining efficiency

Medscheme, South Africa's largest medical scheme administrator, provided the data used in this study. The reason for their generosity reflects their interest in improving the managed-care services they provide to their client medical schemes. Exploring new profiling methods (such as DEA) may provide Medscheme the ability to more effectively evaluate physician performance on the part of their clients. Therefore, even though this study makes use of data provided by a third-party administrator, profiling is still performed from a medical scheme's perspective.

The methodology began with an in-depth discussion with a Medscheme employee in order to gauge key profiling objectives from a funder's perspective. It is important to frame the profiling objectives from the funder perspective and not from the physician perspective. If physicians' were attempting to determine the efficiency of their own practices then the performance issues of concern would be, for example, the cost of their labour in treating their patients, the cost and quantity of materials used, their capital expenditure on tools and appliances, the rent cost of their consultation room; among others (Coelli et al., 2005). The funder, on the other hand, is concerned with the quality and utilisation rate of the healthcare services that they reimburse for the treatment of patients.

Table 6.1 Summary of prominent international studies evaluating physician performance using DEA

Study	*Physicians*	*Inputs*	*Outputs*
Chilingerian and Sherman (1990)	15 cardiac surgeons	Cost of: ▪ Hospital services ▪ Ancillary services	Number of: ▪ High severity discharges ▪ Low severity discharges
Chilingerian and Sherman (1995)	24 internists 12 surgeons	Cost of: ▪ Hospital services ▪ Ancillary services	Number of: ▪ High severity discharges ▪ Low severity discharges
Chilingerian and Sherman (1997)	326 general practitioners	Quantity of: ▪ Hospital days used ▪ Ambulatory units ▪ Office visits ▪ Referrals to sub-specialists ▪ Mental health visits ▪ Therapy units ▪ Tests ▪ Emergency room visits	Number of: ▪ Infants and children ▪ Females 20 – 39 ▪ Males 20 – 39 ▪ Females 40 – 59 ▪ Males 40 – 59 ▪ Females 60+ ▪ Males 60+
Ozcan (1998)	160 general practitioners	Cost of: ▪ GP services ▪ Specialist services ▪ Hospital services ▪ Prescriptions ▪ Laboratory procedures	Number of: ▪ Low severity otitis media patients ▪ Medium severity otitis media patients ▪ High severity otitis media patients
Ozcan et al. (2000)	152 general practitioners 24 otolaryngologists	Quantity of: ▪ Visits to the attending physician ▪ Referrals by the attending physician ▪ Emergency room visits ▪ Prescriptions ▪ Laboratory tests	Number of: ▪ Low severity sinusitis patients ▪ Medium severity sinusitis patients ▪ High severity sinusitis patients

A decision is, therefore, required on the criteria upon which physicians' performance is to be based; healthcare resource utilisation, quality of care, or a combination of both. This forms the starting point in determining how physicians' relative efficiency is defined in this study. The decision is made that physicians' performance is to be assessed solely on their relative efficiency in utilising healthcare resources to treat their patients. The reason for this reflects the desire to avoid the complications surrounding the inclusion of quality measures (discussed in Section 5.2). Performance evaluation reflecting quality of care is, therefore, left as future research. By excluding quality measures from the analysis it is assumed that the quality provided by each profiled physician is constant. This is a very strong assumption and its implications on the interpretation of results are discussed in the Sections 8.

Two separate approaches can be used to assess the efficiency of physicians' resource utilisation. First, resource efficiency can be based on the quantity of resources utilised by physicians to treat their patients. Table 6.1 illustrates that this approach was taken by Chilingerian and Sherman (1997) and Ozcan et al. (2000). When this approach is used, physicians' profiles reflect exclusively their technical efficiency (Sherman & Zhu, 2006). Alternatively, resource efficiency can be based on the reimbursement cost to funders of the resources used by physicians. Table 6.1 illustrates that this approach is taken by Chilingerian and Sherman (1990; 1995) and Ozcan (1998). Performance based on cost of resources utilised reflects the combination of physicians' technical and price efficiency (Ozcan, 2008). In other words, physicians' efficiency is based on their ability to both utilise the least inputs to produce outputs, and to source the cheapest inputs to produce outputs (Sherman & Zhu, 2006). It is decided, based on the discussion with Medscheme, that it is more meaningful from a funder's perspective to base performance on the reimbursement costs incurred. The key advantage justifying this approach is that it allows efficiency improvements to be quantified in monetary terms (Chilingerian, 1995). The disadvantage is that the profiling results

will not reflect the extent to which physicians' inefficiency is a result of being technical as opposed to price inefficient (Sherman & Zhu, 2006). This is due to the complicated nature of price efficiency (discussed in Section 4.2.4).

In addition, it is decided that the DEA analysis is undertaken on a practice level (as discussed in Section 5.2.4). The reason for this is that it is deemed to better investigate the research question of focus in this study; the use of DEA to measure the overall resource efficiency of physician practices. Consequently, it is important to highlight the complications of performing practice-level analyses on the methodology. This is detailed in Section 6.5.1 below.

Therefore, the profiling process in this study compares the relative performance of physicians based on their price and technical efficiency of healthcare resource utilisation.

6.2 Conceptualising the production transformation process

When conceptualising a physician's production transformation process, Chilingerian and Sherman (1990, p. 4) advises picturing the physician as "the general manager of a temporary firm that exists each and every time he/she treats a patient". The inputs of the temporary firm are the healthcare services used by the physician to treat patients. The production output is the physician's entire patient population over a given period (Chilingerian & Sherman, 1990). The conceptualisation of the production process differs, however, between GPs and specialists. Specialist profiling involves comparing the efficiency in which a specialist utilises healthcare services to treat their population of patients (Figure 6.1).

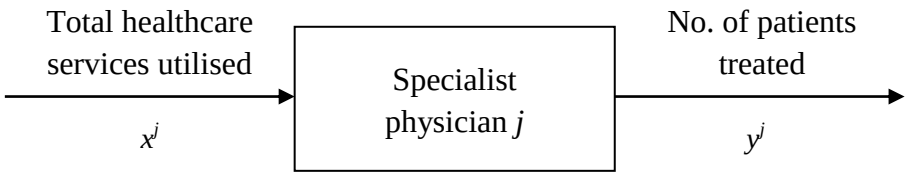

Figure 6.1 Illustration of specialist physician production transformation process

Profiling the performance of GPs, on the other hand, is not only determined by their healthcare resources efficiency but also by the 'downstream costs' that the GP generates (Chilingerian and Sherman, 1997). The GP generates downstream costs when they are unable to fully treat their patients; and thus have to refer patients on to tertiary care provided by specialists. GPs' production process must, therefore, reflect that they are responsible for both the cost of healthcare services they themselves incurred in the treatment of patients, as well as the downstream costs that they generate (Figure 6.2). The above is illustrated in all three GP-profiling studies featured in Table 6.1. Chilingerian and Sherman (1997) and Ozcan et al. (2000) both include the number of referrals as an input. Ozcan (1998) includes the cost of specialist services as an input.

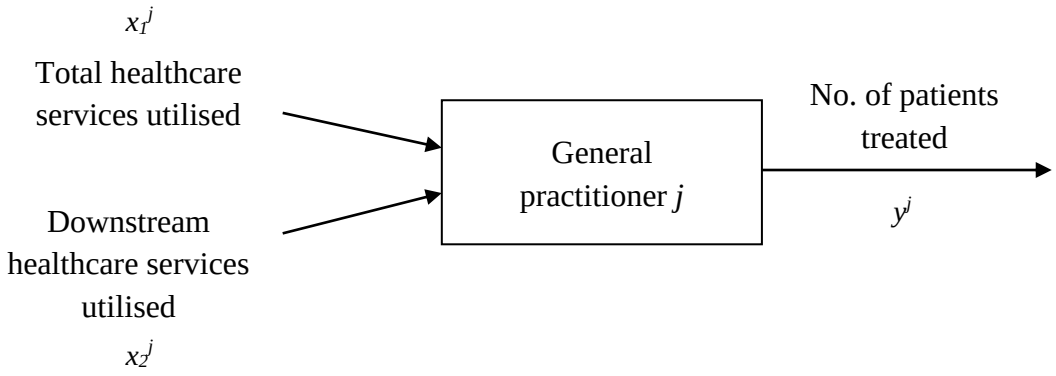

Figure 6.2 Illustration of GP production transformation process

Incorporating downstream costs into a GP profiling process can be complicated. A patient may not always visit the same GP and may seek the opinion of multiple GPs before undergoing tertiary care. The patient may also bypass the GP consultation altogether and go straight to a specialist. Consequently, it is difficult to attribute downstream costs to a particular GP (Thomas et al., 2004). The profiling of GPs is most easily performed in an environment where GPs play the

role of ‘gatekeepers’ (Thomas et al., 2004). This is where all patients are assigned a specific GP, who represents the patient's first contact with the healthcare system. The GP has the sole responsibility to triage patient's further access to the healthcare system (Loudon, 2008). The GP manages their patient’s healthcare services by coordinating referrals, and screening out unnecessary services. All downstream costs can, therefore, be confidently attributed to a particular GP. In the South African private healthcare sector, however, GPs do not act as gatekeepers (McIntrye, 2010).

6.3 Choice of physician type

Physician profiling techniques are performed separately on the different types of physicians providing healthcare services. The reasons for this are discussed in Section 5.2.1. It is decided that the DEA profiling approach will be applied exclusively to a specialist type in order to avoid the complications surrounding GP profiling in a ‘non-gatekeeper’ environment. Figure 6.1, therefore, represents the production process applicable to this study. In addition, it is decided that, due to time constraints in this study, only one specialist type will be profiled. There are, however, a large number of specialities operating in the South African healthcare system. A decision thus needs to be made regarding which speciality to profile.

The choice of speciality needs to be such that the two key requirements of all profiling methods are met (discussed in Section 3.3). First, it must be possible to perform adequate case-mix adjustment on the patient population treated by the chosen physicians. Second, the profiling process needs to exhibit adequate reliability. This requires that there be a sufficient number of physicians of the type chosen and these physicians need to have treated a sufficient number of patients. Consequently, both of these requirements depend on the data available. The steps followed to achieve these two requirements are discussed in detail below in Sections 6.4.1 & 6.5.1 respectively.

Further discussions with Medscheme were held to discuss which specialities best meet the two profiling requirements stated. Medscheme provides administrative and managed care services to (and thus has data on) 18 medical schemes. Many speciality types, therefore, meet the requirement of having a sufficient participation of practicing specialists treating a sufficient number of patients. Focus turned to rather identifying which speciality could most effectively and easily be adjusted for differences in case-mix.

The initial intention was to choose the speciality that treated the most homogenous population of patients with respect to their demographic characteristics, current state of health as well as type and severity of condition being treated. Then only basic risk adjustment would need to be applied to homogenise the resource utilisation of specialists' populations of patients. Unfortunately, even though some specialities are potentially believed to possess a more homogenous case-mix than others it is very difficult to determine this with any degree of certainty.

Consequently, each speciality was individually considered to determine which one most readily allowed for effective and straightforward case-mix adjustment. The speciality type that stood out was general/ paediatric surgeons. The reason for this is that the majority of patients seen by surgeons involve admission to hospital and all hospital admissions are classified into a 'diagnosis-related group' (DRG). Further details regarding DRGs and their use in this investigation for the purpose of case-mix adjustment is discussed in Section 6.5.1 below. Most other specialities treat a large percentage of their patients in private consultation rooms. DRGs are not allocated to procedures performed outside of hospital. Choosing to profile general/ paediatric surgeons, therefore, provided the necessary DRG data, making it particularly suited to effective case-mix adjustment that is straightforward to carry out.

Even though the general/ paediatric surgery is deemed to most suitable speciality type, there are complications posed by this speciality type. These are discussed in Section 6.5.1 below.

6.4 Data

As stated above, DRG case-mix adjustment can only be applied to hospital admissions. To ensure that a DRG could be assigned to all the surgeons' patients, only healthcare services provided to patients admitted to hospital are included in the profiling process. Stated another way, this study profiles the in-hospital services provided by general/ paediatric surgeons. It should be noted that the term 'in-hospital' reflects healthcare services directly related to a hospital admission and not the healthcare facility in which services are provided. Therefore, even though the majority of healthcare services utilised by surgeons are provided in private hospital facilities (e.g. surgery, hospital, anaesthetist, radiology etc.), some may have been provided elsewhere (e.g. pathology and physiotherapy).

Medscheme provided in-hospital data on 504 surgeons treating 81 934 patients in 2012. The data was provided at the beginning of February 2014 and, at that time, the 2012 calendar year was the most current year where full data was available. This is because the 2013 data was not fully run-off[7] and, therefore, approximately three months of 2013 data was not available. The raw data on patients' in-hospital events was provided in the form of claim-line data. This data showed, for each event, the description and amount of all healthcare services claimed for reimbursement. In addition, the DRGs assigned to each in-hospital event were provided. Demographic information was also provided on all the patients.

[7] Claims occurring in a particular year may not be settled by a medical scheme in that year. In addition, claims may not even be reported to the medical scheme by the patient in that year (Bornhuetter & Ferguson, 1972). As a result, claims data will only be complete in the following year, once the data has fully 'run off'; all claims have been reported and settled.

The above three data sets were then condensed into summary data for each in-hospital event. The summary data shows, for each event, the demographic information, DRG and aggregated cost of each of the healthcare services claimed for. The summary data was produced using the data analytics software, Microsoft Access 2007. For samples of the raw and summary data refer to Appendix A. Medscheme reimbursed 38 categories of in-hospital healthcare service utilised by surgeons in the treatment of their patients (Appendix B).

Medscheme runs an algorithm to determine how a surgeon is attributed responsibility for a particular patient's in-hospital event. The details of this algorithm were not, however, shared by Medscheme. It should be noted that this attribution algorithm was relied on and assumed accurate in this study. The summary data for each event was, thereby, assigned to the particular surgeon deemed responsible for that event. The total number of events assigned to a particular surgeon represents the number patients treated by that surgeon in 2012. This is thus the method followed to obtain the pre-case-mix adjusted output used in the profiling process. In discussions that follow, the surgeons' output is sometimes expressed as the number of 'patients' treated. A more accurate description is the number of 'hospital admissions' treated. This is because a patient could have multiple admissions during the study period and this analysis does not aggregate multiple admissions per patient.

It is important to note that some of the surgeons' practices may have multiple surgeons claiming using the same practice identification number. Complicating this is that sometimes they specialise in different types of surgery. These data were not available, but this limitation is not believed to be significant as group practices are not the norm in South Africa. In any case, this is a confounding factor that is very difficult to allow for and additional data are necessary to determine its affect on results obtained.

6.4.1 The use of claims data

The above illustrates that profiling in this study uses 'claims data'. This is detailed data on the healthcare services submitted for reimbursement from the funder (Ferver, Burton, & Jesilow, 2009). It is noted that not all claims will necessarily lead to reimbursement by the funder. This occurs where certain healthcare services do not form part of the cover provided by the funder. Claims data, therefore, represent the cost to the patient of healthcare services utilised in their treatment by the physician. The cost to the funder of patients' treatment may, however, be less. Claims data are used as opposed to 'micro-data'; data collected specifically for the profiling process, such as individual medical records and surveys of physician practices (Ferver et al., 2009).

The reasons for using claims data are the same as those described by McNeil, Pedersen, and Gatsonis (1992, p. 300), that "large claims databases exists and are relatively inexpensive to use. Moreover, they support unobtrusive data collection, support episode-of-illness analysis, and facilitate the longitudinal surveillance of selected patient cohorts". Furthermore, claims data are available in electronic format, good for establishing the cost for certain diagnoses, and avoid the common problem with surveys that individuals may not accurately self-report (Ferver et al., 2009).

The advantage of note is that claims databases provide access to a large volume of clinical data. As discussed above in Section 6.3, in order to ensure statistical reliability profiling requires a large sample size of physicians and for the individual physicians to have seen enough patients (Lasker et al., 1992; McNeil et al., 1992). The consequence of low reliability is that it cannot be discerned whether the measure of physicians' performance is that result of true variation in efficiency or just random chance (Eijkenaar & van Vliet, 2013). There are two explanations as to why this is the case. First, the performance of a physician can be unduly affected by a small number of high-resource use patients. This of

particular effect since case-mix adjustment may not be able to fully allow for extreme-outlier-high-resource patients (McNeil et al., 1992). Second, if a physician sees too few patients it will be difficult to show a significant performance difference between providers, considering that a very small performance difference in either direction will be seen as significant (Luft & Hunt, 1986). Therefore, the use of claims data provides an effective method of achieving the necessary volume of data required to effectively profile the performance of physicians.

In spite of providing access to large volumes of data, claims data will represent only a percentage of each physician's caseload. This is because data is usually only available from a percentage funders providing healthcare financing to only a fraction of the market. In this study, Medscheme provides data on the 18 medical schemes to which they provide services. As such, the data only includes claims information on the patients covered by those medical schemes. Surgeons may have treated many other patients who are covered by other schemes or that paid for treatment out of pocket. Therefore, the profiling process is an analysis of the surgeons' efficiency in treating beneficiaries of these 18 schemes. The results cannot be interpreted as surgeons' efficiency in treating all their patients. Furthermore, efficiency is only measured relative to surgeons that treated beneficiaries of these schemes. Surgeons that exclusively treated beneficiaries of other South African medical schemes and/or patients that paid out of pocket are not included in the profiling process.

A further issue with using claims databases is that they are designed to support patient billing and not for use in physician profiling investigations (Ferver et al., 2009). As a result, claims databases are often missing and/or distort pertinent data necessary for effective profiling. McNeil et al. (1992) provide a number of examples, including incomplete utilisation data, inaccurate physician identifier information and incomplete clinical detail. Incomplete utilisation data is clearly not conducive to comprehensive measurements of physician performance based

on their efficiency of resource utilisation. This issue was discussed with Medscheme. They assured that this is not a significant issue in the data provided due to comprehensive billing practices by hospitals and specialists. The lack of physician identifier information gives rise to attribution problems. This is where more than one physician is involved in the treatment of a patient and it is difficult to determine from the claims data which physician, for example, ordered a particular procedure, prescribed a particular drug or admitted the patient to hospital (Thomas et al., 2004). As discussed in Section 6.2, this is a frequent barrier to the effective profiling of GPs. The outcome of the lack of identifier information is that effective profiling of GPs is often limited to 'gatekeeper' environments (Thomas et al., 2004). The lack of clinical detail can result in the ability to only identify what procedure is performed and not the reason why (Parente, 2002). This makes it impossible to effectively perform case-mix adjustments reflecting the diagnosis of the patient. This is a limited issue in South Africa where regulation requires that healthcare providers assign each claim with an International Classification of Diseases (ICD) 10 code, reflecting the diagnosis for which treatment is provided.

Claims databases are also susceptible to questionable billing practices on the part of physicians (Ferver et al., 2009). In particular, physicians' are often inclined to bill patients in such a way as to ensure reimbursement from the funder. Therefore, the physician may not be billing for services actually provided but for similar services that are covered by the funder and as such are sure to be reimbursed (Ferver et al., 2009). An example of this in the South African context is physicians' incentive for claims to be classified as PMBs because funders are obligated to cover these claims in full (McLeod & Grobler, 2010). The result of this is that claims data may not properly represent the conditions being treated by physicians. Medscheme expressed that effort is exerted to ensure accurate billing by healthcare service providers. They were not, however, able to quantify the effect of poor billing practices on the data provided.

6.4.2 The data-cleaning process

The above limitations require that a comprehensive data-cleaning process is performed to ensure the reliability and accuracy of the data. This is of particular importance when using DEA as it is a non-parametric and does not incorporate a stochastic error term (Coelli et al., 2005). Therefore, DEA is particularly sensitive to the quality of data and any noise present in the data used (Bogetoft & Otto, 2010). The data checking procedures used in this study are those stated by Coelli et al. (2005) as essential to ensure data accuracy. The data were checked for the presence of outliers. This was done by calculating sample means, standard deviations, maximum and minimum values for the total healthcare service costs. In addition, the distributions for each of the 38 individual healthcare services were plotted. All questionable observations were investigated in more detail to ensure their accuracy. The plots of the healthcare service costs were also analysed for the presence of unexpected trends in the data. Furthermore, the zeroes in the data were investigated to determine whether such values were appropriate. For example, having very low or zero costs for hospital services when profiling surgeons' in-hospital admission indicates a problem. Finally, some basic ratios were calculated for all the surgeons, such as the individual healthcare costs per patient. A visual check of plots of these ratios was carried out to reveal further outliers. These data accuracy checks identified 1395 in-hospital admissions as having questionable accuracy. They were thus excluded from the profiling process.

In addition, the claims data was considered on a DRG-level. This was to ensure that there was a sufficient number of each type of condition treated during the period. This is of particular important to ensure an appropriate level of statistical reliability when performing case-mix adjustment (discussed in Section 6.5.1). All DRGs with less than 10 hospital admissions assigned to it were removed. This removed 1322 hospital admission assigned to 379 DRGs; leaving 468 DRGs between which admissions were allocated.

To further ensure the profiling process maintains adequate reliability, the data cleaning process checked that all the surgeons had treated enough patients in 2012. There are two possibilities as to why surgeons may exhibit small caseloads in 2012. First, different surgeons treated different numbers of patients in 2012. There are, therefore, surgeons who, for any number of reasons, treated very few patients in that year. Second, the data provided by Medscheme is claims data and thus represents only a percentage of each surgeon's caseload (as discussed above). To ensure reliability, surgeons that treated less than 21 patients are excluded from the profiling process. The number '21' was chosen as the cut-off by considering the caseload distribution of the 504 surgeons and removing the 20^{th} percentile of surgeons with the smallest caseloads.

The outcome of the data cleaning process revealed that the DEA profiling process undertaken in this study would comprise 403 surgeons who treated 78 135 patients in 2012.

6.4.3 Description of the data

To get a better understanding of the data used it is important to provide a description and summary of the data on the 403 surgeons' practices. This provides context to the analysis as well as allowing for a better understanding of the clinical nature of the data and the complexities surrounding case-mix adjustment, described in the next section. Tables 6.2, 6.3 & 6.4 provides a summary of the data from an admission-, condition/DRG-, and overall practice-level.

Table 6.2 Admission-level summary statistics of data

	Min	*1st Quantile*	*Median*	*Mean*	*3rd Quantile*	*Max*
Patient age	0	31	45	43	58	102
Amount claimed per admission for healthcare service (ZAR)						
Surgeon	400	1 046	1 754	3 235	3 422	341 800
Hospital	500	5105	10 140	22 720	21 350	1 509 000
Pathology	0	0	687	1 799	1 722	204 000
Anaesthetist	0	0	726	1 377	1 663	258 100
Radiology	0	0	0	1 279	949	91 770
Pharmacies	0	0	0	155	129	58 200
Total	1 100	8 400	15 960	33 820	32 470	243 7000

Table 6.3 List of top 30 most common DRGs

DRG	*Quantity*	*% of total admissions*
Other Gastroscopy W/O CC	4681	5.99%
Other Gastroscopy W Major Diagnosis W/O CC	4255	5.45%
Colonoscopy W/O CC	4002	5.12%
Anal and Stomal Procedures W/O CC	3500	4.48%
Circumcision W/O CC	3419	4.38%
Colonoscopy W Major Diagnosis W/O CC	3199	4.09%
Laparoscopic Cholecystectomy CDE W/O CC	2753	3.52%
Appendicectomy W/O CC	2533	3.24%
Other Hernia Procedures (1 + yrs) W/O CC	2334	2.99%
Skin, Subcutaneous Tissue and Breast Plastic W/O CC	2241	2.87%
Minor Procedures for Breast Conditions W/O CC	1951	2.50%
Other Gastroscopy W Major Diagnosis W CC	1895	2.43%
Subcutaneous Tissue and Breast Procedures W/O CC	1877	2.40%
Colonoscopy W Major Diagnosis W CC	1369	1.75%
Laparoscopic Cholecystectomy W Closed CDE W CC	1143	1.46%
Other Digestive System Diagnoses W/O CC	855	1.09%
Fundoplasty W/O CC	753	0.96%
Other Gastroscopy W Major Diagnosis W MCC	664	0.85%
Anal and Stomal Procedures W CC	624	0.80%
Ventral Hernia Procedures W/O CC	594	0.76%
Thyroid Procedures W/O CC	549	0.70%
Major Procedures for Breast Conditions W/O CC	509	0.65%
Vein Ligation and Stripping W/O CC	508	0.65%
Other Hernia Procedures (1 + yrs) W CC	506	0.65%
Other Debridement Procedures W/O CC	430	0.55%
Cellulitis W/O CC	426	0.55%
Major Small and Large Bowel Procedures W MCC	408	0.52%
Major Small and Large Bowel Procedures W CC	397	0.51%
Other Digestive System Diagnoses W CC	395	0.51%
Skin, Subcutaneous Tissue and Breast Plastic W CC	390	0.50%

Table 6.4 Practice-level summary statistics of data

	Min	*1st Quantile*	*Median*	*Mean*	*3rd Quantile*	*Max*
Practice average patient age	0	31	45	43	58	102
Amount claimed per practice for healthcare service (ZAR)						
Surgeon	50 950	269 700	458 700	629 900	734 800	4 726 000
Hospital	351 500	1 797 000	3 340 000	4 422 000	5 644 000	26 090 000
Pathology	13 740	129 300	244 400	350 500	457 700	2 565 000
Anaesthetist	15 430	101 200	214 200	266 700	353 600	2 209 000
Radiology	2 766	90 570	177 200	248 000	330 100	1 383 000
Pharmacies	1 281	8 129	20 500	30 220	41 960	58 200
Total	1 100	8 400	15 960	33 820	32 470	243 7000

The tables above clearly illustrate the high level of complexity that exists when analysing the data from a clinical perspective. There for appropriate case-mix adjustment is essential together with seeking clinical expertise when analysing any results obtained.

6.5 Variable selection process

6.5.1 Choice of output

As stated previously, the profiling process has one output. The precise definition for which is: the DRG case-mix adjusted number of in-hospital events treated by the general/ paediatric surgeons in 2012. Since this analysis is performed on a practice level, the output reflects all the in-hospital patients treated by the surgeon, and not limited to those with a specific condition or DRG.

It should be noted that it is assumed that the surgeons do not have control over number of patients that they treat in the given study period. In addition, it is assumed that a particular surgeon does not have control over the demographic or health characteristics of these patients. In other words, the surgeons treat all the patients who approach them for treatment, regardless of their demographic and

health status, or the severity of the condition needing treatment. This is of course conditional on the surgeon being qualified and capable to treat the patients' conditions. This assumption is made as opposed to assuming that surgeons are able to actively increase the number of patients treated through, for example, marketing and/or actions taken to increase referrals from GPs.

As discussed in Section 6.4, the decision to profile exclusively surgeons' in-hospital services was based on it allowing DRG case-mix adjustment to be performed. All the studies illustrated in Table 6.1 performed case-mix adjustment by using multiple outputs, each representing different risk adjusted quantities of treated patients. The DRG approach is chosen instead in this study because it homogenises the case-mix effectively while still keeping the number of variables as few as possible (Salem-Schatz et al., 1994).

Case-mix adjustment using DRGs was first proposed by Fetter, Shin, Freeman, Averill, and Thompson (1980). DRGs are designed only for the classification of hospital admissions. DRGs can be described as resource homogeneous units of hospital activity based on ICD diagnoses, procedures, age, sex, discharge status, and the presence of complications or co-morbidities (Fetter & Freeman, 1986). Therefore, patients within each category are clinically similar and are expected to use the same level of hospital resources. The DRG approach is the most common case-mix adjustment technique for in-hospital events.

The process of using DRGs for case-mix adjustment involves generating a case-mix index (CMI). The DRG CMI calculated in this study is based on the claims data for the admissions included in the profiling process. The in-hospital events are classified into DRGs. Thereafter, the average cost of the in-hospital events assigned to each DRG are compared against the overall average of treating any hospital admission (Thompson, Fetter, & Mross, 1975). Therefore, each DRG is assigned a relative average value that indicates, on average, the cost of resources required to treat patients in that group, as compared to all the other DRGs. This

relative average value is called the 'DRG case-mix ratio'. The above highlights the importance of cleaning the data on a DRG-level (discussed in Section 6.4.2 above). If there are very few admissions in a particular DRG, then the average costs of hospital events calculated for that DRG will be based on insufficient data leading to potential inaccuracy. Making sure that only the DRGs with at least 10 admissions attempts to avoid this problem and thus maintain the statistical reliability of the case-mix adjustment process.

A ratio with a value greater than one means that in-hospital admissions classified with that DRG are, on average, more resource cost intensive than the norm (Fetter & Freeman, 1986). The opposite applies for a DRG case-mix ratio with a value of less than one. For example, a DRG with a case-mix ratio of value two indicates that, on average, patients assigned that DRG utilise twice the healthcare service costs than a patient assigned a DRG with ratio value one. Another way to think about this is that treating a patient with a case-mix ratio of value two is equivalent, from a resource cost perspective, to treating two single patients both with ratios of value one. The list of all case-mix ratios, one for each DRG, forms the CMI. The case-mix adjusted number of patients is then determined for each surgeon by summing the values of the DRG case-mix ratios allocated to each of their patients (Fetter & Freeman, 1986; Salem-Schatz et al., 1994).

The potential downside of choosing surgeons as the speciality type to profile is the large number of surgery sub-specialities. As this analysis is a practice-level analysis, choosing a speciality that provides treatment to a large number of conditions results in the study being susceptible to the case-mix adjustment process not being able to allow for all the heterogeneity among the surgeons practices. In addition, even if the DRG case-mix adjustment is able to homogenise surgeons' practices based on resource utilisation, it is still necessary to take careful note of the surgeons' case-mix when interpreting the results in order to ensure that surgeons' efficiency is comparable from a clinical perspective. As noted in Section 5.2.4, a surgeon exclusively performing organ transplants should

not be compared to one that performs predominantly colonoscopies. Therefore, when applying the results of a practice-level analysis is essential to interpret the results with an individual with expert clinical knowledge of the surgeons' practices.

The above highlights the significant complication that case-mix adjustment has in the results of a practice-level analysis and the interpretation thereof. It is, therefore, important to reiterate that a significant implicit assumption is being made. It is being assumed that the above case-mix adjustment adequately homogenises surgeons' practices regarding the resource intensity of the conditions treated. Any persisting heterogeneity will subject the results obtained to potential reduced accuracy and interpretability.

6.5.2 Choice of inputs

When treating a patient it is noted that the surgeon chooses, for example, the hospital in which treatment is performed, the anaesthetist used, the amount of pharmaceuticals utilised etc. Therefore, even though surgeons are only reimbursed for the services they personally provide; from the funder's perspective, the surgeon determines the level of the total costs reimbursed for all services utilised to treat a patient. Chilingerian and Sherman (1990, p. 4) explain that "it is the physician who is ultimately in charge of the patient's care and recovery, and depending on the requirements of the patient as interpreted by the physician, each patient receives a unique, highly customized bundle of products and services". Therefore, it is assumed in this study that the surgeons being profiled are responsible for the total cost of healthcare services utilised to treat the patients.

As a result of the above, the profiling methodology need only have one exhaustive input: the total cost of services utilised to treat a particular physician's patient population. The issue with only using the total cost is that the methodology will not be able to provide information to the funder as to the source of the surgeon's inefficiency. This is because it will not be possible to decompose individual

healthcare service cost efficiency from the overall efficiency scores. As a result, the total cost needs to be disaggregated into its component healthcare service costs before the model is run. This is done by assigning individual healthcare service costs as distinct inputs. The question arises as to what is the appropriate level of disaggregation. The model cannot be fully disaggregated, as it will have very little power to discern efficiency if all 38 categories of healthcare service are included as individual inputs.

The starting point of the disaggregation process was to consider which services are most frequently used by surgeons in treating patients as well as which services make up the greatest percentage of the total cost. The cost of the remaining services is aggregated in a single input called 'total other costs'. The rationale here is to include as inputs the services that form the most significant part of physicians' treatment process and will therefore represent important potential sources of inefficiency. Table 6.5 shows the services that are most utilised and which account for the greatest percentages of the total cost.

Table 6.5 Disaggregated healthcare services that are the most significant in the treatment of patients

	Percentage of patients treated with this service	*Cost attributed to this service*	*Percentage of total treatment costs*
Surgeon	100%	R253 867 644.47	9.58%
Hospital	100%	R1 782 151 950.69	67.22%
Pathology	74.03%	R141 234 145.62	5.33%
Anaesthetist	54.46%	R107 477 100.19	4.05%
Radiology	41.37%	R99 962 952.58	3.77%
General Medical Practice	39.55%	R32 240 352.32	1.22%
Pharmacies	38.44%	R12 177 502.63	0.46%
Independent Specialist Practice	16.47%	R49 296 630.79	1.86%
Total Other	52.13%	R172 819 321.21	6.51%

The health services identified in Table 6.5 can only be incorporated as inputs into the DEA profiling approach, however, if they satisfy the variable requirements discussed in Section 5.2[8]. All these inputs are conceptually distinct as they represent services provided by different types of healthcare service provider. Therefore, each input explains the separate role played by that healthcare service in treating surgeons' populations of patients. However, considering the inter-correlations of the services (illustrated in Table 6.6), it is revealed that radiology and pathology services are highly correlated to hospital services as well as to one another. This intuitively makes sense because high hospital service costs are often the result of more complicated and/or severe surgeries needing longer time in the operating room and longer recovery time in the ward. Complicated and/or severe surgeries often also need more tests to diagnose and monitor the patient (such as blood tests and x-rays), both before and after surgery.

Table 6.6 Input variable inter-correlations

Surgeon	1.00								
Hospital	0.78	1.00							
Pathology	0.68	**0.91**	1.00						
Anaesthetist	0.76	0.70	0.58	1.00					
Radiology	0.75	**0.86**	**0.89**	0.62	1.00				
General Medical Practice	0.40	0.52	0.54	0.02	0.52	1.00			
Pharmacies	0.63	0.71	0.69	0.58	0.70	0.45	1.00		
Independent Specialist Practice	0.55	0.73	0.79	0.48	0.76	0.47	0.54	1.00	
Total Other	0.62	0.88	0.86	0.58	0.78	0.43	0.63	0.74	1.00

[8] The requirement of *homogeneity* is ensured by case-mix adjustment discussed in Section 6.5.1. The requirement of *accuracy* is met through the data accuracy checks discussed in Section 6.4.2. The requirements of *measurability* and *quantifiability* are also obviously met. Therefore, only the remaining requirements (*distinct* and *exhaustive*) are considered.

Hence, both higher pathology and radiology services are coupled with higher hospital service costs and, as such, pathology and radiology services are also correlated to one another. In spite of this, it is important to monitor the efficiency of pathology and radiology services to ensure that surgeons are not conducting unnecessary tests on their patients. Therefore, it is decided that it is necessary to keep both of these as inputs in the model, as they represent important potential sources of inefficiency.

The inputs in Table 6.5 are by definition collectively exhaustive. This is because together they represent the total costs of health services utilised by surgeons to treat their patients. However, the correlations between the inputs and output need to be considered. This is in order to ensure that all the healthcare services have individual impact on the case-mix adjusted number of patients treated. Table 6.7 illustrates that 'general practice' and 'independent specialist' services have low correlations with the case-mix adjusted number of patients. Furthermore, both these services contribute small percentages of total costs (Table 6.4). As a result, these services are interpreted to not have significant individual effect in determining the level of outputs. They are re-aggregated into the input 'total other costs'.

Table 6.7 Input – output variable correlations

	Case-mix adjusted no. of patients
Surgeon	0.79
Hospital	0.95
Pathology	0.85
Anaesthetist	0.70
Radiology	0.84
General Medical Practice	**0.57**
Pharmacies	0.74
Independent Specialist Practice	**0.68**
Total Other	0.81

Six disaggregated healthcare services therefore meet the required variable requirements. A parsimonious model is, thereby, achieved with seven inputs and one output. The result of the variable selection process is illustrated in Figure 6.3.

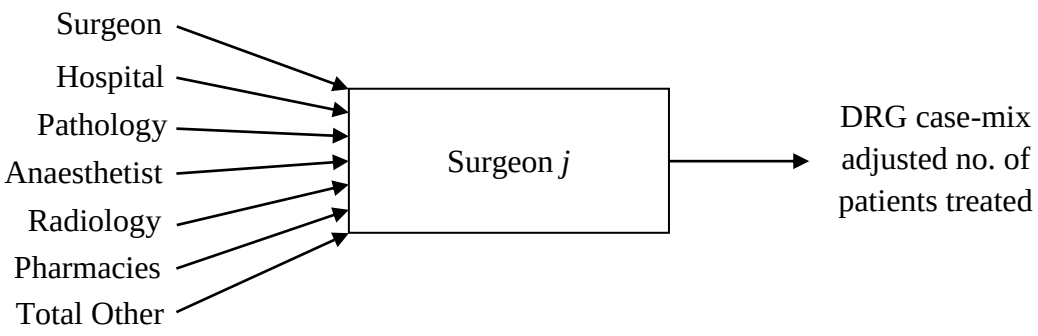

Figure 6.3 Illustration of 7-input-1-output surgeon production process

6.6 Defining the DEA model used

The final step of the profiling methodology is determining which of the classic DEA models, discussed in Section 5, is most appropriate in this study. It has been previously stated that this study is the first to investigate of use of a DEA approach for physician profiling in South Africa. Therefore, only the classic DEA models are considered in this study as it would be premature at this point to investigate the use of more complicated DEA models. The choice of the model to use involves determining which production assumptions can realistically be made when defining the technology set, as well as deciding on the method followed to measure efficiency.

The appropriateness of the production assumptions is considered intuitively. No rigorous mathematical proof of the suitably of particular assumptions is undertaken. The basic determinism postulate holds because all surgeons profiled will be included in the DEA model. The assumptions of non-negativity and weak

essentiality are also easily shown to apply. Non-negativity holds because the inputs are monetary amounts reimbursed for healthcare services and are, therefore, finite, non-negative, real amounts with a minimum value of zero. Weak essentiality holds because a patient can only by treated using of some combination of healthcare services. Therefore, if all the inputs are zero the surgeon will have treated no patients, and thus produced zero output.

The return to scale property that is assumed is that of variable returns to scale (VRS). The reason for this is that it is the easiest return-to-scale assumption to make, as it is the weakest. It therefore avoids the need to prove *a priori* that the surgeon is able to rescale his production in order to optimise his/her efficiency. A consequence of using a VRS DEA model is that the profiling process does not reflect scale efficiency.

The determination of whether the free disposability and convexity assumptions hold is more complicated as both hinge on proving two further assumptions; that the inputs and outputs are divisible and not subject to congestion. The costs of services are by their nature divisible. The case-mix adjusted number of patients is allowed to be any decimal amount, with any rounding necessary performed in the analysis of the results. Therefore, the inputs and output chosen in this investigation are divisible as they need not be integers. The assumption of congestion is simpler to prove in a single output model (as is the case in this study). This is because the congestion assumption becomes the converse of the free disposability assumption. Intuitively, if a surgeon is able to treat his/her patient population with the observed level of healthcare services then increasing his/her utilisation has two possible outcomes. First, the extra services can be used to provide more extensive treatment to same number of patients. Second, the extra services can be used to treat a greater number of patients. Utilising a greater amount of services cannot result in a surgeon treating fewer patients. Therefore, free disposability is intuitively expected to hold. Since all the above production assumptions are met the DEA model exhibits the minimum extrapolation principle

and thus produces a conservative measure of efficiency.

The next it needs to be decided which of the two DEA methods of efficiency measurement will be used. An envelopment model is used because it provides information relative to the reference set of each physician. Therefore, the model will not only provide a measure of efficiency but also provides the set of peers against which a particular surgeon is deemed inefficient. This provides the inefficient surgeon with efficient peers whose practices may be analysed in order to potentially achieve efficiency improvements. The combination of assuming VRS and using an envelopment model means that the efficiency scores are determined using the BCC model discussed in Section 5.3.

Furthermore, an input orientated DEA model is used. This is because it is assumed that the surgeon is in control of all the healthcare services used to treat their patients. Additionally, it assumed that the surgeon has no control of the number, characteristics or types of patients that are treated (discussed in Section 6.5.1 and 6.5.2 above).

Therefore, to summarise, the profiling methodology uses a 7-input 1-output input-orientated BCC model to determine the combination of relative technical and price efficiency of 403 general/ paediatric surgeons in providing in-hospital services in 2012. The model was run using the statistical software R *version 3.0.3*[9]. The DEA algorithms used were obtained from the R-package called 'Benchmarking', developed by Peter Bogetoft and Lars Otto.

[9] R is a language and environment for statistical computing and graphics. It was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and his colleagues.

7 Results

This section discusses the results obtained from the profiling methodology described above. A summary of the profiling results is presented and thereafter a more detailed analysis is conducted. The aim of the analysis is to inform the funder's initial steps in designing and implementing efficiency improvement interventions. It is important to note that the results below are analysed solely from a statistical perspective. Access was not available to an individual with clinical expertise regarding the data and surgeons' practices. Therefore, all results obtained will require additional interpretation together with a clinical expert to determine if they are applicable in practice. In addition it should be noted that graphical representation was found to be of little use in displaying results. This is due to the large number of surgeons profiled. Consequently, the results below are represented using summary statistics.

7.1 High-level summary of surgeon performance

Table 7.1 illustrates a summary of the 403 surgeons' efficiency scores determined by the DEA profiling process. There are 58 surgeons perceived efficient, constituting approximately 14.4% of those profiled. The remaining 345 are found to be able to improve their efficiency to some extent relative to their peers. The profiling process, thus, reveals substantial scope for efficiency improvements. More that 85% of surgeons profiled are potentially able to increase their efficiency by reducing the cost and/or quantity of healthcare services utilised to treat their populations of patients. The mean efficiency score of the inefficient surgeons is 0.68. This indicates that, on average, the surgeons found to be inefficient are required to proportionally reduce their input levels by 32% in order to reach the optimal level of efficiency. The individual surgeon determined most inefficient by the model attained an efficiency score of 0.35. This means that the cost and/or quantity of healthcare services utilised by this surgeon would have to

be decreased by a massive 65% in order to achieve optimal technical and price efficiency.

Since a dual model was used to obtain the efficiency scores, each inefficient surgeon is also assigned their efficient peers and the corresponding weights λ_j. These illustrate the efficient surgeons that each inefficient surgeon was compared to in the determination their efficiency score.

It should be noted that these results do not illustrate the efficiency of surgeons' whole practices. This is because only surgeons' efficiency in providing in-hospital services is profiled. In addition, only price and technical efficiency is assessed. The efficiency scores do not provide insight in surgeons other types of efficiency. The process also may not have included all the patients treated by a particular surgeon. This is due to the feature of claims data (discussed in Section 6.4.1) that only beneficiaries of the medical schemes administrated by Medscheme are present in the process.

Furthermore, these results do not portray surgeons' efficiency relative to all the other surgeons practicing in the South African private healthcare sector. Efficiency scores only reflect efficiency relative to the 403 surgeons included in the process. It is, therefore, also clear that these results are not a reflection of the efficiency of surgeons operating in the South African public sector.

Table 7.1 Summary of surgeon performance

	Efficiency score classification					
	Overall	*Efficient*	*Inefficient*	*0.75 - 1*	*0.5 - 0.75*	*>0.5*
No. of surgeons	403	58	**345**	112	196	**37**
Mean efficiency score	0.72	1	**0.68**	0.84	0.63	**0.44**

7.2 Further analysis

The next step is to consider whether the high-level summary above can be broken down into more detail in order to determine if the surgeons present any clinical characteristics explaining their efficiency, or lack thereof. Unfortunately data were not available regarding the particular characteristics of the surgeons. This would have allowed detailed analysis of the efficiency score based on such external factors as: age, expertise and sub-speciality of the surgeon as well as details on the facilities in which treatments were provided.

It is, however, possible to analyse the results according to the size of the case-load of the surgeon and the variety of patients treated by a surgeon. Table 7.2 illustrates summary statistics of surgeons' efficiency scores according to the number of admissions treated. Observation of the mean efficiency scores in the respective categories points to potential scale efficiencies. Future research to include the measurement of scale efficiency in the DEA profiling process is necessary to investigate this result in more detail.

Table 7.2 Summary of surgeon performance according to case-load

	Case-load ranges			
	1st Quartile	*2nd Quartile*	*3rd Quartile*	*4th Quartile*
Size of case-load	*21-76*	*77-145*	*146-246*	*246-1061*
No. of surgeons	104	98	101	100
No. of efficient surgeons	18	9	9	22
No. of inefficient surgeons	86	89	92	78
Mean efficiency score	0.69	0.68	0.72	0.80
Mean score of inefficient surgeons	**0.63**	**0.65**	**0.69**	**0.74**

Table 7.3 provides a summary of the surgeons' performance based on the variety of admissions treated. The variety of surgeons' admissions is based on the number of unique DRGs treated by the surgeon. It is observed from Table 7.3 that the efficiency score increases as the number of admission with unique DRGs treated increased. This result is unexpected because it is intuitively hypothesised that surgeons specialising in the treatment of particular types of conditions would be able to treat those conditions more efficiently. This result may, however, be distorted by the scale effect illustrated in Table 7.2. This is because it is observed that the number of unique DRGs treated increased with the size of surgeons' case-load.

Table 7.3 Summary of surgeon performance according to variety of admissions treated

	Ranges of unique DRGs treated			
	1st Quartile	*2nd Quartile*	*3rd Quartile*	*4th Quartile*
No. of unique DRGs	*7-37*	*38-61*	*62-86*	*87-215*
No. of surgeons	102	102	98	101
No. of efficient surgeons	22	8	4	24
No. of inefficient surgeons	80	94	94	77
Mean efficiency score	0.71	0.67	0.71	0.80
Mean score of inefficient surgeons	**0.64**	**0.64**	**0.69**	**0.74**

When considering both of the results above it is important to remember that only a portion of surgeons' case-load forms part of the analysis (discussed in Section 6.4.1). As a result, the surgeons may have treated additional patients,

which may have potentially been more varied in nature. The results above can therefore only be interpreted to represent the admissions included in the analysis and not surgeons overall practices.

The above to results are, however, successful at illustrating the clinical complexity that exists between varying surgeon practices. The statistical analysis must consider these complexities in order to form actionable interventions.

7.3 Focusing on the 'problem' cases

When analysing the performance results from a physician profiling process it is important for the funder to consider the extent of inefficiency. Efficiency improvement interventions focused on physicians deemed only slightly inefficient will be difficult to design, for the following reasons. The validity of low inefficiency levels may be questionable since DEA is a non-parametric approach with no stochastic error term. A physician exhibiting low levels of inefficiency may just be reflecting the unaccounted for random variation in the variables chosen or just noise in the data (Cook & Zhu, 2006). This includes persisting heterogeneity in the case-mix of physicians profiled. In addition, it is difficult to determine the source of low levels of inefficiency (Eijkenaar & van Vliet, 2013). Furthermore, what the DEA model deems theoretically possible and what is practically possible may be different. In reality these physicians may often be operating at or very close to optimal efficiency (Bogetoft & Otto, 2010). In addition, getting the cooperation of a physician to agree to interventions with very small potential gain may be difficult and the physician may feel like their practices are being micro-managed by the funder. This can quickly lead to the deterioration of the relationship between the physician and the funder. In any case, it is likely that the funder will save far more financially from focusing on the physicians portraying severe deficiencies in their ability to operate efficiently.

Table 7.1 also breaks the efficiency scores into bands in order to determine those surgeons whose efficiency levels are of most concern. It can be seen that there are 37 surgeons who, in order to be deemed efficient, need to at least halve the quantity and/or price of the healthcare services utilised. Interventions to improve the efficiency should begin by focusing on these 37 surgeons that are of most concern; before moving on to interventions aimed at increasing the efficiency of the 196 surgeons exhibiting approximately average efficiency scores. Finally, interventions aimed at fine-tuning the efficiency of the rest of the surgeons exhibiting low levels of efficiency should be considered, keeping in mind the challenges addressed at the start of this subsection.

7.4 Financial savings of efficiency

Since the inputs are expressed in monetary terms, the analysis can be used to determine the reduction in claims potentially achievable from surgeons increasing their level of efficiency to that of their peers. However, determining the financial savings in this context will only really be a theoretical exercise. The reason for this is that the savings achievable are dictated by the varying types of procedures performed by the surgeons as well as the particular characteristics of the surgeons themselves. Consequently, the claims reduction figures are not included in this analysis because without clinical verification there is no way of providing a sense as to whether these figures are realistically achievable by the surgeons profiled.

It is also important to revisit the discussion from Section 3.4 that the funder has no direct ability to impact on how a physician practices. The funder can only use incentives, education and relationships with the surgeons in order to encourage them to change their behaviour to achieve more efficient practices. Any savings from interventions informed by this study rests on the funder being able to engage effectively with the surgeons. Effective engagement with surgeons will take time in order to generate sufficient trust between the parties. It is also likely that

interventions will need the backing of the professional societies to which surgeons belong.

7.5 Drivers of efficiency

In order to design interventions to improve efficiency it is important to understand which healthcare services are most responsible for whether a surgeon is deemed efficient or inefficient. It is not a given that the healthcare services that offer the greatest scope for financial savings will represent the major determinant of efficiency in the model. This is because it may be the case that surgeons' efficient utilisation of one of the other services drives more efficient utilisation of the remaining healthcare services used.

Sensitivity analysis is performed in order to determine which inputs in the model are most significant in determining the efficiency score. This is done using two approaches. Both make the assumption that the 7-input-1-output DEA model is the 'true' model determining a surgeon's 'actual' level of efficiency. Changes are then made to the nature and form of the inputs used and the sensitivity of the 'true' model to these changes are then analysed. In first approach a 6-input 1-output DEA model is run six times and in each iteration one of the six distinct inputs in the 'true' model is re-aggregated to form part of 'Total Other' costs. (Recall for Section 6.5.2 that the 'Total Other' costs represent the healthcare service costs not initially deemed to be significant enough in the treatment of patients to justify them being disaggregated into a distinct input). The set of six new efficiency scores are then analysed to reveal the service whose re-aggregation had the greatest impact on the efficiency scores, relative to the 'true' model. The idea here is to determine which of the healthcare services initially deemed most significant in the treatment of patients would result in the least impact to the efficiency scores if it were re-aggregated with the 'Total Other' costs. In other words, the healthcare services that result in the least change in efficiency scores relative to the 'true' model have the least descriptive power over the 'Total Other'

costs in discerning the level of efficiency. The healthcare services that result in the biggest change in efficiency scores most need to be distinct inputs in their own right, considering the high level of descriptive power they have. Note that the inputs and outputs in each 6-input-1-output DEA model still meet all the necessary requirements discussed in Section 5.

The second sensitivity analysis approach runs another seven 6-input-1-output DEA models. This time, each iteration completely removes one input from those included in the 'true' model. Therefore, the cost of treatment attributable to that healthcare service no longer features in the model and is no longer a part of surgeons' production transformation process yielding treated populations of patients. The results are studied to reveal the healthcare service that induces the most changed efficiency scores. The models run no longer meet the essential requirement of DEA that the model inputs be exhaustive. In other words, the inputs in the new model no longer capture all the resources having bearing on the output produced and the type of efficiency being assessed. This allows for the measurement of the effect that the excluded input has in determining the efficiency levels of surgeons in the 'true' model.

The above two approaches were conducted by Kittelsen (1993). He explains that the combination of the above two sensitivity analyses provide an indication of the inputs having the greatest effect in altering the 'goodness of fit' of the 'true' model. The first approach, therefore, can be thought of altering the inputs to determine which input upon re-aggregation most alters the *shape* of the production transformation process from that in the 'true' model. The second approach determines which inputs most describe the *actual* production transformation process itself used in the 'true' model.

The healthcare service costs found to be the greatest driver of efficiency are those of the profiled surgeons themselves (Table 7.4). This is an important result and can be explained intuitively. It has previously been explained that surgeons are

not only responsible for the services they provide but also control the treatment process and decide on the particular bundle of healthcare services that their patients will receive. The fact that the surgeon's own services are the greatest driver of efficiency illustrate that surgeons who exhibit efficiency in the provision of their own services relay this onto the rest of their decisions made in the treatment of patients.

Table 7.4 Sensitivity analysis results

	Mean score of inefficient surgeons
'True' model	0.68
Re-aggregated category of healthcare service costs	*Change*
Surgeon	**0.037**
Hospital	0.014
Pathology	0.011
Anaesthetist	0.010
Radiology	0.006
Pharmacies	0.010
Removed category of healthcare service costs	
Surgeon	**0.043**
Hospital	0.058
Pathology	0.012
Anaesthetist	0.013
Radiology	0.002
Pharmacies	0.009
Other costs	0.006

For example, the surgeon may choose other service providers, such as anaesthetists, that demonstrate efficient provision of care. The surgeon may not recommend the patient undergo excessive or expensive radiology and pathology tests or prescribe unnecessary pharmaceuticals. The surgeon may ensure that the

care provided in hospital is not inefficient. This includes making sure that patients are discharged on time and not provided unnecessarily wasteful services, such as, placing a patient in an ICU or high-care ward when adequate recovery can take place in a general ward.

7.6 Dominant peers

Finally, it is useful for the funder to identify the most 'dominant' peers out of the surgeons deemed efficient. 'Dominant' is represented by two features of a peer. First, by how many surgeons are deemed inefficient based on the comparison with that efficient surgeon. Second, by how many surgeons were deemed inefficient based predominantly on the optimal efficiency that surgeon was able to achieve. This is done by determining which of the peers in the reference unit had the highest weight, λ_j[10] .

The benefit of isolating these dominant peers arises when attempting to design efficiency improvement interventions. The 7-input-1-output DEA model in this investigation produced 58 surgeons that were perceived to be operating efficiently in 2012. However, analysing the practices of 58 individual surgeons in order to reveal commonalities that can inform interventions is incredibly difficult. Table 7.5 identifies 8 dominant peers from those profiled. These surgeons together represent the peers that are most significant in determining the efficiency scores of 80% of the inefficient surgeons. This simplifies the job of designing interventions by only having to analyse the efficient practices of these 8 surgeons. At the very least, it provides the 8 surgeons whose practices should form the starting point of any efficiency improvement interventions.

[10] The details surrounding 'reference units' and 'peers' are discussed in Section 5.5.2

Table 7.5 Dominant peers

Peer reference no.	*No. for which significant peer*	*No. for which peer*
236	54	116
187	52	145
209	51	97
186	48	242
61	31	87
400	15	101
382	13	37
211	12	54

8 Conclusions and discussions

The profiling of physicians locally and abroad is widespread, testament to its value to funders in achieving managed care objectives. This dissertation sets out to investigate the potential use of a DEA profiling approach by South African funders. This is done by applying a DEA profiling methodology to evaluate the price and technical efficiency of 403 general/ paediatric surgeons' utilisation of in-hospital services in 2012. This section analyses the DEA profiling methodology followed and the results obtained in order to conclude on its usefulness to South African healthcare funders. This section, first, highlights the attributes of the DEA profiling approach that support its use by South African funders. This is followed by a discussion on the limitations of the DEA approach as well as the further research necessary to address these limitations.

8.1 Attributes of a DEA profiling approach supporting its use

8.1.1 The requirement to conceptualise the production process

Potentially the most valuable aspect of profiling using a DEA approach is that it is entirely based on the physicians' production transformation process. The DEA profiling approach, therefore, necessitates that the profiler fully understand the process by which physicians utilise healthcare services to treat patients. In particular, this involves appreciation of all the healthcare services the physician is responsible for in the treatment of their patients. This ensures the basic requirement that the profiler be clear on the responsibilities of the physicians being assessed before determining whether these are being performed efficiently. In addition, it provides a holistic view of the healthcare services utilised by a physician and prevents misleading outcomes resulting from considering resource efficiency from only one perspective. The risk of not properly understanding physicians' production process is the potential misclassification of efficiency as well as ineffective efficiency improvement interventions.

The above benefit is illustrated by considering the discussion regarding a GP's production process in Section 6.2. If the GPs' production process were not initially considered then profiling may have be performed without allowing for the fact that they are responsible for the level of downstream costs. Consequently, the profiling process may have deemed certain GPs efficient when in fact their efficiency is generated by a high referral rates to more-expensive specialists.

In terms of the profiling of surgeons in this study, the profiler is forced to express efficiency in terms of the specific production process of the surgeons analysed. This ensures that healthcare costs included in the DEA profiling approach represent the particular bundle of services that the surgeon decides their patients need in order to be treated effectively. Therefore, the 403 surgeons' efficiency scores are based on the healthcare services that they are responsible for in the treatment of patients and which they can affect in order to improve efficiency.

This is particularly beneficial to funders in South Africa who outsource managed care functions to administrators or MCOs. Conceptualising the surgeons' production process prevents the blind use, by these 3rd parties, of performance evaluation techniques without determining their suitability for profiling surgeon performance. Funders can, therefore, be more confident that the outcomes of profiling processes and the interventions they inform are effective and achievable.

A proper understanding of the production transformation process is also a major advantage when it comes to presenting inefficient surgeons with their profiles. A surgeon is more likely to cooperate with any interventions if the funder is able to portray an understanding of surgeons' practices. This assures the surgeon that the funder is providing information that they can utilise to improve their efficiency.

8.1.2 Multiple inputs and outputs

DEA explicitly allows for the inclusion of multiple inputs and outputs. This allows for the conceptualised production process to include multiple healthcare

services as inputs. In this study, this allowed for the important healthcare services utilised by surgeons in the treatment of patients to be disaggregated and included as separate inputs in the model. The result of this is that these healthcare services could be analysed to determine their significance in driving surgeons' overall efficiency, and the scope for savings related to each service.

This information is very useful in the design of efficiency improvement interventions. It revealed that the efficient utilisation of hospital services provides the greatest potential for financial savings. However, the surgeons own practice is the greatest driver of efficiency. This means that interventions aimed at motivating surgeons to provide their own services efficiently may lead to more efficient decision-making regarding the other healthcare service utilised in the treatment of their patients.

In addition, the ability to include multiple inputs further allows for a holistic profiling process and averts the complication of profiling individual factors separately. DEA accounts for the inherent relationships among the inputs themselves as well as between the inputs and the outputs. Other profiling techniques may require the individual analysis of each input e.g. risk assessment models. This may provide a distorted view of physician efficiency if the relationships between these inputs and outputs are not understood and defined correctly.

8.1.3 Conservative efficiency estimates

As discussed in Section 5, DEA is unique in the way it estimates the technology set. In this study, the minimum extrapolation principle is shown to hold and thus surgeons are given the highest efficiency score possible based on the input-output combination of their peers. The benefits of this are best explained by considering the 37 surgeons found to be most inefficient. These surgeons are found to be severely inefficient even after the DEA approach provides them with the highest efficiency score possible.

The benefit of a cautiousness measure is that it provides greater confidence that the profiling approach is actually identifying inefficiency. These surgeons' inefficiency cannot be interpreted as a consequence of the standard being set too high against which efficiency is being assessed. On the contrary, these 37 surgeons have achieved poor efficiency despite the efficiency standard being as favourable to them as possible. If the model was not so 'kind' their efficiency scores would have been far worse. Therefore, the DEA profiling approach is particularly suited at identifying those surgeons whose inefficiency is of particular concern. The DEA approach highlights the surgeons that are required to form part of initial efficiency-improvement interventions.

A cautious measure of efficiency provides further benefit when it comes to presenting inefficient surgeons with their profiles. Cautious efficiency scores can help avoid an inefficient surgeon from contesting the results based on being compared to an 'unfair' standard of efficiency. A surgeon may be more obliged to acknowledge their inefficiency if the profiler is able to assure the surgeon that the approach provided them with the highest score possible. This, in turn, may result in improved cooperation by the surgeon to engage in funder's proposed interventions aimed at increasing their efficiency.

8.1.4 Comparison with best-practice peers

As previously stated, DEA does not only provide an efficiency score it also provides reference units of best-practice peers from whom inefficient physicians' scores are determined. Each of the 345 surgeons found to be inefficient can be provided with the set of peers against whom they were perceived to operate inefficiently. Funders can thus provide each inefficient surgeon with personalised information regarding the set of efficient surgeons whose practices they can attempt to emulate in order to increase efficiency.

Comparison with a best-practice frontier also provides further benefits when presenting the results to surgeons. First, the DEA profiling approach frames

surgeons' efficiency relative to what their peers are able to achieve. As previously expressed, recommendations by the funder based on efficiency in this form may lend itself to greater cooperation by surgeons. Second, from a funder perspective, the interventions informed from the DEA profiling approach will encourage surgeons to increase their efficiency to levels achieved by the best performing surgeons. This is opposed to encouraging surgeons to just demonstrate above average efficiency.

In addition, the results of the profiling of surgeons in this study illustrate that the DEA approach allows for the determination of dominant peers. This simplifies the process by which the funder is able to design interventions that individually improve the efficiency of a large percentage of surgeons. Identifying the dominant peers in this study decreased the number of surgeons required to consider when designing interventions by up to 50. The results of the DEA approach reflect that analysis of the efficient practices of just the 8 dominant surgeons has the potential to improve the efficiency of up to 80% of the surgeons profiled.

8.2 Limitations and further research

8.2.1 Quality

The major limitation of this study is the absence of quality measures in the DEA profiling approach. Quality is excluded due to the complication of defining quality measures and the uncertainty regarding where to include quality within surgeons' production transformation process (discussed in Section 5.2). As stated previously, the absence of quality measures results in the implicit assumptions that the care provided by all profiled surgeons is of the same quality. This is a very strong assumption that is at the very least subjective, however, likely to be untrue. Therefore, profiling is only being evaluated from one dimension of performance; the technical and price efficiency of resource utilisation.

There are two competing views regarding the relationship between physicians' quality of care and healthcare service utilisation. One view proposed by Chilingerian and Sherman (1990) is that the quality of physician care is directly proportional to the healthcare resources invested in a patient. This can be explained as a 'more is better' proposition. More medical care, represented by physicians investing more healthcare resources in their patients, will always improve the quality of treatment outcomes. An alternative view expressed by Enthoven (2002) suggests that the utilisation of healthcare resources is subject to diminishing marginal returns with respect to the quality of care provided. For example, assume a patient undergoes a complex surgery. The first day of recovery in a hospital ward is essential and so is the second. However, each subsequent day makes less difference to the patient's health status. In time, an additional day in the ward will yield no further benefit at all (Enthoven, 2002).

If the first point of view is assumed representative of reality, the exclusion of quality from the DEA approach exposes the variation of efficiency scores to be representative of differences in quality of care, and not inefficient resource utilisation in the provision of care. For example, surgeons' resource inefficiency may be sanctioned by the fact that through increased healthcare service costs superior treatment outcomes were possible; represented by, for example, lower mortality and/or readmission rates. In this study, the 37 surgeons perceived to be highly resource inefficient relative to their peers may be able to justify the use of more resources. Their extra utilisation may not be wasteful but representative of the superior quality of treatment they have provided to their populations of patients. Furthermore, the 58 efficient surgeons may be achieving their resource utilisation efficiency at the expense of providing substandard quality of care relative to that of inefficient surgeons. Chilingerian and Sherman (1990) do concede that this point of view is overly simplistic. Intuitively, the actual relationship between quality of care and healthcare service utilisation lies closer to that proposed by Enthoven (2002). It is, however, difficult without empirical

evaluation to be certain as to the effect of this view on the interpretation of surgeons' observed efficiency scores.

Further research is thus required in order to construct a two-dimensional profiling methodology that considers both resource utilisation and quality of care. This requires investigation of the most appropriate quality measures as well as the manner in which these measures should be incorporated in the DEA profiling methodology.

8.2.2 Case-mix

A theme throughout this study is the significance of effective case-mix adjustment to homogenise the DEA process. This is a particular limitation of practice-level analyses. Even though the methodology followed undertook to apply the most appropriate and effective case-mix adjustment process possible, the results are still vulnerable to heterogeneity from any case-mix differences unaccounted for. Further research is necessary on alternative case-mix adjustment processes to determine whether the one performed in this study can be improved upon.

In addition, the results obtained need to be assessed for clinical applicable by an individual with appropriate expertise. This is essential to transform the analysis in this study from a theoretical exercise into one that is able to inform actual efficiency improvement interventions.

Further research on applying procedural-level DEA profiling analyses is also necessary. These analyses are less sensitive to case-mix complications. Comparing the results of a procedural analysis to the ones obtained in this study may be able to provide valuable support to the accuracy of results obtained above.

8.2.3 Non-parametric

One of the defining features of DEA is that it is a non-parametric frontier analysis approach. As previously mentioned, the advantage of this is that it does not

require *a priori* assumptions regarding the distribution characteristics of the inputs, outputs or the production frontier. This avoids jeopardising the validity of profiling results, through efficiency scores determined based on an incorrect choice of distribution function.

There are, however, consequences to using a non-parametric approach. Parametric approaches incorporate an error term assumed to have a particular distribution function. This error term is incorporated to capture any random fluctuation arising due to incorrect model-fit and/or any noise in the data. In the absence of an error term, the implicit assumption is made that the model is the 'true' representation of reality and the data is absent of any noise. If this is not the case then variations in efficiency scores observed from the DEA profiling approach are not representative of actual differences in efficiency but rather random error that is not being properly allowed for.

When interpreting the efficiency scores of surgeons in this study, it is assumed that there is some random fluctuation unaccounted for. This is expressed as one of the justifications for focusing on surgeons that are observed to be highly inefficient. Surgeons with low levels of efficiency are recognised as potentially being misclassified due to random fluctuation present in the DEA approach. Nevertheless, further research into 'stochastic' DEA models is required to determine if superior results can be obtained by the incorporation of a stochastic error term.

8.2.4 Data

Characteristics of the data used in this study affecting the interpretation of surgeons' efficiency scores are (1) the profiling process measures exclusively surgeons' efficiency at treating Medscheme patients, (2) only surgeons treating Medscheme patients form part of the profiling process, and (3) the profiling process only determines the efficiency with which in-hospital services are utilised by surgeons in the treatment of their patients. A limitation of a DEA profiling

approach is that efficiency scores are heavily dependent on the data used. The consequence of this is that the approaches exhibit low levels of robustness to extensions of the process to include the additional relevant data on surgeons' practices. Therefore, surgeons' efficiency scores are vulnerable to significant change if more data were available. The efficiency scores of surgeons determined in this study thus cannot be interpreted as reflecting the surgeons overall technical and price efficiency. This is both regarding the full range of services the surgeons provide as well as relative to all the surgeons practicing within the private healthcare sector in South Africa.

Further research is required to investigate the robustness of the DEA profiling approach followed in this study in reflecting its appropriateness to represent a more holistic view of surgeon performance. In addition, it would be interesting to analyse how the results of this study would change over time. This is one of the methods of illustrating the robustness of this type of DEA profiling approach.

8.2.5 Using classic DEA model

Finally, the profiling approach limited choice to classic DEA models. As discussed in Section 6.6, this is owing to the fact that it would be premature at this early stage of South African DEA profiling research to use more complicated DEA models.

As stated in Section 5, however, a profiling approach based on these models assumes that inefficient surgeons will be able to reduce the amount of all their healthcare service costs proportionally in order to improve efficiency. In practice, surgeons may, however, not be able to proportionally reduce the cost of all healthcare services utilised by the same multiple.

This possibility thus needs further investigation. If found to be the case then 'non-classic' DEA models (such as additive and non-orientated models) need to be applied and the resulting surgeon profiles analysed.

8.3 Final comments

This section illustrates that DEA provides numerous benefits for use by South African funders to profile the efficiency of surgeons providing in-hospital services. These benefits reflect both enhancements in the accuracy of efficiency estimates and improvements in facilitating the presentation of profiles to inefficient surgeons. However, extensive further research is required in order to address some of the major potential limitations in the interpretation of the results. Development of the approach is, therefore, necessary before its extended use to profile all types of physicians. The body of international literature on DEA is, however, extensive providing the necessary material needed for these limitations to be addressed. As such, it can be concluded that DEA represents a significant opportunity for South African healthcare funders to achieve greater understanding regarding physician efficiency and the efficiency improvements achieved as a result.

References

Agrell, Per J, & Bogetoft, Peter. (2001). Should health regulators use DEA. *Coordinacion e Incentivos en Sanidad, Asociasion de Economia de la Salud, Barcelona*, 133-154.

Andes, Steven, Metzger, Lawrence M, Kralewski, John, & Gans, David. (2002). Measuring efficiency of physician practices using data envelopment analysis. *Managed care (Langhorne, Pa.), 11*(11), 48.

Arrow, Kenneth J. (1963). Uncertainty and the welfare economics of medical care. *The American economic review*, 941-973.

Balk, Bert M. (2001). Scale efficiency and productivity change. *Journal of Productivity Analysis, 15*(3), 159-183.

Banker, Rajiv D, Charnes, Abraham, & Cooper, William W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science, 30*(9), 1078-1092.

Bogetoft, Peter, & Otto, Lars. (2010). *Benchmarking with DEA, SFA, and R* (Vol. 157): Springer.

Bornhuetter, R. L., & Ferguson, R. E. (1972). The actuary and IBNR. *Proc. Cas. Act. Soc*, 181-195.

Boyd, Stephen, & Vandenberghe, Lieven. (2009). *Convex optimization*: Cambridge university press.

Centre for Development and Enterprise. (2011). Reforming healthcare in South Africa - What role for the private sector? *Research no. 18*.

Charnes, Abraham, & Cooper, William W. (1962). Programming with linear fractional functionals. *Naval Research logistics quarterly, 9*(3-4), 181-186.

Charnes, Abraham, Cooper, William W, & Rhodes, Edwardo. (1978). Measuring the efficiency of decision making units. *European journal of operational research, 2*(6), 429-444.

Charvet, Heidi. (2009). The Problems with Physician Profiling: What Have We Learned? *Quill and Scope, 2*.

Chilingerian, Jon A. (1995). Evaluating physician efficiency in hospitals: A multivariate analysis of best practices. *European journal of operational research, 80*(3), 548-574.

Chilingerian, Jon A, & Sherman, H David. (1990). Managing physician efficiency and effectiveness in providing hospital services. *Health Services Management Research, 3*(1), 3-15.

Chilingerian, Jon A, & Sherman, H David. (1997). DEA and primary care physician report cards: Deriving preferred practice cones from managed care service concepts and operating strategies. *Annals of operations Research, 73*, 35-66.

Christiansen, Cindy L, & Morris, Carl N. (1997). Improving the statistical approach to health care provider profiling. *Annals of Internal Medicine, 127*(8_Part_2), 764-768.

Coelli, Timothy J, Rao, Dodla Sai Prasada, O'Donnell, Christopher J, & Battese, George Edward. (2005). *An introduction to efficiency and productivity analysis*: Springer.

Competition Commission. (2013). *Terms of reference for market inquiry into the private healthcare sector*. Pretoria: Government Gazette.

Cook, Wade D, & Zhu, Joe. (2006). *Modeling performance measurement: applications and implementation issues in DEA* (Vol. 566): Springer.

Cooper, William W, Seiford, Lawrence M, & Zhu, Joe. (2011a). Data envelopment analysis: history, models, and interpretations *Handbook on data envelopment analysis* (pp. 1-39): Springer.

Cooper, William W, Seiford, Lawrence M, & Zhu, Joe. (2011b). *Handbook on data envelopment analysis*: Springer Science+ Business Media.

Council for Medical Schemes. (2010). Requirements of medical scheme administrators. Pretoria.

Council for Medical Schemes. (2013). Annual Report 2012-2013. Pretoria.

Council for Medical Schemes. (2014). Quarterly report for the period ending 30 September 2013. Pretoria.

Cumming, Robert B, Knutson, David, Cameron, Brian A, & Derrick, Brian. (2002). A comparative analysis of claims-based methods of health risk assessment for commercial populations. *Final report to the Society of Actuaries*.

Debreu, Gerard. (1951). The coefficient of resource utilization. *Econometrica: Journal of the Econometric Society*, 273-292.

Deloitte. (2013). 2014 Global health care outlook. United Kingdom: Deloitte Touche Tohmatsu Life Sciences and Health Care group.

du Preez, Laura. (2013, 7 September). Medical schemes' nasty symptoms hurt members, *Weekend Argus*.

Eckermann, Simon, & Coelli, Tim. (2008). Including quality attributes in a model of health care efficiency: A net benefit approach. *Centre for Efficiency and Productivity Analysis, Working paper no. WP03/2008 University of Queensland*.

Econex. (2013). The South African Private Healthcare Sector: Role and Contribution to the Economy. Stellenbosch.

Edmunds, Margaret. (1997). *Managing managed care: quality improvements in behavioral health*. Institute of Medicine (US). Committee on Quality Assurance Accreditation Guidelines for Managed Behavioral Health Care: National Academies.

Eijkenaar, Frank, & van Vliet, René CJA. (2013). Profiling individual physicians using administrative data from a single insurer: Variance components, reliability, and implications for performance improvement efforts. *Medical care*, *51*(8), 731-739.

Enthoven, Alain C. (2002). *Health plan: the practical solution to the soaring cost of medical care*: Beard Books.

Färe, R, Grosskopf, S, & Lovell, CAK. (1985). The Measurement of Efficiency of Production. *Studies in Productivity Analysis, Kluwer-Nijhoff Publishing, Dordrecht*.

Färe, R, Grosskopf, S, & Lovell, CAK. (1994). *Production frontiers*: Cambridge University Press.

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

Ferver, Kari, Burton, Bryan, & Jesilow, Paul. (2009). The use of claims data in healthcare research. *Open Publ Health J, 2*, 11-24.

Fetter, Robert B, & Freeman, Jean L. (1986). Diagnosis related groups: product line management within hospitals. *Academy of Management Review, 11*(1), 41-54.

Fetter, Robert B, Shin, Youngsoo, Freeman, Jean L, Averill, Richard F, & Thompson, John D. (1980). Case mix definition by diagnosis-related groups. *Medical care*, i-53.

Findlay, Steven. (1993). Profiling systems aim to eliminate second-guessing of doctors. *Business and health, 11*(5), 58.

Garnick, Deborah W, Fowles, Jinnet, Lawthers, Ann G, Weiner, Jonathan P, Parente, Steve T, & Palmer, R Heather. (1994). Focus on quality: profiling physicians' practice patterns. *The Journal of Ambulatory Care Management, 17*(3), 44-75.

Gillis, Kurt D, & Hixson, Jesse S. (1991). Efficacy of statistical outlier analysis for monitoring quality of care. *Journal of Business & Economic Statistics, 9*(3), 241-252.

Golany, Boaz, & Roll, Yaakov. (1989). An application procedure for DEA. *Omega, 17*(3), 237-250.

Hodge, J., Fiandeiro, F., Lynch, S., & Mohamed, R. (2012). Healthcare market backgroung paper: Genesis Analytics.

Hollingsworth, Bruce. (2008). The measurement of efficiency and productivity of health care delivery. *Health economics, 17*(10), 1107-1128.

InterStudy. (1989). *The InterStudy Edge*.

Jencks, Stephen F, & Dobson, Allen. (1987). Refining case-mix adjustment. The research evidence. *The New England journal of medicine, 317*(11), 679-686.

Jenkins, Larry, & Anderson, Murray. (2003). A multivariate statistical approach to reducing the number of variables in data envelopment analysis. *European journal of operational research, 147*(1), 51-61.

Kassirer, Jerome P. (1994). The use and abuse of practice profiles. *The New England journal of medicine, 330*(9), 634.

Kittelsen, S. (1993). Stepwise DEA. *Choosing variables for measuring technical efficiency in Norwegian electricity distribution. Memorandum, 6*, 93.

Koopmans, Tjalling C. (1951). Analysis of production as an efficient combination of activities. *Activity analysis of production and allocation, 13*, 33-37.

Kuosmanen, Timo. (2001a). DEA with efficiency classification preserving conditional convexity. *European journal of operational research, 132*(2), 326-342.

Kuosmanen, Timo. (2001b). *The role of production assumptions in Data Envelopment Analysis.* (Doctoral Dissertations), Helsinki School of Economics and Business Administration. (Series A, 188)

Lasker, Roz Diane, Shapiro, David W, & Tucker, Anthony M. (1992). Realizing the potential of practice pattern profiling. *Inquiry*, 287-297.

Lodh, Mita, Raleigh, Michelle L., Uccello, Cori E., & Winkelman, Ross A. (2010). Risk assessment and risk adjustment: American Academy of Actuaries.

Loudon, Irvine. (2008). The principle of referral: the gatekeeping role of the GP. *British Journal of General Practice, 58*(547), 128-130.

Lovell, CA Knox. (1993). Production frontiers and productive efficiency. *The measurement of productive efficiency: techniques and applications*, 3-67.

Lovell, CA Knox. (2006). Frontier analysis in healthcare. *International journal of healthcare technology and management, 7*(1), 5-14.

Luft, Harold S, & Hunt, Sandra S. (1986). Evaluating individual hospital quality through outcome statistics. *JAMA: the journal of the American Medical Association, 255*(20), 2780-2784.

Mataconis, Doug. (2011). Health Care Costs And The Third-Party Payer Problem. Retrieved 15 July, 2014, from http://www.outsidethebeltway.com/health-care-costs-and-the-third-party-payer-problem/

Matsoso, Malebona Precious, & Fryatt, R. (2013). National Health Insurance: the first 18 months. *SAMJ: South African Medical Journal, 103*(3), 154-155.

McIntyre, Diane. (2010). *Private sector involvement in funding and providing health services in South Africa: implications for equity and access to health care*: Regional network for equity in health in east and southern Africa (EQUINET).

McLeod, H. (2005). Mutuality and solidarity in healthcare in South Africa. *South African Actuarial Journal, 5*(1), 135-167.

McLeod, H., & Grobler, P. (2010). Risk equalisation and voluntary health insurance: The South Africa experience. *Health policy, 98*(1), 27-38.

McLeod, H., Mubangizi, D. B., Rothberg, A., & Fish, T. (2003). *The Impact of Prescribed Minimum Benefits on the Affordability of Contributions*: Centre for Actuarial Research, University of Cape Town.

McLeod, H., & Ramjee, S. (2007). Medical schemes: pooling of resources and purchasing of health care. *South African health review*, 47-70.

McNeil, Barbara J, Pedersen, Sarah H, & Gatsonis, Constantine. (1992). Current issues in profiling quality of care. *Inquiry*, 298-307.

Mills, Anne, Ataguba, John E, Akazili, James, Borghi, Jo, Garshong, Bertha, Makawia, Suzan, . . . Meheus, Filip. (2012). Equity in financing and use of health care in Ghana, South Africa, and Tanzania: implications for paths to universal coverage. *The Lancet, 380*(9837), 126-133.

Mushkin, Selma J. (1958). Toward a definition of health economics. *Public health reports, 73*(9), 785.

Normand, Sharon-Lise T, Glickman, Mark E, & Gatsonis, Constantine A. (1997). Statistical methods for profiling providers of medical care: issues and applications. *Journal of the American Statistical Association, 92*(439), 803-814.

Ozcan, Yasar A. (1998). Physician benchmarking: measuring variation in practice behavior in treatment of otitis media. *Health Care Management Science, 1*(1), 5-17.

Ozcan, Yasar A. (2008). *Health care benchmarking and performance evaluation: an assessment using Data Envelopment Analysis (DEA)*: Springer Berlin.

Ozcan, Yasar A, Jiang, HJ, & Pai, CW. (2000). Do primary care physicians or specialists provide more efficient care? *Health services management research: an official journal of the Association of University Programs in Health Administration/HSMC, AUPHA, 13*(2), 90.

Parente, Stephen T. (2002). *Fixed and random effects econometric modeling for provider profiling*. Presentation. Department of Healthcare Management. University of Minnesota.

Pope, Gregory C, & Kautter, John. (2007). Profiling efficiency and quality of physician organizations in Medicare. *Health Care Financing Review, 29*(1), 31.

Ramjee, S, Kooverjee, A, & Dreyer, KA. (2013). The construction of a price index for contributions to South African open medical schemes. *South African Actuarial Journal, 13*(1), 1-20.

Rosenthal, Meredith B, Landon, Bruce E, Normand, Sharon-Lise T, Frank, Richard G, & Epstein, Arnold M. (2006). Pay for performance in commercial HMOs. *New England Journal of Medicine, 355*(18), 1895-1902.

Salem-Schatz, Susanne, Moore, Gordon, Rucker, Malcolm, & Pearson, Steven D. (1994). The case for case-mix adjustment in practice profiling: when good apples look bad. *Jama, 272*(11), 871-874.

Shapiro, DW, Lasker, RD, Bindman, AB, & Lee, PR. (1993). Containing costs while improving quality of care: the role of profiling and practice guidelines. *Annual Review of Public Health, 14*(1), 219-241.

Sherman, H David, & Zhu, Joe. (2006). *Service productivity management: Improving service performance using data envelopment analysis (DEA)*: Springer.

Smith, David W. (1994). Evaluating risk adjustment by partitioning variation in hospital mortality rates. *Statistics in medicine, 13*(10), 1001-1013.

Taylor, Bettina, Taylor, A, Burns, D, Rust, JD, & Grobler, P. (2007). prescribed minimum benefits-quagmire or foundation for social health reform?: original article. *South African Medical Journal, 97*(6), 446-450.

Thanassoulis, Emmanuel. (2001). *Introduction to the theory and application of data envelopment analysis: a foundation text with integrated software*: Springer.

Thomas, J William, Grazier, Kyle L, & Ward, Kathleen. (2004). Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures. *Health services research, 39*(4p1), 985-1004.

Thompson, John D, Fetter, Robert B, & Mross, Charles D. (1975). Case mix and resource use. *Inquiry*, 300-312.

Twiss, A. Kirk, Yamamoto, Dale H., Pyenson, Bruce S., Allen, Jeffrey G., & Fredericks, Melissa A. (1998). The Actuary's Role in Managed Care- The Actuary's Role in Managed Care Working Group. *North American Actuarial Journal, 2*(3), 128-136.

van den Heever, Alex. (2012). Review of competition in the South African health system. South Africa: University of Witwaterstrand.

Wagner, Janet M, & Shimshak, Daniel G. (2007). Stepwise selection of variables in data envelopment analysis: Procedures and managerial perspectives. *European journal of operational research, 180*(1), 57-67.

Wagner, Janet M, Shimshak, Daniel G, & Novak, Michael A. (2003). Advances in physician profiling: the use of DEA. *Socio-Economic Planning Sciences, 37*(2), 141-163.

WHO. (2010). Classifying health workers. Geneva: World Health Organisation.

Winkelman, Ross, & Mehmud, Syed. (2007). A comparative analysis of claims-based tools for health risk assessment. *Society of Actuaries*, 1-70.

Zweifel, Peter, Breyer, Friedrich, & Kifmann, Mathias. (2009). *Health economics*: Springer.

Appendix A

Claim_lines_2012

STG_HAS_KEY	patient_id	SCHEME_CODE	PLAN_KEY	year	treatment_date	ms_std_code	code_description	sub_group	group	account_amount	tariff_amount	practice_number	practice_type	ref_pract_number
337562070	4486562	SAU	9	2012	16MAR2012	1415	Combined procedur	Varicose vein p	Varicose Vein	1231.46	824.92	196967	10	4208714
337562070	4486562	SAU	9	2012	16MAR2012	61731				120	120	8700672	87	0226114
337562070	4486562	SAU	9	2012	16MAR2012	97647	Age Modifier	Inpatient stay:	Inpatient Stay	238.52	238.52	5808650	58	
337562070	4486562	SAU	9	2012	16MAR2012	97631	Base Rate,Same Day	Day patient: S	Daypatient sta	644	644	5808650	58	
337562070	4486562	SAU	9	2012	16MAR2012	58273				356.39	356.39	5808650	58	
337562070	4486562	SAU	9	2012	16MAR2012	P2861328	MEDI MEDIVEN COM	Other prosthes	Supplies	550	550	8700672	87	0226114
337562070	4486562	SAU	9	2012	16MAR2012	0151	Pre-anaesthetic ass	Visits: Medica	Visits	362.09	242.5	196967	10	4208714
337562070	4486562	SAU	9	2012	16MAR2012	1417	Extensive sub-fascia	Varicose vein p	Varicose Vein	1332.32	1056.16	4208714	42	
337562070	4486562	SAU	9	2012	16MAR2012	1413	Combined procedur	Varicose vein p	Varicose Vein	1502.84	1502.84	4208714	42	
337562070	4486562	SAU	9	2012	16MAR2012	97200	Theatre Time, per 1	Operating roor	Operating Roo	8100	8100	5808650	58	
337562092	5552745	SAU	9	2012	02MAR2012	07017	Medicine dispensec	TTO/acute dru	Drug Administ	148.2	148.2	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	1807	Add to open proced	Laparoscopy w	Laparoscopy w	836.2	422.5	4208153	42	
337562092	5552745	SAU	9	2012	03MAR2012	P12301	Physical modalities:	Nebulisations	Respiratory Th	114.6	114.7	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	97817	Surgical Category K	Operating roor	Operating Roo	41828	41828	5808111	58	
337562092	5552745	SAU	9	2012	03MAR2012	P2931100	X-ray barium swallo	Radiology - Fla	Radiology: Fla	591.3	591.2	3802590	38	4208153
337562092	5552745	SAU	9	2012	02MAR2012	1563	With anti-reflux pro	Hiatus hernia p	Hiatus hernia s	3272.18	1826.5	1012339	10	4208153
337562092	5552745	SAU	9	2012	03MAR2012	P12301	Physical modalities:	Nebulisations	Respiratory Th	114.6	114.7	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	1563	With anti-reflux pro	Hiatus hernia p	Hiatus hernia s	5574.2	2816.3	4208153	42	
337562092	5552745	SAU	9	2012	02MAR2012	1807	Add to open proced	Laparoscopy w	Laparoscopy w	1017.97	633.75	4208153	42	NS
337562092	5552745	SAU	9	2012	02MAR2012	1563	With anti-reflux pro	Hiatus hernia p	Hiatus hernia s	6785.93	4224.45	4208153	42	NS
337562092	5552745	SAU	9	2012	02MAR2012	97106	TTO (Revenue code	TTO/acute dru	Drug Administ	66.17	66.17	5808111	58	
337562092	5552745	SAU	9	2012	02MAR2012	97823	Surgical > One Day S	Inpatient stay:	Inpatient Stay	2609	2609	5808111	58	
337562092	5552745	SAU	9	2012	03MAR2012	P2900090	Consumables used i	Radiology: Mis	Radiology: Fla	572.51	572.51	3802590	38	4208153
337562092	5552745	SAU	9	2012	02MAR2012	0151	Pre-anaesthetic ass	Visits: Medica	Visits	434.5	242.5	1012339	10	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P12301	Physical modalities:	Nebulisations	Respiratory Th	114.6	114.7	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	2802	Procedures for pain	Alcohol/anesth	Injections/bloc	464.62	234.7	4208153	42	
337562092	5552745	SAU	9	2012	02MAR2012	P12321	Physical modalities:	Nebulisations	Respiratory Th	35.6	35.6	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P12319	Physical modalities:	Nebulisations	Respiratory Th	35.6	35.6	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P12901	Visiting codes: Trea	Physical treatn	Rehabilitation	71.2	71.2	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P12321	Physical modalities:	Nebulisations	Respiratory Th	71.2	71.2	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P12701	Evaluation: Simple	Physical treatn	Rehabilitation	106.8	106.8	7218087	72	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P2930100	X-ray of the chest, s	Radiology - Fla	Radiology: Fla	272.4	272.3	3802590	38	4208153
337562092	5552745	SAU	9	2012	02MAR2012	P2900130	X-ray with mobile u	Radiology - Fla	Radiology: Fla	170.2	170.2	3802590	38	4208153
337562092	5552745	SAU	9	2012	03MAR2012	P12307	Physical modalities:	Physical treatn	Rehabilitation	35.6	35.6	7218087	72	4208153
337562092	5552745	SAU	9	2012	03MAR2012	P12321	Physical modalities:	Nebulisations	Respiratory Th	35.6	35.6	7218087	72	4208153
337562092	5552745	SAU	9	2012	03MAR2012	P12901	Visiting codes: Trea	Physical treatn	Rehabilitation	71.2	71.2	7218087	72	4208153
337562092	5552745	SAU	9	2012	03MAR2012	P12307	Physical modalities:	Physical treatn	Rehabilitation	35.6	35.6	7218087	72	4208153
337562092	5552745	SAU	9	2012	03MAR2012	P12321	Physical modalities:	Nebulisations	Respiratory Th	35.6	35.6	7218087	72	4208153

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	Q	Z	AA	AD
1	STG_HAS_KEY	Patient_ID	SCHEME_CODE	PLAN_KEY	Year	Month	DATE_OF_BIRTH	AGE	PATIENT	PROV_PF	ADMISSION_CATEGORY	mdc	base_drg	drg	SURGERY/PAEDIATRIC SURGERY	TOTAL HOSPITAL AMOUNT	ANAESTHETISTS	TOTAL RADIOLOGY AMOUNT
2	338537908	7368078	GMT	679	2012	3	1960/09/21	52	F	4207777	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	727.6	2618.95	0	0
3	338537922	2340886	GMT	679	2012	2	1959/10/04	53	F	4204883	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	692.5	2624.1	0	0
4	338537962	8569857	GMT	681	2012	11	1952/11/22	60	F	4207211	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	462.6	1717.72	0	0
5	338537973	2356036	GMT	679	2012	2	2000/07/06	12	M	0162647	Debridement of wound/slou	MDC 21 In	Procedures	Procedur	477.7	7121.31	905.54	0
6	338538006	7695922	GMT	679	2012	3	1962/01/26	50	M	4204883	Laparotomy: Liver/stomach	MDC 06 D	Other Diges	Other Dig	1415.1	10577.79	950.5	0
7	338538029	9334563	GMT	679	2012	7	1956/12/22	56	F	0109401	Lower GI Endoscopy	MDC 16 D	Red Blood	Red Bloo	1574.6	3436.94	0	0
8	338538045	9181372	GMT	681	2012	11	1968/10/10	44	M	0033227	Debridement of wound/slou	MDC 05 D	Other Circul	Other Circ	2357.69	38541.09	1257.46	1094.96
9	338538053	9916468	GMT	679	2012	10	1930/09/10	82	F	0033227	Haematemesis, melaena, h	MDC 06 D	Other Gastr	Other Gas	939.8	67765.52	0	10431.82
10	338538056	9916468	GMT	679	2012	9	1930/09/10	82	F	0316032	Acute Renal Failure	MDC 11 D	Acute Rena	Acute Rer	724.3	15567.27	0	0
11	338538063	2371123	GMT	679	2012	10	1971/04/15	41	F	4207211	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	462.6	1659.96	0	0
12	338538101	6332190	GMT	680	2012	4	1995/12/07	17	F	0312983	Appendectomy	MDC 06 D	Appendice	Appendic	9307.12	45007.21	6211.51	750.6
13	338538164	2397791	GMT	679	2012	4	1997/12/23	15	F	0207330	Ovarian/Tube/Menopause I	MDC 13 D	Menstrual a	Menstrual	1632.7	10030.93	0	676.2
14	338538180	8570491	GMT	681	2012	3	1963/08/01	49	F	4203372	Cholecystectomy	MDC 07 D	Laparosco	Laparosc	709.1	3090.84	0	637.9
15	338538203	2403269	GMT	680	2012	10	1955/05/25	57	M	0018678	Cholecystectomy	MDC 06 D	Other Herni	Other Her	9372.54	33859.16	4647.8	0
16	338538208	2403847	GMT	681	2012	5	1968/05/25	44	F	0263133	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	462.6	819.73	0	0
17	337562070	4486562	SAU	9	2012	3	1945/03/09	67	F	4208714	Varicose Vein Procedures	MDC 05 D	Vein Ligatic	Vein Liga	2836.16	9339.91	1594.55	0
18	337562092	5552745	SAU	9	2012	3	1969/06/02	43	M	4208153	Hiatus hernia surgery	MDC 06 D	Fundoplast	Fundopla	8269.52	44504.17	3707.68	1607.41
19	337562095	5551785	SAU	9	2012	4	1971/09/11	41	F	4208153	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	3562.96	2921	2850.68	0
20	337562114	2138835	SAU	9	2012	9	1955/07/21	57	F	4207599	Skin Lesion Biopsy/Excisior	MDC 09 D	Skin, Subcu	Skin, Sub	1720.6	5465	1456.49	0
21	337562167	5550860	SAU	9	2012	2	1973/06/22	39	F	4207696	Gastritis/Dyspepsia/Heartbi	MDC 06 D	Other Gastr	Other Gas	2220.63	16117.36	1998.9	1353.8
22	337562177	1747982	SAU	9	2012	5	1932/03/18	80	F	4207319	Haematemesis, melaena, h	MDC 06 D	Other Gastr	Other Gas	3397.35	29282.5	0	0
23	337562180	384822	SAU	9	2012	1	1948/01/06	64	F	4203933	Diverticulitis/Diverticulosis	MDC 06 D	Diverticular	Diverticul	2161	19245	0	7699.04
24	337562181	384822	SAU	9	2012	5	1948/01/06	64	F	4203933	Diverticulitis/Diverticulosis	MDC 06 D	Diverticular	Diverticul	2138.8	13678.29	0	6812.66
25	337562213	4319431	SAU	9	2012	7	1961/03/19	51	F	4208811	Inguinal Hernia Repair	MDC 06 D	Ventral Her	Ventral He	3815.15	28529.28	2141.31	1452.3
26	337562229	5080449	SAU	9	2012	4	1990/11/21	22	F	0069884	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	1062.9	2989.09	0	1219.4
27	337562291	759711	SAU	9	2012	10	1930/05/08	82	F	0109924	Upper GI Endoscopy	MDC 06 D	Colonosco	Colonosc	916.6	4172.61	718.29	3415.5
28	337562324	2838411	SAU	9	2012	4	1962/02/28	50	F	4206150	Lower GI Endoscopy	MDC 06 D	Colonosco	Colonosc	1303.7	4284.19	1279.92	0
29	337562333	2816301	SAU	9	2012	1	1971/06/14	41	M	0177768	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	458.7	2819.23	0	0
30	337562361	5017809	SAU	9	2012	2	1970/02/24	42	F	4208889	Lower GI Endoscopy	MDC 06 D	Colonosco	Colonosc	1765.6	6787.54	714.9	0
31	337562374	2784805	SAU	9	2012	8	1995/03/30	17	M	0465712	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	1325.54	3669.36	714.94	854.16
32	337562389	2189731	SAU	9	2012	10	1957/04/29	55	F	4204336	Breast Lesion Biopsy/Excisi	MDC 09 D	Minor Proc	Minor Pro	880	10230	2637.54	0
33	337562421	7890536	SAU	9	2012	11	2009/07/06	3	M	0195065	Circumcision	MDC 12 D	Circumcisic	Circumcis	620.6	5912.79	1933	0
34	337562458	756029	SAU	9	2012	7	2001/04/18	11	M	0052981	Intestinal infectious disease	MDC 06 D	Oesophagi	Oesopha	470.93	3849.77	0	0
35	337562487	1786324	SAU	9	2012	7	1971/07/17	41	M	0361429	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	1156.13	5037.08	0	7697.42
36	337562546	5007309	SAU	9	2012	7	2004/11/12	8	M	0277940	Circumcision	MDC 12 D	Circumcisic	Circumcis	822	4936.8	0	0
37	337562562	823778	SAU	9	2012	2	1956/08/01	56	M	4205944	Incision/Drainage Skin Absc	MDC 09 D	Skin, Subcu	Skin, Sub	1450.2	7547.81	1310.96	0
38	337562596	5709299	SAU	9	2012	11	1994/07/07	18	M	4208811	Appendectomy	MDC 06 D	Appendice	Appendic	2079.2	14780.74	940.47	0
39	337562605	5761294	SAU	9	2012	2	1991/06/07	21	F	4200551	Breast Lesion Biopsy/Excisi	MDC 09 D	Skin, Subcu	Skin, Sub	885.4	9992	2716.76	0
40	337562606	5761294	SAU	9	2012	2	1991/06/07	21	F	4200551	Debridement of wound/slou	MDC 09 D	Skin, Subcu	Skin, Sub	1399.86	38296	5327.05	443.5
41	337562634	3700378	SAU	9	2012	5	1959/03/15	53	M	0177768	Peptic Ulcer Disease	MDC 06 D	Other Gastr	Other Gas	1156	9473.75	0	0
42	337562636	3614434	SAU	9	2012	1	1965/01/11	47	M	4203178	Hernias	MDC 06 D	Other Gastr	Other Gas	1631	11591.06	0	802
43	337562637	3614434	SAU	9	2012	4	1965/01/11	47	M	4203178	Upper GI Endoscopy	MDC 06 D	Other Gastr	Other Gas	928.1	3323.61	0	0
44	337562648	2443037	SAU	9	2012	10	1967/09/29	45	F	0307653	Debridement of wound/slou	MDC 09 D	Skin, Subcu	Skin, Sub	4216.09	8685.99	1579.36	0
45	337562700	2335548	SAU	9	2012	9	1971/04/09	41	M	4208943	Injury/Trauma (Medical Stay	MDC 09 D	Trauma to t	Trauma to	471.1	1357.48	0	0
46	337562702	2335686	SAU	9	2012	6	2003/09/25	9	M	4208943	Circumcision	MDC 12 D	Circumcisic	Circumcis	839.95	5574.53	0	0
47	337562774	9542107	SAU	9	2012	5	1971/10/13	41	F	0105260	Injections/blocks for pain co	MDC 09 D	Other Skin,	Other Skir	1263.2	11855.78	891.66	0
48	337562790	6074326	SAU	9	2012	1	1972/04/11	40	F	0092436	Incision/Drainage Skin Absc	MDC 14 P	Antenatal a	Antenatal	1407.88	17508.22	1716.86	0

Appendix B

DISCIPLINE_DESCRIPTION
AMBULANCE SERVICES
ANAESTHETISTS
DERMATOLOGY
GENERAL MEDICAL PRACTICE
OBSTETRICS AND GYNAECOLOGY
PULMONOLOGY
INDEPENDENT PRACTICE SPECIALIST MEDICINE
GASTROENTEROLOGY
NEUROLOGY
CARDIOLOGY
PSYCHIATRY
MEDICAL ONCOLOGY
NEUROSURGERY
OPTHALMOLOGY
HAEMATOLOGY
ORTHOPAEDICS
OTORHINOLARYNGOLOGY
RHEUMATOLOGY
PAEDIATRICS
PAED. CARDIOLOGY
PLASTIC AND RECONSTRUCTIVE SURGERY
RADIOGRAPHY
SURGERY/PAEDIATRIC SURGERY
THORACIC SURGERY
UROLOGY
PATHOLOGY
GENERAL DENTAL PRACTICE
PRIVATE HOSPITALS
PHARMACIES
MAXILLO-FACIAL AND ORAL SURGERY
ORTHODONTICS
OCCUPATIONAL THERAPY
OPTOMETRISTS
PHYSIOTHERAPISTS
ORTHOPTISTS
SPEECH THERAPY / AUDIOLOGY
PSYCHOLOGISTS
ORTHOTISTS & PROSTHETISTS