ALLOCATING HUMAN RESOURCES FOR HEALTHCARE IN THE MEDICAL SCHEMES INDUSTRY
PER DISCIPLINE

Preparing for National Health Insurance & Amended Medical Schemes Act

Research and Monitoring

Working Paper Series: Policy Brief No. 3 of 2019

Senior Policy Analyst:
Phakamile Nkomo.
EXECUTIVE SUMMARY

INTRODUCTION

Garret Harding hypothesised the solution for Thomas Malthus's population problem when he wrote; “the population problem has no technical solution; it requires a fundamental extension in morality” (Hardin, 1968). In that same spirit, a “benevolent social planner” should provide a coordinated response to market inefficiencies. This gives credence to the National Department of Health’s drive for universal health coverage. In this instance, the “Tragedy of the Commons” is that so much human capital in healthcare is invested at diminishing marginal returns to healthcare outcomes. The tragedy of the common healthcare market is that, 80% of the population is impoverished through not accessing the heavily endowed private healthcare sector. This policy brief finds the diminishing returns to scale in geographic healthcare markets and proposes a method to divest from residual efficiency outlays with the objective of reducing inequalities in accessing private healthcare services. Recommendations are discussed with the intention to complement other policy objectives, in support of implementing a set of mixed policy instruments.

PROBLEM STATEMENT

There are inequalities in the access of healthcare providers across geographic markets in the private healthcare industry. This has an impact on primary healthcare re-engineering drives for National Health Insurance (NHI), implementation of appropriate integrated care networks, and guaranteed access to healthcare benefit for all. A solution to the potential coordination problem, amongst other, is to identify markets where resources are underproductive, and redirect them to geographic markets where increasing returns to scale are present.

RECOMMENDATIONS

1. Redistribution of Healthcare Providers

1.1 The issue of resolving unequal access to healthcare:

i. The issue should be resolved through regulatory coordinating actions. The regular operation of the market cannot resolve the unequal distribution of healthcare providers.

ii. In resolving issues on the technical allocation of healthcare resources, Data Envelopment Analysis (DEA) models should be used in evaluating the economic interaction of market agents and resulting healthcare market prices. Relevant market identification models measure market concentration, but the market forces nor market agent interactions behind high utilisation costs or efficiencies in healthcare production. These are merely assumed from the concentration indices. An economic evaluation model based on DEA
analysis has been used to decompose technical efficiency into its pure efficiency and residual efficiency (technical efficiency components).

iii. DEA models should identify areas where there are technical inefficiencies in healthcare delivery and redirect these resources where there are positive returns to scale.

iv. The Virtual Risk Pooling Fund could be used to incentivise needed healthcare resources to relocate to underserved areas. This could also act as an indirect income cross-subsidization mechanism for deprived communities.

v. Re-imbursement contracts for healthcare providers who are subsidized for relocating, should be based on a good knowledge of the providers true efficiency. Providers could pretend to be less efficient than what they are, and benefit from fixed price contract arrangements like highly priced capitation agreements.

1.2 The issue of non-existing markets creating market failure in certain geographic markets:

i. The Health Market Inquiry (HMI) has been receiving recommendations on how to resolve the issues of arm in the private healthcare market.

ii. The lack of market access, or lack of an existing market, in certain areas is a market failure. This is because the normal operation of competition has not allocated resources where there is demand, for lack of incentives.

iii. The analysis findings and methods in this document could assist in making HMI recommendations where relevant market methodology has not assisted.

1.3 The issue of impediments to efficient operation of competition in geographic markets, as an “issue of harm”:

i. In some instances, merger evaluations in healthcare markets have under-estimated the anticompetitive effects of price increases. This very particular to healthcare market.

ii. This is because market price mechanisms have been very difficult assess with market delineation methods.
iii. Understanding the nature of existing technical efficiencies in a geographical healthcare market, provides a useful inference on a driving force average costs resulting from healthcare facility production functions.

iv. The DEA approach to economic evaluation may help in monitoring market behaviour.

1.4 The issue of where to intervene and how to intervene through provider networks and integrated healthcare delivery models:
   i. Integrated healthcare networks and networks in benefit designs should prioritise underserved areas
   
   ii. Integrated healthcare networks and networks in benefit design should not be allowed in geographic markets with residual efficiencies (technical inefficiency). If Efficiency Discount Option (EDO) and low-cost integrated care networks are registered in these areas, policy objectives will not be met.

2 Considerations for a Mixed Bag of Policy Interventions
2.1 The issue on whether to accredit healthcare facilities based on geographic markets or a certificate of needs:
   i. It may be easier to delineate and standardise product markets for primary and secondary healthcare than it may be for tertiary and quaternary services. As the latter may prove to be more complex product offerings.
   
   ii. It may be difficult to delineate markets for certain complex healthcare interventions by geographic, regional or district level. Market forces will respond to factors that are external to markets that are delineated at sub-national or provincial level. Therefore, the policy intervention or regulated equilibrium will not hold.

2.2 The issue on tariff regulation or bargaining councils/chambers:
   i. Standardising benefit designs could go a long to making contracting more transparent in bargaining chambers. Consumer groups could ensure that consumers are protected as they are passive participants in the primary healthcare market
   
   ii. Standardised benefit options could remove the complexity of delineating geographic healthcare markets
## CONTENT

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>PROBLEM STATEMENT</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>BACKGROUND</td>
<td>1</td>
</tr>
<tr>
<td>1.3</td>
<td>PURPOSE</td>
<td>3</td>
</tr>
<tr>
<td>1.4</td>
<td>OBJECTIVES</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>METHODS</td>
<td>5</td>
</tr>
<tr>
<td>2.1</td>
<td>LITERATURE REVIEW</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>EMPIRICAL METHODS</td>
<td>10</td>
</tr>
<tr>
<td>2.3</td>
<td>LIMITATIONS</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>RESULTS</td>
<td>12</td>
</tr>
<tr>
<td>3.1</td>
<td>OVERALL MARKET STRUCTURE: DISTRIBUTION OF PROVIDERS &amp; BENEFICIARIES</td>
<td>12</td>
</tr>
<tr>
<td>3.2</td>
<td>PRIMARY CARE: GENERAL PRACTITIONANS</td>
<td>14</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Efficiency Analysis: GP</td>
<td>14</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Equality, pata optimality &amp; trade-offs: GP</td>
<td>17</td>
</tr>
<tr>
<td>3.3</td>
<td>MATERNITY BENEFITS: GYNAECOLOGISTS</td>
<td>18</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Efficiency analysis: Gynaecologists</td>
<td>18</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Equity, pata optimality &amp; trade-offs: Gynaecologists</td>
<td>20</td>
</tr>
<tr>
<td>3.4</td>
<td>CHILD HEALTHCARE: PEDIATRICIANS</td>
<td>22</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Efficiency analysis: Paediatricians</td>
<td>22</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Equity, pata optimality &amp; trade-offs: Paediatricists</td>
<td>25</td>
</tr>
<tr>
<td>3.5</td>
<td>MENTAL HEALTHCARE: PSYCHIATRISTS</td>
<td>26</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Efficiency analysis: Psychiatrists</td>
<td>26</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Equity, pata optimality &amp; trade-offs: Psychiatrists</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>FINDINGS &amp; POLICY DISCUSSION</td>
<td>31</td>
</tr>
<tr>
<td>4.1</td>
<td>EMERGING POLICY THEMES FROM LITERATURE REVIEW</td>
<td>31</td>
</tr>
<tr>
<td>4.1.1</td>
<td>The problem with geographic market delineation at tertiary healthcare level</td>
<td>31</td>
</tr>
<tr>
<td>4.1.2</td>
<td>The new medical arms race hypothesis</td>
<td>31</td>
</tr>
<tr>
<td>4.1.3</td>
<td>Co-dependence between efficiency and equality</td>
<td>32</td>
</tr>
<tr>
<td>4.2</td>
<td>SIGNIFICANT FINDINGS FROM EMPIRICAL ANALYSIS</td>
<td>33</td>
</tr>
<tr>
<td>4.3</td>
<td>POLICY DISCUSSION</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>RECOMMENDATIONS</td>
<td>36</td>
</tr>
<tr>
<td>5.1</td>
<td>REDISTRIBUTION OF HEALTHCARE PROVIDERS</td>
<td>36</td>
</tr>
<tr>
<td>5.2</td>
<td>CONSIDERATIONS FOR A MIXED BAG OF POLICY INTERVENTIONS</td>
<td>38</td>
</tr>
<tr>
<td>6</td>
<td>REFERENCES</td>
<td>40</td>
</tr>
<tr>
<td>7</td>
<td>TECHNICAL NOTE</td>
<td>43</td>
</tr>
<tr>
<td>7.1</td>
<td>GINI CO-efficient</td>
<td>43</td>
</tr>
<tr>
<td>7.2</td>
<td>HOOVER INDEX</td>
<td>43</td>
</tr>
<tr>
<td>7.3</td>
<td>DECILE RATIOS</td>
<td>44</td>
</tr>
</tbody>
</table>
LIST OF TABLES

Table 1: General practitioners -- equity indices and efficiency scores by province (2016).................................17
Table 2: Gynaecologists -- equity indices and efficiency scores by province (2016)..............................................21
Table 3: Paediatricians -- equity indices and efficiency scores by province (2016)...............................................25
Table 4: Psychiatrists -- equity indices and efficiency scores by province (2016)...............................................29

LIST OF FIGURES

Figure 1: Strategic objectives 2018/2019.................................................................1
Figure 2: Percentage distribution of medical scheme beneficiaries by province (Dec. 2016)........................................12
Figure 3: Availability of healthcare providers by discipline (Dec. 2016).................................................................13
Figure 4: Percentage distribution of providers by province -- private vs. public sector (2016)............................13
Figure 5: GPs -- production returns to scale (Dec. 2016)......................................................................................14
Figure 6: Returns to scale -- GPs (Dec. 2016).........................................................................................................15
Figure 7: Technical efficiency -- GPs (Dec. 2016)....................................................................................................16
Figure 8: Overall, pure & residual efficiency -- GPs (Dec. 2016)........................................................................16
Figure 9: Scale efficiency -- Gynaecologists (Dec. 2016)......................................................................................18
Figure 10: Residual efficiency -- Gynaecologists (Dec. 2016)............................................................................19
Figure 11: Overall, pure & residual efficiency -- Gynaecologists (Dec. 2016)......................................................20
Figure 12: Scale efficiency -- Paediatricians (Dec. 2016).......................................................................................22
Figure 13: Residual efficiency -- Paediatricians (Dec. 2016)..................................................................................23
Figure 14: Overall, pure & residual efficiency -- Paediatricians (Dec. 2016)......................................................24
Figure 15: Scale efficiency -- Psychiatrists (Dec. 2016).......................................................................................26
Figure 16: Residual efficiency -- Psychiatrists (Dec. 2016)..................................................................................27
Figure 17: Overall, pure & residual efficiency -- Psychiatrists (Dec. 2016)...........................................................28
Figure 18: Illustration of elements used for calculating the hoover index.................................................................43
Figure 19: Illustration of elements used for calculating the hoover index.................................................................44
Figure 20: Illustration of elements used for calculating the decile ratios.................................................................44
Figure 21: Production frontiers, efficiency and pareto allocation..............................................................................45
Figure 22: Implementation of CRS model in excel solver (input method)..............................................................46
Figure 23: Implementation of VRS model in excel solver (input method).............................................................48
Figure 24: Implementation of non-increasing returns model in excel solver (input method)...............................50
Figure 25: Implementation of Mini-max model in excel solver (additive method)................................................52
1. **INTRODUCTION**

1.1 **Problem Statement**

There are inequalities in the access of healthcare providers across geographic markets in the private healthcare industry. This has an impact on primary healthcare re-engineering drives for National Health Insurance (NHI), implementation of appropriate integrated care networks, and guaranteed access to healthcare benefit for all. A solution to the potential coordination problem, amongst other, is to identify markets where resources are underproductive, and redirect them to geographic markets where increasing returns to scale are present.

1.2 **Background**

The Council for Medical Schemes (CMS) and the National Department of Health (NDoH) are in a collaborate drive to increase health care for all in South Africa. The NDoH has released its White Policy Paper on NHI (NDoH, 2016). The CMS has aligned its 2017/2018 (CMS, 2018) strategic objectives with those of the NDoH to prepare the market for NHI.

The Competition Commission’s Health Market Inquiry (HMI), seeks to identify market interventions that increase the efficient operation of the health market (HMI, 2017). The CMS participates in the inquiry and seeks to make recommendations and implement the HMI’s recommendations.

Figure 1 below reflects an alignment of the CMS strategic objectives with those of the NDoH, and also shows the strategic objectives which are aligned with market structure issues.

---

**Figure 1: Strategic objectives 2018/2019**

*Note: National Department of Health strategic goals aligned with National Development Plan and CMS strategic goals for 2018. (See NDoH Strategic Plan 2015-2019, p.22).*
The distribution of healthcare providers has been identified as a project that is the feed into the NDoH strategic goal of re-engineering the primary healthcare system in preparation of NHI. That said; as all policies should coalesce, the provision of health services in the different provincial/geographic healthcare markets should complement other initiatives. This policy synergy should increase the yield in policy outcomes, as returns to scale from policy initiatives increase.

Therefore; figure 1 identifies some of the other policy areas in which the technical efficiency of allocated of healthcare resources, should efficiently provide services for the health financing component in the demand side. This is imperative the objective of the HMI, which is to recommend initiatives that are to increase the efficiency of healthcare market, both on the demand and supply side.

The mixed policy bag interlinks as follows:

i. Healthcare providers should provide quality healthcare services for medical scheme beneficiaries in all geographic markets. If consumer engagements with healthcare providers are based on a positive patient experience, healthcare outcomes will improve. This will have a positive impact on returns to scale.

ii. If medical scheme beneficiaries are able to purchase healthcare services with the purchasing power of a consolidated demand side, through scheme and option consolidation, returns to scale in the system should improve.

iii. If medical scheme beneficiaries are able optimise the purchase of healthcare services with aid of standardised benefit options. The bargaining process should be more transparent and yield higher returns to scale in the private healthcare system. This is also an issue in which CMS is to support the HMI price regulation process.

iv. If healthcare resources in geographic markets with diminishing returns to scale are redirected to geographic markets with increasing returns to scale, then equality in healthcare access will be achieved at lower average costs. This should support the process of making healthcare costs lower and attract lower-middle income beneficiaries into the medical schemes industry. Preparing the country for universal health cover.

An economic evaluation of healthcare resource efficiency will be used to describe efficiency of health providers (inputs) in delivering healthcare services that are paid by medical schemes (outputs). The healthcare provider variables that are factored into the economic evaluation model, are informed by the CMS position (CMS, 2017) on: i) its revision of the Prescribed Minimum Benefits (PMBs); and ii) the role out of the essential benefit package for the NHI.
1.3 Purpose

The purpose of market delineation or relevant market identification models are to not only assess the possibility of market power, but to also to describe the market forces behind prices. Otherwise, the issue of whether efficient resources are efficiently allocated across needs (demand), cannot be properly addressed.

It has been recently confirmed that market delineation models do not adequately explain the rationale behind the location of service providers within geographic markets. Nor do they adequately provide for product market realities in healthcare delivery markets (this is explained in the literature review).

As is mentioned in the background, both the issues in which market definition models are lacking are at the cornerstone the strategic objectives. Re-engineering the primary healthcare system is about accessibility to healthcare, and thus redistributing health resources to where are needed and more productively used (where there are increasing returns to scale). The standardisation of benefit options can render consumers more active in disciplining utilised resources or technologies to rein in costs.

This analysis intends to be more than just a mere stock-taking exercise on human capital in the healthcare delivery market. It seeks to identify and implement a methodology for identifying residual capacity in over served geographic healthcare markets and freeing these resources for underserved areas to participate in the healthcare market.

The analysis provides a baseline. That said; the baseline supports and must therefore be discussed within the policy paradigm in which the right market incentives need to be provided to resolve inequality issues. Ultimately; the document seeks to discuss the policy impact of the residual efficiency, on other policy issues faced in the healthcare policy agenda. Using the NHI and PMB prioritised service benefits provides an empirical study on where efficiencies lie for the in the private healthcare market.

1.4 Objectives

The study's objectives are to:

i. Conduct an efficiency analysis by discipline and geographic market, using an economic evaluation tool.

ii. Describe resource efficiency capacities by healthcare discipline across provinces, and by relevant peer group.
iii. Decompose the overall efficiency of healthcare human resources into their pure and residual efficiency components.

iv. Identify which healthcare markets have increasing or decreasing returns to scale.

v. Identify and discuss supporting policy interventions for redistributing healthcare resources to underserved geographic healthcare markets.
2. METHODS

2.1 Literature Review

The difficulty that the Federal Trade Commission (FTC) (in the United States) has faced identifying ‘relevant markets’ for merging hospitals, is relevant for any jurisdiction seeking to make an impact on competitive forces within geographic spaces. This is equally relevant when healthcare resources need to be provided on a more equitable basis than what market efficiency can provide.

Elzinga and Hogarty (Elzinger & Hogarty, 1973) formulated a method, called the E-H method, for delineating geographic markets. The method relies on hospital customer or service area data. The method has subsequently been found wanting in accurately identifying ‘relevant markets’ when dealing hospital mergers (Gaynor & Pflum, 2017). The E-H approach does not appropriately account substitutive geographic markets. It understates the anti-competitive impact of potential post-merger price increases, by overestimating the size of the relevant geographic market of healthcare providers. Particularly hospitals.

Overstating the size of a geographic market results in the following inappropriate conclusion by competition authorities or courts: i) Patients/consumers in an internal market are assumed to be far from the proximity of substitute products in alternative geographic markets; ii) Therefore, if there to be a small but significant post-merger hospital increase, the impact of the magnitude would of the price increase would be understated; iii) This is because there would not be an alternative hospital/competitor within close proximity of consumers, that could perhaps offer a lower price to correct anti-competitive behaviour.

Consumers are not able to optimally rationalize a price with the first willing provider. Equal access to healthcare should not be presumed to be a given in the private healthcare market, because consumers play a passive role in the actual contracting arrangements between managed care organisations (MCOs), and healthcare providers. Therefore, the outcomes of the movements of the ‘invisible hand’ in healthcare markets, might not necessarily be Pareto Optimally fair.

Elzinga and Swisher (2011) cite Deborah Haas-Wilson (2003, pp. 125–6), when they say there two healthcare markets in the supply-side. These are the primary and secondary markets.

In the primary market; MCO’s contract on behalf of patients when forming networks and pricing services with provider groups. The patient is passive in the pricing decision and in terms of the E-H model, this does not impact the rational behaviour of patients. The prices are borne by health funders, and not directly paid by consumers. Therefore, these prices have no impact on the decision of patients to move to other providers’ geographic areas.
It is only in the second stage market when patients choose among available providers in the managed care networks devised by MCOs and provider groups. Therefore E-H models say something about competition within geographic markets, but very little about the how consumers respond to the pricing mechanism set in the primary market.

The E-H approach does not do so because it overstates the true size of the geographical market, and the true effect of price changes are understated. The standard merger guidelines assume that healthcare consumers actively participate in price setting in the primary market, through finding substitute product in adjacent healthcare markets.

The flaw here is that geographic market models should capture the demand and supply mechanisms underlying price, and thus efficiency or constraints to efficient outcomes. Market definition methods capture market concentration, but they are inadequate when dealing with the forces underlying that market concentration.

Another reason why the E-H model may be inappropriately applied in the hospital or healthcare market, is that products are not appropriately bundled. This creates an incorrect representation of the true market structure, and inappropriate market correcting interventions. Clustering healthcare services/products together which are not substitutable, renders an assessment of two separate as though they were a single market.

To be specific, combining acute and inpatient hospital care into a single bundle makes differentiated healthcare services, one service. For example; patients requiring primary and secondary are less likely to have long referral paths, relative to patients requiring specialised surgery. This means that markets for hospitals are inappropriately delineated within the same geographic market as primary and secondary care markets. Therefore, one may infer; reorganising the healthcare system through primary healthcare networks, may make the identification of relevant (geographic) markets easier for managing anti-competitive behaviour.

GPs are the first point of contact for beneficiaries in a primary care setting. Belgium implemented their Impulse Fund policy to make GP healthcare services more accessible in underserved communities (Gerkens & Merkur, 2010: 158).

There are some literature which have: i) grappled with technical efficiencies behind healthcare facility costs across different geographical healthcare markets; ii) explained the existence of residual efficiencies in healthcare markets by way of a ‘healthcare arms race’ hypothesis, and that this may lead to ‘tragedy of the commons’ scenario; and iii) sought to understand why healthcare facilities, or service providers tend to operate in specific locations; pursuant to recommending appropriate interventions for a more equitable distribution of healthcare resources.
Devers et al. (2003) review the medical arms race hypothesis and explain its impact on that it has on the behaviour hospitals in healthcare markets. (Devers, et al., 2003) make a significant observation. They find that although hospital markets are concentrated, hospitals compete.

The authors find that hospitals compete by attracting the best specialists and purchasing the most technologically advanced equipment to differentiate themselves from competitors. Hospital competition is not perceived through prices, but rather, on the quality perceptions made by targeted consumers/patients.

The truth is that quality in healthcare markets might not be perceptible by patients/consumers. That judgement is made by the referral system that prefers specific facilities. The patient plays a role once again. This time it's not in the primary healthcare market (Elzinga & Swisher, 2011), but also in the secondary healthcare market (Devers, et al., 2003).

Devers et al (Devers, et al., 2003) find that the nature of the competition between healthcare facilities, is such that they each player responds as best they can. The best they can in a market the best secondary healthcare resources and equipment win market share.

My concern is whether this “one-upmanship” creates residual efficiency in markets with decreasing returns to scale, at the expensive of underserved markets where increasing returns to scale exist. This is of concern because such a scenario tells of unhealth laissez fair competition is in-efficient and glaringly unfair. The in-efficient aspect suggests that there is no pareto-optimal allocation, both in overserved and underserved geographical markets.

This misallocation introduces a notion of waist, one that is all too common when there is an argument for regulatory coordination. A discussion that has taken place within the theoretical framework that Hardin (Hardin, 1968) called the “Tragedy of the Commons”.

The notion of the Tragedy of the Commons emanates from Harding's hypothesis on what the implications of over utilising land has on the marginal utility gained from exploiting it. His exposition suggested that the quality and satisfaction gained from overutilization would decrease over time. Such that the common resources will grow obsolete or society would become impoverished. Harding's contribution to what the solution of such social problems would is coined in the following phrase: “the population problem has no technical solution; it requires a fundamental extension in morality” (Hardin, 1968).

A similar sentiment was held by David Hume (A treatise on human nature, 1739). Hume said the natural behaviour of human beings are to maximize their utility in gaining pleasure. Gibbons (Gibbons, 1992) provides a game theory driven analogy of Hume's argument., Gibbons (1922) shows that, when resources are limited, competitive behaviour
tends to overutilize or produce goods. This happens at the expense of quality outcomes. He showed seeking a social optimum, provides society with better utility than in perfect competitive circumstances. Gibbon's method has been used to craft social equilibriums that reduce waste and optimise social efficiency.

The identification of geographic markets that have diminishing returns to scale and residual efficiency, may invest in healthcare resources at the cost of receiving diminishing or negative utility for society. Data envelopment or economic optimization techniques have used to identify economic units that operate at increasing or decreasing returns to scale. These methods have used to study the efficiency of public or private healthcare facilities, in order to free up resources to invest in areas that are underserved areas.

The type of economic solutions that applied to this type of ‘wicked policy problem’ are the ‘Hotelling linear city’ (Hotelling, 1929) or ‘Salop circular city’ (Salop, 1979) spatial economics models. Some literature has also introduced product market competition to the analysis competition in geographic markets (Economides, 1993). Some healthcare economists have applied these models to healthcare geographic markets (Brekke, et al., 2010), (Brekke, et al., 2006). Others have considered market forces behind prices, when location decisions are made in geographic markets (Anderson & Engers, 1994). The objective of these game theory models to: i) understand how the demand and supply market forces generate competitive equilibriums within spatial or geographic markets; and ii) some health economists have used these to formulate policy recommendations for deriving socially optimal and equal resource allocation.

The review will provide only passing commentary on these spatial economics methods, as they will not be applied in the analysis section. That said, they important to mention. This is because they provide insight on market forces in geographical markets where market delineation methods or merger guideline techniques are not able. Most importantly, they are relevant in explaining why healthcare resources are concentrated in specific geographic markets. For example, the workings of product market competition in geographic markets. They also support the argument of “supply induced demand” and the “new medical arms race hypothesis”. These concepts are important, as they may explain why these are high residual efficiencies in some geographic healthcare markets.

The significance of Hotelling’s (Hotelling, 1929) paper in terms of why sellers will operate in close proximity, confirms the concentration in healthcare markets identified in Devers et al ‘new medical arms race’ hypothesis. An important aspect of market competition from a spatial perspective, is the equitable distribution of service providers, relative to, the distribution of consumers. Hotelling (1929) introduced the basis upon which service providers compete within spatial locations. Hotelling (1929) said that service providers will locate close to each other to optimise their profits. If healthcare providers are, near each other, the more concentrated the market in certain parts of a province. This
means that Gini-coefficients will describe poor equity outcomes. This means that healthcare providers will not be
distributed in a socially optimal fashion.

The contribution of Brekke et al (Brekke, et al., 2010) and Economides explain Devers et al (Devers, et al., 2003)
“one-upmanship” strategic behaviour of healthcare provider groups in out competing each other. Their contribution
imputes into geographic market analysis, the significance of consumption bundles and product market competition
that is absent in ‘relevant market’ models.

Brekke et al, Economides, and Anderson & Engers have extended to his work, by considering the socially optimal
distribution of providers of goods and services. Anderson and Engers (1994) and Economides (1993) suggest that
regulator intervention may have an impact on spatial competition. They argue that price interventions and minimal
quality regulations may result in spatial competition which is socially optimal (Anderson & Engers, 1994) and
(Economides, 1993).

Kutzin (2000) provides a framework for health system planning. He states that; the efficiency of health delivery
function is essential for achieving equity in accessing healthcare services (Kutzin, 2000: 41).

Kutzin (2000) explains -- if resource capacity is adjusted to achieve allocative efficiency, excess capacity could be
redistributed to underserved regions. The criteria for redistributing healthcare resources is, deserving regions should
have increasing returns to scale (Australia. Steering Committee for the Review of Commonwealth/State Service
Provision, 1997), (Zhang, et al., 2015), and (CIHI, 2016).

Reorganising risk pooling arrangements on a geographical basis allows for more effective healthcare cover. Large
risk pools, covering large geographical areas, are beneficial for reducing utilisation costs and improving benefit
allocation. That said; differences in equal healthcare access, and effective cover (real access to benefits and financial
protection) across regions, hampers the existence of markets for medical scheme beneficiaries. Disparities in
accessing healthcare cover result in market failure.

The coordination problem lies in the multiple objective scenario of optimising resource capacity, and access to health
services/benefits. Efficiency in the provision of health services across geographically designated healthcare markets,
may release healthcare resources from richer to more deprived provinces.

Economic evaluations have been used to measure the efficiency of healthcare resources. They have also been used
to evaluate government programs. These are listed as follows: (Banker, et al., 1984), (Banker, et al., 2004), (CIHI,
2016), (Jat, et al., 2013), (Munisamy, 2016), (Zhang, et al., 2015).
2.2 Empirical Methods

The method for deriving the returns to scale/economic efficiency of different geographic markets is based on the data envelopment (DEA) method (Banker, et al., 1984) and (Ragsdale, 2008). The method is appealing as it can be used to derive economic meaning from results if adapted into an economic problem-solving model (Dixit, 1991). The DEA can be represented as a Lagrange economic optimisation method for optimising the production function different economic units, subject to a series of constraints.

Kuhn-Tucker slack constraints can then be substituted into the objective functions of the economic units. This done through the Lagrange lambda ($\lambda$) shadow prices taking on the property of the efficiency weights in the DEA model. This gives the system of equations the property of a goal programming multiple objective function, with pareto optimisation characteristics associated with the objective functions (Ragsdale, 2008) and (Jablonsky, 2014). This explain the mini-max objective function for determining the trade-offs (see technical notes in section 7).

The economic evaluation models are implemented using the Excel Solver Add-In. The equations and implementation methodology are explained in the technical note (see technical notes in section 7).

The economic evaluation model assesses efficiency from the following perspectives of returns to scale:

i. Constant Returns to Scale

The system of equations is evaluated as though each of the geographic markets were similar in scale and economic activity (Industry Commission, 1997). Therefore, each market's overall efficiency score is compared with the market highest efficiency score, assuming that all other markets are within the same peer group. The Constant Returns to Scale (CRS) frontier is assumed to be log-linear in functional form (see fig. 25 in technical note 7.4).

ii. Variable Returns to Scale

The Variable returns to Scale (VRS) method is used to relax the assumptions of the CRS method. Therefore, the scale and size of the different geographic healthcare markets are not assumed to be the same. The size characteristic is thus made an exogenous factor, and relative market efficiency scores are derived using the VRS efficiency frontier (see fig. 25 in technical note 7.4). The VRS frontier is assumed to have a logarithmic functional form. The VRS method endogenizes the competitive characteristics within each geographic market, making it easy to say something unique about specific geographic markets. This is also nice feature to have, when making inferences about competition in different geographic markets (Industry Commission, 1997). Overall efficiency scores can also be decomposed to assess pure technical efficiency and residual technical efficiency in individual geographic markets (Industry Commission, 1997).
iii. Increasing & Decreasing Returns to Scale

The determination of Increasing Returns to Scale (IRS) and Decreasing Returns to Scale (DRS) are explained in technical note 7.9 (also see figure 25 in technical note 7.4).

iv. Pareto Optimality/Trade-Off

See technical note 7.10 and figure 25.

Equality in accessing healthcare services are derived through the calculation of inequality indicators. The equations and meaning of the indicators are explained in the technical note (see section 7). The methodologies have been sourced from the following research papers: (Kondor, 1997), (Maio, 2007), (Wagstaff, et al., 1989), (Zhang, et al., 2015), (Taylor, et al., 2000), (Watkins, et al., 2014).

2.3 Limitations

The limitations of the analysis are:

i. Data not adjusted for severity of risk pools, that said; differences in severity could also be caused by inefficient management of chronic diseases.

ii. GP counts are based on GPs who have claimed and been paid by medical schemes.

iii. The inequality indicators are driven by the postal based on the cumulative distribution amounts paid to GPs by postal code, relative to, the cumulative distribution of beneficiaries by postal code. The data for describing GPs at postal code level, are new data. Therefore, the Gini-coefficient should be interpreted with due consideration.
3. RESULTS

3.1 Overall Market Structure: Distribution of Providers & Beneficiaries

Figure 2 shows the percentage distribution of medical scheme beneficiaries across the nine South African provinces, in December 2016. The distribution of medical scheme beneficiaries across the provinces/provincial healthcare markets, is as follows:

- Gauteng 39.4%
- Western Cape 15.1%
- KwaZulu-Natal 14.5
- Eastern Cape 7.5%
- Mpumalanga 6.5%
- North West 5.6%
- Limpopo 4.7%
- Free State 4.5%; and
- Northern Cape 2.1%.

Figure 2: Percentage distribution of medical scheme beneficiaries by province (Dec. 2016)
Source: Data are sourced from CMS utilisation statutory return.
NB: The unclassified component of beneficiary location is excluded

Figure 3 describes the availability of healthcare provider disciplines across the nine South African provinces, in December 2016. The availability of healthcare provider disciplines across the provinces/provincial healthcare markets, is as reflected in figure 2.
Figure 3: Availability of healthcare providers by discipline (Dec. 2016)
NB: The unclassified component of healthcare provider location is excluded

Figure 4 shows contrasts the percentage provincial distribution of provider disciplines in the private sector, with that of the public sector. It is clear that there are large differences in the endowment between the private and public sector. The imbalances are even gaping when specialities are included.

Figure 4: Percentage distribution of providers by province -- private vs. public sector (2016)
Data Source: Utilization returns CMS and (Day & Gray, 2016) SA Health Review 2016.
3.2 Primary Care: General Practitioners

3.2.1 Efficiency Analysis: GPs

Figure 5 below provides an approximation of increasing and decreasing returns to scale across the nine provinces. The x-axis describes the number of GPs per province, and the y-axis reports the number of GP visits paid for by medical schemes.

A logarithmic function maps the number of GPs per province, to total GP visits paid for by medical schemes (fig. 5). The trend line is a crude representation of increasing and decreasing returns to scale. The trend line shows that there are increasing returns to scale for Northern Cape, North West, Limpopo and Free State. There are decreasing returns to scale for Western Cape, KwaZulu-Natal and Gauteng (fig. 5). This is a crude representation; the figure does not represent results that are generated by DEA model.

![Figure 5: GPs – Production returns to scale (Dec. 2016). Source: Data are sourced from CMS utilisation statutory returns.](image)

Figure 6 describes the efficiency of provinces at the scale at which they are operating. The x-axis shows the number of GPs per province. The y-axis shows the returns to scale; the marginal rate of efficiency associated with an increase in GP resources. The slope of the curve reduces from left to right (fig. 6).

There are benefits in increasing the number of GPs that are available in Northern Cape, North West, Free-State, Limpopo and Mpumalanga. The highest benefit in increasing GPs accrues to Northern Cape, and the lowest benefit accrues to Mpumalanga (fig. 6).
There are no benefits in increasing the number of GPs that are available in KwaZulu-Natal, Western Cape, and Gauteng. The provinces’ marginal rates of return are zero, which is reflected by flat slope running from KwaZulu-Natal to Gauteng (fig. 6).

Figure 6: Returns to scale – GPs (Dec. 2016)
NB: The graph is generated using Variable Returns to Scale

Figure 7 describes the technical efficiency of GPs across the nine South African provinces, using December 2016 data. It is the result of decomposing overall efficiency of GPs per provinces, into their pure efficiency and residual efficiency components. Pure and residual efficiency describe the nature of the technical efficiency that is present.

Pure efficiency is defined by the ratio between CRS and VRS scores. This ratio is called Scale Efficiency (SE) the score (see note 8.7 for derivation). “Residual efficiency is technical efficiency that result from factors not efficiency related” (Industry Commission, 1997, p. 17). Therefore, it is “technical inefficiency” (Industry Commission, 1997, p. 17).

The x-axis in figure 7 is the SE score derived from comparing CRS and VRS scores using an input method of the DEA model. The y-axis in figure 7 is generated by decomposing actual visits paid into the SE and residual efficiency components. The y-axis reflects the residual efficiency component of total number of GP visits paid for. Therefore figure 7 shows the relationship between the two components describing the nature of technical efficiency for GPs. There is a negative relationship between the pure/SE and residual efficiency.

Gauteng, KwaZulu-Natal and Western Cape has the greatest residual efficiency, and least pure efficiency. The other provinces show very minimal residual efficiency and thus high SE or pure efficiency (fig. 7).
Figure 7: Technical efficiency -- GPs (Dec. 2016)

NB: The graph is generated using Variable Returns to Scale

Figure 8 is a web diagram that shows the relative relationships between overall efficiency, pure efficiency and residual efficiency. The overall efficiency is the CRS efficiency score. The pure efficiency represents the distance between the CRS and VRS frontiers. The residual efficiency represents the distance between the VRS-frontier, and the actual level of efficiency generated by providers.

In Eastern Cape, GPs are relatively more efficient than other provinces and thus the performs at 100% pure efficiency, and 0% inefficiency (residual efficiency). This is reflected by the zero distance in the fractions representing pure efficiency and residual efficiency (fig. 8).

In Gauteng, GPs operate at 83% overall efficiency (CRS score). Their pure efficiency is at 83%, but their residual efficiency (inefficiency) 17%. Gauteng has residual efficiency even though its VRS score is 100% because on non-scale factors (Industry Commission, 1997, p. 17). Non-scale factors result from technical in-efficiencies not associated with scale efficiencies (SE) (fig. 8).

Figure 8: Overall, pure & residual efficiency -- GPs (Dec. 2016)

Note: Overall efficiency assumes constant returns to scale
3.2.2 Equality, pareto optimality & trade-offs: GPs

Table 1 describes: i) the number of GPs per 10,000 beneficiaries (density ratio); ii) equity in accessing GPs for all provinces (Gini Co-efficient (Gini), Hoover Index (HI) and decile ratios); and iii) the relative efficiency of GPs for all provinces (efficiency score).

Table 1 shows that the provinces that have Decreasing returns to Scale (DRS) are Gauteng, KwaZulu-Natal and Western Cape. The other provinces have Increasing Returns to Scale (IRS). At the province level, IRS means that further investment in producing or transferring GPs from other provinces will increase productivity. DRS means that further investment in producing or transferring GPs from other provinces will not be productive. The only source of productivity will be to re-allocate resources within the provinces, depending on the level of needs reflected in the inequality indicators.

Table 1: General practitioners -- equity indices and efficiency scores by province (2016)

<table>
<thead>
<tr>
<th>PROVINCE</th>
<th>GPs per 10,000 beneficiaries</th>
<th>GINI CO-EFFICIENT</th>
<th>HOOVER INDEX</th>
<th>DECILE RATIO (20:80)</th>
<th>DECILE RATIO (40:60)</th>
<th>CRS EFFICIENCY SCORE (%)</th>
<th>VRS EFFICIENCY SCORE (%)</th>
<th>SCALE EFFICIENCY SCORE (%)</th>
<th>TYPE OF SCALE EFFICIENCY</th>
<th>TRADE-OFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern Cape</td>
<td>15.9</td>
<td>0.529</td>
<td>0.41</td>
<td>37</td>
<td>5</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>IRS</td>
<td>12.0</td>
</tr>
<tr>
<td>Free State</td>
<td>18.8</td>
<td>0.445</td>
<td>0.34</td>
<td>16</td>
<td>3</td>
<td>79.0</td>
<td>79.4</td>
<td>99.2</td>
<td>IRS</td>
<td>34.0</td>
</tr>
<tr>
<td>Gauteng</td>
<td>13.3</td>
<td>0.634</td>
<td>0.53</td>
<td>14</td>
<td>4</td>
<td>83.0</td>
<td>100.0</td>
<td>83.4</td>
<td>DRS</td>
<td>5.4</td>
</tr>
<tr>
<td>KwaZulu-Natal</td>
<td>16.5</td>
<td>0.492</td>
<td>0.38</td>
<td>37</td>
<td>6</td>
<td>89.0</td>
<td>100.0</td>
<td>89.4</td>
<td>DRS</td>
<td>12.0</td>
</tr>
<tr>
<td>Limpopo</td>
<td>18.9</td>
<td>0.416</td>
<td>0.31</td>
<td>77</td>
<td>8</td>
<td>81.0</td>
<td>81.6</td>
<td>99.4</td>
<td>IRS</td>
<td>18.1</td>
</tr>
<tr>
<td>Mpumalanga</td>
<td>16.8</td>
<td>0.358</td>
<td>0.28</td>
<td>48</td>
<td>5</td>
<td>88.0</td>
<td>88.4</td>
<td>99.7</td>
<td>IRS</td>
<td>27.0</td>
</tr>
<tr>
<td>Northern Cape</td>
<td>13.8</td>
<td>0.503</td>
<td>0.37</td>
<td>93</td>
<td>8</td>
<td>97.0</td>
<td>100.0</td>
<td>96.9</td>
<td>IRS</td>
<td>0.0</td>
</tr>
<tr>
<td>North West</td>
<td>11.4</td>
<td>0.403</td>
<td>0.35</td>
<td>70</td>
<td>6</td>
<td>95.0</td>
<td>95.9</td>
<td>99.0</td>
<td>IRS</td>
<td>47.3</td>
</tr>
<tr>
<td>Western Cape</td>
<td>17.2</td>
<td>0.51</td>
<td>0.41</td>
<td>9</td>
<td>2</td>
<td>71.0</td>
<td>70.0</td>
<td>91.2</td>
<td>DRS</td>
<td>11.0</td>
</tr>
</tbody>
</table>

NB: The decile ratios should be interpreted as rand amounts.

Table 1 shows that the provinces that have Decreasing returns to Scale (DRS) are Gauteng, KwaZulu-Natal and Western Cape. The other provinces have Increasing Returns to Scale (IRS). At the province level, IRS means that further investment in producing or transferring GPs from other provinces will increase productivity. DRS means that further investment in producing or transferring GPs from other provinces will not be productive. The only source of productivity will be to re-allocate resources within the provinces, depending on the level of needs reflected in the inequality indicators.

The trade-offs for adjusting to GP visit levels relative to the number of available GPs, provides the scope for reallocating resources from over resourced parts of provinces to underserved regions in provinces. The obvious re-allocations should be identified by the 20:80 decile ratios (table 1). The highest 20:80 decile ratios are for Northern Cape, Limpopo, and North West. The lowest 20:80 decile ratios are for Western Cape, Gauteng and Free State (table 1).

In Northern Cape R93 per capita is paid in the highest healthcare expenditure quintile for GP visits, relative to R1 in the lowest healthcare expenditure quintile (table 1). Because the Northern Cape’s VRS is close to its CRS, it is already performing at it’s true SE. Therefore, there is no room for adjusting its operating scale, the trade-off is 0 and is likely to be at Pareto optimality. Northern Cape would benefit from a Virtual Risking transfer from outside the province and into the province.
In Western Cape R9 per capita is paid in the highest healthcare expenditure quintile for GP visits, relative to R1 in the lowest healthcare expenditure quintile (table 1). Because the Western Cape's combined trade-off adjustment of GP visits and number of available GPs is 11%, transfers within Western Cape might improve the returns to scale from being DRS to being IRS. This may also improve the Hoover Index which requires utilisation expenditure of 41% from the most endowed to the least endowed parts of Western Cape (table 1).

3.3 Maternity Benefits: Gynaecologists

3.3.1 Efficiency analysis: Gynaecologists

Figure 9 describes the efficiency of provinces at the scale at which they are operating. The x-axis shows the number of Gynaecologists per province. The y-axis shows the scale efficiency (SE) derived from applying an input method to deriving CRS efficiency and VRS efficiency scores. The interpretation is that the pure efficiency achieved by providers decreases, the more the providers in a province. This can be seen in slope of the curve; however, the rate of decrease reduces at a decreasing rate from left to right (fig. 9). However, the greater the number of available providers, the greater the likelihood of residual efficiency (inefficiency).

Northern Cape has a scale efficiency of 1, which makes it the most productive province. Limpopo, Free State, Mpumalanga and Eastern Cape cluster together at less than 100 gynaecologists per province. Followed by KwaZulu-Natal and Western Cape, which have 200-300 gynaecologists per province. Gauteng has the highest number of gynaecologists, but the lowest SE (pure efficiency) (fig. 9).

Figure 9: Scale Efficiency – Gynaecologists (Dec. 2016)

Figure 10 describes the technical efficiency of Gynaecologists across the nine South African provinces, using December 2016 data. The figure is generated by decomposing overall efficiency of Gynaecologists per provinces,
into their pure efficiency and residual efficiency components. Pure and residual efficiency collectively describe the nature of the technical efficiency that is present.

Pure efficiency is defined by the ratio between CRS and VRS scores. This ratio is called the Scale Efficiency (SE) score (see note 7.7 for derivation). “Residual efficiency is technical inefficiency that result from factors that are not efficiency related” (Industry Commission, 1997, p. 17). Therefore, it is “technical inefficiency” (Industry Commission, 1997, p. 17).

The x-axis in figure 10 is the SE score derived from comparing CRS and VRS scores using an input method of the DEA model. The y-axis in figure 10 is generated by decomposing actual visits paid into the SE and residual efficiency components. The y-axis reflects the residual efficiency (technical inefficiency) component of total number of Gynaecologist visits paid for. Therefore figure 10 shows the relationship between the two components describing the nature of technical efficiency for Gynaecologists. There is a negative relationship between the pure efficiency and residual efficiency (technical inefficiency).

Gauteng, and KwaZulu-Natal and Western Cape have the greatest residual efficiency (technical inefficiency), and least pure efficiency. The other provinces show residual efficiency that is lower than 50,000 visits per province. Northern Cape, Free State and Limpopo have the highest pure efficiency (scale efficiency) (fig. 10).

Figure 10: Residual Efficiency -- Gynaecologists (Dec. 2016)

Figure 11 is a web diagram that shows the relative relationships between overall efficiency, pure efficiency and residual efficiency. The overall efficiency is the CRS efficiency score. The pure efficiency represents the distance between the CRS and VRS frontiers. The residual efficiency represents the distance between the VRS frontier, and the actual level of efficiency generated by providers. For figure 11, pure and residual efficiency take on a number in the range from 0 to 1 ($d \in [0,1]$). They reflect the efficiency operating distance from the CRS and VRS efficiency frontiers (fig. 11).
The largest residual efficiencies (technical inefficiencies) are found on the left-hand side of figure 11. These provinces are Gauteng, KwaZulu-Natal, and Eastern Cape. The provinces that enjoy the largest pure efficiency from technical allocations combining input and outputs are at the right-hand side of figure 11. These provinces are Free State and Western Cape.

In Northern Cape, Gynaecologists are relatively more efficient than other provinces, they perform 100% pure efficiency, and 0% residual efficiency (technical inefficiency). This is reflected by the zero distance in the fractions representing pure efficiency and residual efficiency (fig. 11).

In Gauteng, Gynaecologists operate at 41% overall efficiency (CRS score). The distance between the CRS and VRS efficiency frontiers is zero, thus having a pure efficiency of 0 (fig. 11). They therefore operate at scale efficiency. That said, there are non-efficiencies that creep into this market. Residual efficiency (technical inefficiency) results in Gynaecologists operating 0.4 away from the markets' VRS frontier (fig. 11).

In Free State, Gynaecologists operate at 46% overall efficiency (CRS score). The distance between the CRS and VRS efficiency frontiers is 0.12. Thus; having a pure efficiency and scope for technical allocations to increase efficiency to SE level in certain parts of the province (fig. 11).

### Figure 11: Overall, pure & residual efficiency -- Gynaecologists (Dec. 2016)

*Note: Pure and residual efficiency are distances from efficiency frontiers.*

#### 3.3.2 Equity, pareto optimality & trade-offs: Gynaecologists

Table 2 describes: i) the number of Gynaecologists per 10,000 beneficiaries (density ratio); ii) equity in accessing Gynaecologists for all provinces (Gini Co-efficient (Gini), Hoover Index (HI) and decile ratios); and iii) the relative efficiency of Gynaecologists for all provinces (efficiency score).
Table 2: Gynaecologists -- equity indices and efficiency scores by province (2016)

<table>
<thead>
<tr>
<th>PROVINCE</th>
<th>Gynaecologists (per 10,000 beneficiaries)</th>
<th>GINI CO EFFICIENT</th>
<th>HOOVER INDEX</th>
<th>DECILE RATIO (20:80)</th>
<th>VRS EFFICIENCY SCORE (%)</th>
<th>VRS EFFICIENCY SCORE (%)</th>
<th>SCALE EFFICIENCY SCORE (%)</th>
<th>TYPE OF SCALE EFFICIENCY</th>
<th>TRADE-OFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern Cape</td>
<td>1.2</td>
<td>0.75</td>
<td>0.60</td>
<td>300</td>
<td>65</td>
<td>52.6</td>
<td>51.4</td>
<td>57.6</td>
<td>DRS</td>
</tr>
<tr>
<td>Free State</td>
<td>1.6</td>
<td>0.74</td>
<td>0.65</td>
<td>455</td>
<td>178</td>
<td>34.0</td>
<td>34.0</td>
<td>34.0</td>
<td>IRS</td>
</tr>
<tr>
<td>Gauteng</td>
<td>1.7</td>
<td>0.73</td>
<td>0.65</td>
<td>119</td>
<td>28</td>
<td>41.0</td>
<td>41.0</td>
<td>41.0</td>
<td>DRS</td>
</tr>
<tr>
<td>KwaZulu Natal</td>
<td>1.5</td>
<td>0.70</td>
<td>0.60</td>
<td>450</td>
<td>76</td>
<td>49.7</td>
<td>49.7</td>
<td>49.7</td>
<td>DRS</td>
</tr>
<tr>
<td>Limpopo</td>
<td>0.7</td>
<td>0.71</td>
<td>0.59</td>
<td>651</td>
<td>127</td>
<td>73.6</td>
<td>73.6</td>
<td>73.6</td>
<td>DRS</td>
</tr>
<tr>
<td>Mpumalanga</td>
<td>0.9</td>
<td>0.73</td>
<td>0.64</td>
<td>111</td>
<td>46</td>
<td>57.2</td>
<td>57.2</td>
<td>57.2</td>
<td>DRS</td>
</tr>
<tr>
<td>Northern Cape</td>
<td>0.1</td>
<td>0.78</td>
<td>0.66</td>
<td>477</td>
<td>147</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>...</td>
</tr>
<tr>
<td>North West</td>
<td>2.0</td>
<td>0.57</td>
<td>0.39</td>
<td>355</td>
<td>75</td>
<td>59.6</td>
<td>59.6</td>
<td>59.6</td>
<td>IRS</td>
</tr>
<tr>
<td>Western Cape</td>
<td>2.1</td>
<td>0.71</td>
<td>0.58</td>
<td>450</td>
<td>49</td>
<td>27.3</td>
<td>27.3</td>
<td>27.3</td>
<td>DRS</td>
</tr>
</tbody>
</table>

NB: The decile ratios should be interpreted as rand amounts.

Table 2 shows that the provinces that have Decreasing returns to Scale (DRS) are Eastern Cape, Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga and Western Cape. Free-State and Northern Cape have Increasing Returns to Scale (IRS). At the province level, IRS means that further investment in producing or transferring Gynaecologists from other provinces will increase productivity. DRS means that further investment in producing or transferring Gynaecologists from other provinces will not be productive. The only source of productivity will be to re-allocate resources within the provinces, depending on the level of needs reflected in the inequality indicators.

The trade-offs for adjusting to Gynaecology visit levels relative to the number of available Gynaecologists, provides the scope for reallocating resources from over resourced parts of provinces to underserved regions in provinces. The obvious re-allocations should be identified by the 20:80 decile ratios (table 2). The highest 20:80 decile ratios are for, Limpopo, Free State and Northern Cape. The lowest 20:80 decile ratios are for Mpumalanga, Gauteng and Eastern Cape (table 2).

In Limpopo R551 per capita is paid in the highest healthcare expenditure quintile for Gynaecology visits, relative to R1 in the lowest healthcare expenditure quintile (table 2). Because Limpopo’s combined trade-off adjustment of Gynaecology visits and number of available Gynaecologists is 19.4%, transfers within Limpopo might improve the returns to scale from being DRS to being IRS. This may also improve the Hoover Index which requires utilisation expenditure of 59% from the most endowed to the least endowed parts of Limpopo (table 2).
In Gauteng R119 per capita is paid in the highest healthcare expenditure quintile for Gynaecology visits, relative to R1 in the lowest healthcare expenditure quintile (table 2). Because the Gauteng's combined trade-off adjustment of Gynaecology visits and number of available Gynaecologists is 1%, a transfer within Gauteng might improve the returns to scale from being DRS to being IRS. A 1% transfer of Gynaecology visits in Gauteng to less endowed areas could approximate 1,600 visits per month. This may also improve the Hoover Index which requires utilisation expenditure of 58% from the most endowed to the least endowed parts of Gauteng (table 2).

3.4 Child Healthcare: Paediatricians

3.4.1 Efficiency analysis: Paediatricians

Figure 12 describes the efficiency of provinces at the scale at which they are operating. The x-axis shows the number of Paediatricians per province. The y-axis shows the scale efficiency (SE) derived from applying an input method to deriving CRS efficiency and VRS efficiency scores. The interpretation is that the pure efficiency achieved by providers decreases, the more the providers in a province. This can be seen in slope of the curve; however, the rate of decrease reduces at a decreasing rate from left to right (fig. 12). However, the greater the number of available providers, the greater the likelihood of residual efficiency (inefficiency).

Northern Cape has a scale efficiency of 1, which makes it the most productive province. Limpopo, Free State, Mpumalanga, North West and Eastern Cape cluster together at less than 50 Paediatricians per province. Followed by KwaZulu-Natal (less than 150 Paediatricians) and Western Cape (less than 250 Paediatricians. Gauteng has the highest number of Paediatricians (over 400 Paediatricians), but the lowest SE (fig. 12).

Figure 12: Scale Efficiency – Paediatricians (Dec. 2016)
Figure 13 describes the technical efficiency of Paediatricians across the nine South African provinces, using December 2016 data. The figure is generated by decomposing overall efficiency of Paediatricians per provinces, into their pure efficiency and residual efficiency components. Pure and residual efficiency collectively describe the nature of the technical efficiency that are present.

Pure efficiency is defined by the ratio between CRS and VRS scores. This ratio is called the Scale Efficiency (SE) score (see note 7.7 for derivation). “Residual efficiency is technical inefficiency that result from factors that are not efficiency related” (Industry Commission, 1997, p. 17). Therefore, it is “technical inefficiency” (Industry Commission, 1997, p. 17).

The x-axis in figure 13 is the SE score derived from comparing CRS and VRS scores using an input method of the DEA model. The y-axis in figure 13 is generated by decomposing actual visits paid into the pure and residual efficiency components. The y-axis reflects the residual efficiency (technical inefficiency) component of total number of Paediatric visits paid for. Therefore figure 13 shows the relationship between the two components describing the nature of technical efficiency for Paediatricians. There is a negative relationship between the pure efficiency and residual efficiency (technical inefficiency) (fig. 13).

Gauteng, and KwaZulu-Natal greatest residual efficiency (technical inefficiency), and least pure efficiency. The other provinces show residual efficiency that is lower than 50,000 visits per province. Northern Cape, North West, Free State, Eastern Cape and Limpopo have the highest pure efficiency (scale efficiency) (fig. 13).
Figure 14 is a web diagram that shows the relative relationships between overall efficiency, pure efficiency and residual efficiency. The overall efficiency is the CRS efficiency score. The pure efficiency represents the distance between the CRS and VRS frontiers. The residual efficiency represents the distance between the VRS-frontier and the actual level of efficiency generated by providers. For figure 14, pure and residual efficiency take on a number in the range from 0 to 1 ($d \in [0,1]$). They reflect the efficiency operating distance from the CRS and VRS efficiency frontiers (fig. 14).

The largest residual efficiencies (technical inefficiencies) are found on the left-hand side of figure 14. These provinces are Gauteng, KwaZulu-Natal, and Eastern Cape. The provinces that enjoy the largest pure efficiency from technical allocations from combining input and outputs are at the right-hand side of figure 14. These provinces are Limpopo, North West, Free State and Western Cape.

In Northern Cape, Paediatricians are relatively more efficient than other provinces, they perform 100% pure efficiency, and 0% residual efficiency (technical inefficiency). This is reflected by the zero distance in the fractions representing pure efficiency and residual efficiency (fig. 14).

In Gauteng, Paediatricians operate at 34.6% overall efficiency (CRS score). The distance between the CRS and VRS efficiency frontiers is zero, thus having a pure efficiency of 0 (fig. 14). They therefore operate at scale efficiency. That said, there are non-efficiencies that creep into this market. Residual efficiency (technical inefficiency) results in Paediatricians operating 0.65 away from the markets' VRS frontier (fig. 14).

In Free State, Paediatricians operate at 36.3% overall efficiency (CRS score). The distance between the CRS and VRS efficiency frontiers is 0.46. Thus; having a pure efficiency and scope for technical allocations to increase efficiency to SE level in certain parts of the province (fig. 14).
3.4.2 Equity, pareto optimality & trade-offs: Paediatricians

Table 3 describes: i) the number of Paediatricians per 10,000 beneficiaries (density ratio); ii) equity in accessing Paediatricians for all provinces (Gini Co-efficient (Gini), Hoover Index (HI) and decile ratios); and iii) the relative efficiency of Paediatricians for all provinces (efficiency score).

Table 3: Paediatricians -- equity indices and efficiency scores by province (2016)

<table>
<thead>
<tr>
<th>PROVINCE</th>
<th>PAEDIATRICIANS (per 10,000 beneficiaries)</th>
<th>GINI COEFFICIENT</th>
<th>HOOVER INDEX</th>
<th>DECILE RATIO (20:89)</th>
<th>DECILE RATIO (40:60)</th>
<th>CRS EFFICIENCY SCORE (%)</th>
<th>VRS EFFICIENCY SCORE (%)</th>
<th>SCALE EFFICIENCY SCORE (%)</th>
<th>TYPE OF SCALE EFFICIENCY</th>
<th>TRADE OFF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern Cape</td>
<td>0.4</td>
<td>0.79</td>
<td>0.73</td>
<td>7.20</td>
<td>299</td>
<td>65.3</td>
<td>100.0</td>
<td>65.3</td>
<td>IRS</td>
<td>17.2</td>
</tr>
<tr>
<td>Free State</td>
<td>1.0</td>
<td>0.76</td>
<td>0.59</td>
<td>1.368</td>
<td>556</td>
<td>36.3</td>
<td>53.6</td>
<td>67.6</td>
<td>IRS</td>
<td>12.2</td>
</tr>
<tr>
<td>Gauteng</td>
<td>1.2</td>
<td>0.74</td>
<td>0.88</td>
<td>3.43</td>
<td>33</td>
<td>34.6</td>
<td>100.0</td>
<td>34.6</td>
<td>IRS</td>
<td>1.2</td>
</tr>
<tr>
<td>KwaZulu-Natal</td>
<td>0.9</td>
<td>0.70</td>
<td>0.59</td>
<td>1.085</td>
<td>258</td>
<td>44.7</td>
<td>100.0</td>
<td>44.7</td>
<td>IRS</td>
<td>4.2</td>
</tr>
<tr>
<td>Limpopo</td>
<td>0.6</td>
<td>0.73</td>
<td>0.58</td>
<td>8.28</td>
<td>215</td>
<td>40.0</td>
<td>55.5</td>
<td>72.1</td>
<td>IRS</td>
<td>18.5</td>
</tr>
<tr>
<td>Mpumalanga</td>
<td>0.5</td>
<td>0.76</td>
<td>0.60</td>
<td>3.77</td>
<td>157</td>
<td>62.5</td>
<td>55.1</td>
<td>66.7</td>
<td>IRS</td>
<td>17.2</td>
</tr>
<tr>
<td>Northern Cape</td>
<td>0.1</td>
<td>0.76</td>
<td>0.62</td>
<td>3.47</td>
<td>157</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>IRS</td>
<td>0.0</td>
</tr>
<tr>
<td>North West</td>
<td>1.6</td>
<td>0.64</td>
<td>0.47</td>
<td>1.00</td>
<td>129</td>
<td>46.7</td>
<td>68.3</td>
<td>68.3</td>
<td>IRS</td>
<td>16.7</td>
</tr>
<tr>
<td>Western Cape</td>
<td>1.7</td>
<td>0.71</td>
<td>0.60</td>
<td>1.80</td>
<td>159</td>
<td>21.0</td>
<td>45.9</td>
<td>45.8</td>
<td>IRS</td>
<td>2.3</td>
</tr>
</tbody>
</table>

NB: The decile ratios should be interpreted as rand amounts.

Table 3 shows that the provinces that have Decreasing returns to Scale (DRS) are Gauteng, KwaZulu-Natal, Limpopo, Mpumalanga and Western Cape. Eastern Cape, Free State and Northern Cape, have Increasing Returns to Scale (IRS). At the province level, IRS means that further investment in producing or transferring Paediatricians from other provinces will increase productivity. DRS means that further investment in producing or transferring Paediatricians from other provinces will not be productive. The only source of productivity will be to re-allocate resources within the provinces, depending on the level of needs reflected in the inequality indicators.

The trade-offs for adjusting to Paediatricians visit levels relative to the number of available Paediatricians, provides the scope for reallocating resources from over resourced parts of provinces to underserved regions in provinces. The obvious re-allocations should be identified by the 20:80 decile ratios (table 3). The highest 20:80 decile ratios are for, Limpopo, KwaZulu-Natal, and Free State. The lowest 20:80 decile ratios are for Gauteng, Northern Cape, and Mpumalanga (table 3).

In Limpopo R826 per capita is paid in the highest healthcare expenditure quintile for Paediatric visits, relative to R1 in the lowest healthcare expenditure quintile (table 3). Because Limpopo’s combined trade-off adjustment of Gynaecology visits and number of available Gynaecologists is 18.5%, transfers within Limpopo might improve the returns to scale from being DRS to being IRS. This may also improve the Hoover Index which requires utilisation expenditure of 58% from the most endowed to the least endowed parts of Limpopo (table 3).
In Gauteng R343 per capita is paid in the highest healthcare expenditure quintile for Paediatrician visits, relative to R1 in the lowest healthcare expenditure quintile (table 3). Because the Gauteng’s combined trade-off adjustment of Paediatric visits and number of available Paediatricians is 1%, a transfer within Gauteng might improve the returns to scale from being DRS to being IRS. A 1.2% transfer of Paediatric visits in Gauteng to less endowed areas could approximate 8,600 visits per month. This may also improve the Hoover Index which requires utilisation expenditure of 68% from the most endowed to the least endowed parts of Gauteng (table 3).

3.5 Mental Healthcare: Psychiatrists

3.5.1 Efficiency analysis: Psychiatrists

Figure 15 describes the efficiency of provinces at the scale at which they are operating. The x-axis shows the number of Psychiatrists per province. The y-axis shows the scale efficiency (SE) derived from applying an input method to deriving CRS efficiency and VRS efficiency scores. The interpretation is that the pure efficiency achieved by providers decreases, the more the providers in a province. This can be seen in slope of the curve; however, the rate of decrease reduces at a decreasing rate from left to right (fig. 15). However, the greater the number of available providers, the greater the likelihood of residual efficiency (technical inefficiency).

Northern Cape has a scale efficiency of 1, which makes it the most productive province. Limpopo, Mpumalanga, North West and Eastern Cape cluster together at less than 50 Psychiatrists per province. Followed by Free State, KwaZulu-Natal. Gauteng (above 350 Psychiatrists) and Western Cape (just below 300 Psychiatrists), but the lowest SE (fig. 15).

Figure 15: Scale Efficiency – Psychiatrists (Dec. 2016)
Figure 16 describes the technical efficiency of Psychiatrists across the nine South African provinces, using December 2016 data. The figure is generated by decomposing overall efficiency of Psychiatrists per provinces, into their pure efficiency and residual efficiency components. Pure and residual efficiency collectively describe the nature of the technical efficiency that are present.

Pure efficiency is defined by the ratio between CRS and VRS scores. This ratio is called the Scale Efficiency (SE) score (see note 7.7 for derivation). "Residual efficiency is technical inefficiency that result from factors that are not efficiency related" (Industry Commission, 1997, p. 17). Therefore, it is "technical inefficiency" (Industry Commission, 1997, p. 17).

The x-axis in figure 16 is the SE score derived from comparing CRS and VRS scores using an input method of the DEA model. The y-axis in figure 16 is generated by decomposing actual visits paid into the pure and residual efficiency components. The y-axis reflects the residual efficiency (technical inefficiency) component of total number of Psychiatric visits paid for. Therefore figure 16 shows the relationship between the two components describing the nature of technical efficiency for Psychiatric. There is a negative relationship between the pure efficiency and residual efficiency (technical inefficiency) (fig. 16).

Gauteng has the highest residual efficiency (technical inefficiency), and least pure efficiency. The other provinces show residual efficiency that is lower than 150,000 visits per province. Northern Cape, Limpopo, Mpumalanga, Eastern Cape, North West, and Free State have the highest pure efficiency (scale efficiency) (fig. 16).
Figure 17 is a web diagram that shows the relative relationships between overall efficiency, pure efficiency and residual efficiency. The overall efficiency is the CRS efficiency score. The pure efficiency represents the distance between the CRS and VRS frontiers. The residual efficiency represents the distance between the VRS-frontier and the actual level of efficiency generated by providers. For figure 17, pure and residual efficiency take on a number in the range from 0 to 1 ($d \in [0,1]$). They reflect the efficiency operating distance from the CRS and VRS efficiency frontiers (fig. 17).

The largest residual efficiencies (technical inefficiencies) are found on the left-hand side of figure 17. These provinces are Gauteng, KwaZulu-Natal, and Eastern Cape. The provinces that enjoy the largest pure efficiency from technical allocations from combining input and outputs are at the right-hand side of figure 17. These provinces are Limpopo, North West, Free State and Western Cape.

In Northern Cape, Psychiatrists are relatively more efficient than other provinces, they perform 100% pure efficiency, and 0% residual efficiency (technical inefficiency). This is reflected by the zero distance in the fractions representing pure efficiency and residual efficiency (fig. 17).

In Gauteng, Psychiatrists operate at 19.4% overall efficiency (CRS score). The distance between the CRS and VRS efficiency frontiers is zero, thus having a pure efficiency of 0 (fig. 14). They therefore operate at scale efficiency. That said, there are non-efficiencies that creep into this market. Residual efficiency (technical inefficiency) results in Psychiatrists operating 0.81 away from the markets’ VRS frontier (fig. 17).

In Free State, Paediatricians operate at 15.4% overall efficiency (CRS score). The distance between the CRS and VRS efficiency frontiers is 0.49. Thus; having a pure efficiency and scope for technical allocations to increase efficiency to SE level in certain parts of the province (fig. 17).
3.5.2 Equity, pareto optimality & trade-offs: Psychiatrists

Table 4 describes: i) the number of Psychiatrists per 10,000 beneficiaries (density ratio); ii) equity in accessing Psychiatrists for all provinces (Gini Co-efficient (Gini), Hoover Index (HI) and decile ratios); and iii) the relative efficiency of Psychiatrists for all provinces (efficiency score).

Table 4: Psychiatrists -- equity indices and efficiency scores by province (2016)

<table>
<thead>
<tr>
<th>PROVINCE</th>
<th>Psychiatrists (per 10,000 beneficiaries)</th>
<th>Gini Co-efficient</th>
<th>Hoover Index</th>
<th>Decile Ratio (20:80)</th>
<th>Decile Ratio (40:60)</th>
<th>CRS Efficiency Score (%)</th>
<th>VRS Efficiency Score (%)</th>
<th>Scale Efficiency Score (%)</th>
<th>Type of Scale Efficiency</th>
<th>Trade-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eastern Cape</td>
<td>0.4</td>
<td>0.74</td>
<td>0.68</td>
<td>182</td>
<td>66</td>
<td>28.8</td>
<td>69.8</td>
<td>38.4</td>
<td>DRS</td>
<td>14.3</td>
</tr>
<tr>
<td>Free State</td>
<td>1.3</td>
<td>0.71</td>
<td>0.62</td>
<td>799</td>
<td>330</td>
<td>15.4</td>
<td>51.2</td>
<td>30.2</td>
<td>DRS</td>
<td>5.5</td>
</tr>
<tr>
<td>Gauteng</td>
<td>1.1</td>
<td>0.75</td>
<td>0.70</td>
<td>160</td>
<td>47</td>
<td>19.3</td>
<td>100.0</td>
<td>19.3</td>
<td>DRS</td>
<td>1.1</td>
</tr>
<tr>
<td>KwaZulu-Natal</td>
<td>0.9</td>
<td>0.67</td>
<td>0.56</td>
<td>408</td>
<td>136</td>
<td>24.0</td>
<td>100.0</td>
<td>24.0</td>
<td>DRS</td>
<td>3.4</td>
</tr>
<tr>
<td>Limpopo</td>
<td>0.5</td>
<td>0.72</td>
<td>0.56</td>
<td>113</td>
<td>74</td>
<td>19.5</td>
<td>21.1</td>
<td>92.4</td>
<td>DRS</td>
<td>21.1</td>
</tr>
<tr>
<td>Mpumalanga</td>
<td>0.2</td>
<td>0.76</td>
<td>0.66</td>
<td>21</td>
<td>13</td>
<td>58.6</td>
<td>100.0</td>
<td>58.6</td>
<td>error</td>
<td>44.4</td>
</tr>
<tr>
<td>Northern Cape</td>
<td>0.1</td>
<td>0.70</td>
<td>0.50</td>
<td>73</td>
<td>53</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>--</td>
<td>0.0</td>
</tr>
<tr>
<td>North West</td>
<td>1.4</td>
<td>0.71</td>
<td>0.62</td>
<td>51</td>
<td>18</td>
<td>33.3</td>
<td>96.1</td>
<td>34.6</td>
<td>error</td>
<td>error</td>
</tr>
<tr>
<td>Western Cape</td>
<td>2.1</td>
<td>0.69</td>
<td>0.58</td>
<td>200</td>
<td>41</td>
<td>19.9</td>
<td>47.3</td>
<td>23.1</td>
<td>DRS</td>
<td>error</td>
</tr>
</tbody>
</table>

NB: The decile ratios should be interpreted as rand amounts.

Table 4 shows that all the provinces that have Decreasing Returns to Scale (DRS) except Northern Cape, which has Increasing Returns to Scale (IRS). At the province level, IRS means that further investment in producing or transferring Psychiatrists from other provinces will increase productivity. DRS means that further investment in producing or transferring Psychiatrists from other provinces will not be productive. The only source of productivity will be to re-allocate resources within the provinces, depending on the level of needs reflected in the inequality indicators.

The trade-offs for adjusting to Psychiatrists visit levels relative to the number of available Psychiatrists, provides the scope for reallocating resources from over resourced parts of provinces to underserved regions in provinces. The obvious re-allocations should be identified by the 20:80 decile ratios (table 4). The highest 20:80 decile ratios are for Western Cape, KwaZulu-Natal, and Free State. The lowest 20:80 decile ratios are for Mpumalanga, North West, and Northern Cape (table 4).

In KwaZulu-Natal R408 per capita is paid in the highest healthcare expenditure quintile for Psychiatric visits, relative to R1 in the lowest healthcare expenditure quintile (table 4). Because Limpopo’s combined trade-off adjustment of Psychiatric visits and number of available Psychiatrists is 3.4%, transfers within KwaZulu-Natal might improve the returns to scale from being DRS to being IRS. A 3.4% transfer of Psychiatric visits in KwaZulu-Natal to less endowed areas could approximate 5,300 visits per month. This may also improve the Hoover Index which requires utilisation expenditure of 56% from the most endowed to the least endowed parts of KwaZulu-Natal (table 4).
In Gauteng R160 per capita is paid in the highest healthcare expenditure quintile for Psychiatric visits, relative to R1 in the lowest healthcare expenditure quintile (table 4). Because the Gauteng’s combined trade-off adjustment of Psychiatric visits and number of available Psychiatrists is 1.1%, a transfer within Gauteng might improve the returns to scale from being DRS to being IRS. A 1.1% transfer of Psychiatric visits in Gauteng to less endowed areas could approximate 4,300 visits per month. This may also improve the Hoover Index which requires utilisation expenditure of 70% from the most endowed to the least endowed parts of Gauteng (table 4).
4. FINDINGS & POLICY DISCUSSION

4.1 Emerging Policy Themes from Literature Review

4.1.1 The problem with geographic market delineation at tertiary healthcare level

The following significant observations can be drawn from the literature pertaining to the market delineation problem:

i. Patients/consumers within the same geographic market respond differently when demanding primary and secondary healthcare benefits, relative to, specialised surgical procedures.

ii. Consumers are passive participants of the primary healthcare market. This is the realm of MCOs and contracting provider groups.

iii. Hospitals have far longer referral destinations than primary care or secondary care level services. In certain instances; geographic markets for certain specialised services at the tertiary or quaternary level of healthcare, could be at national scale.

iv. Healthcare services that are not substitutes should not be clustered together within a single consumption bundle. This distorts market structure and makes it difficult to intervene where market forces create distortions. This could make it difficult to ensure access to effective cover of similar healthcare package in different geographic markets. It can make it difficult for consumers to respond to small but significant price increases, by seeking substitute healthcare bundles in neighbouring healthcare markets.

4.1.2 The new medical arms race hypothesis

The following significant observations can be drawn from the literature pertaining to the competitive strategies of market agents:

i. The nature of competition in the primary healthcare delivery market has shifted from only being determined by the contracting relationship between MCOs and provider groups.

ii. Strategic competition now includes a strong presence in attracting secondary healthcare resources and state of purchasing state of the art healthcare equipment.

iii. This phenomenon is what is called the ‘new medical arms race’ in healthcare markets.
iv. Although price competition does exist in concentrated healthcare environments, the 'new medical arms race' may create residual efficiencies in healthcare delivery.

v. This recent development requires a good understanding of the forces that inform the competitive drive of healthcare markets at the geographical level.

vi. This is important, otherwise, supporting regulatory interventions for redistributing healthcare resources across geographic markets will not be effective. Rendering attempts at making healthcare delivery more equal, ineffective.

4.1.3 Co-dependence between efficiency and equality

The following significant observations can be drawn from the literature pertaining to the co-dependence between efficiency analysis and improving equal access to healthcare:

i. To achieve equitable healthcare cover, requires an assessment of the efficiency at which healthcare services are produced at different geographic locations.

ii. If healthcare resources are invested in geographic markets with decreasing returns to scale, waste and inefficiency are likely to occur.

iii. Resources in geographic markets with decreasing returns to scale could be freed up and redistributed to geographic market with deficiencies in healthcare resources.

iv. Efficient healthcare service delivery enhances the accessibility healthcare benefits. Effective cover improves the depth/enrichment of benefits in medical schemes and allows large risk pools to have access to healthcare services on an equitable basis.
4.2 Significant Findings from Empirical Analysis

The distribution of medical schemes beneficiaries, and healthcare providers; seem to be positively related to patterns of economic performance. The three provinces with the largest enrolment of beneficiaries are Gauteng, Western Cape and KwaZulu-Natal.

Because the data could be describing patterns of economics of conurbation, rather than healthcare demand and delivery characteristics. Therefore, there is validates the need for implementing the Variable Returns to Scale (VRS). This should isolate exogenous factors that associated with conurbation effects. This is extremely important as others have used conurbation economics as a basis of righting off any analysis on supply-induced demand in a spatial context (Newhouse, et al., 1982).

Here again, there is a benefit in using DEA VRS models, over market delineation/relevant market models, for describing the performance of geographic markets. The reason is that VRS methods can compare and benchmark geographic markets within their appropriate peer groups (Industry Commission, 1997).

There is an imbalance in the availability of healthcare providers between the public and private healthcare sector. This makes it more important to re-allocate healthcare resources with diminishing returns to locations, and sectors of the healthcare market where they will be more productive. This is particularly relevant for identifying disciplines that should be prioritised for rolling out public-private partnerships (PPPs), in the view of preparing the healthcare market for NHI.

The efficiency analysis shows that although relatively larger geographic healthcare markets demonstrate scale efficiency, in many instances residual efficiency (technical inefficiency) creeps into the operations of the markets. There have also been Increasing Returns to Scale (IRS). That said; IRS have been in existence in the relatively smaller geographic markets. It is important to mention at this point, that there are benefits in investing more human resources for healthcare in these markets. This is because, exogenous intervention will increase needed healthcare access. This should have a formidable effect on reducing healthcare inequality indicators.

The trade-offs are significant because it gives an indication of what can be done within the ambit of current Medical Scheme Act. More specifically, through the registration of provider networks in scheme rules or accreditation of integrated healthcare and Managed Care Organisations (MCOs).

The same redistribution effects can also be achieved through incentivising re-allocation through supporting mechanisms that have been identified rolling-out NHI. An example would the Virtual Risk Pooling Framework. The risk adjustment financial flows could also allow for income-cross subsidization.
The income redistribution could also correct market structure issues. Market structure issues have also been identified as one of the causes of healthcare utilisation costs (van den Ven & Ellis, 2000).

4.3 Policy Discussion

Optimisation of consumer healthcare utility through a social equilibrium as opposed to a competitive equilibrium, when the market cannot reallocate resources where the demand is greatest, may increase healthcare outcomes and net social benefit. The coordination of the optimal allocation is by its nature a collaborative equilibrium. Therefore, the policy needs to be perceived as being credible when consensus is sought from market agents, who will need to commit to it.

There are opportunities to reduce the imbalance in healthcare provider distribution across geographic healthcare markets within the ambit of the current Medical Schemes Act. This could be done by identifying the pockets within provinces where residual efficiency (technical inefficiencies) exist. Petitions to grow operations in these markets should not be entertained by the regulator. Rather, the regulator should focus on the areas where there Increasing Returns to Scale (IRS) and pure efficiencies (technical efficiencies). Applications for new MCOs and network arrangements should be prioritised for these specific areas.

This is also important for the strategic objectives of implementing Efficiency Discount Options (EDOs), and other variations of network options, to make low cost networks available across geographic markets. If services are obtained in network areas with high residual efficiency (technical inefficiencies), then the policy goal will not be met or poorly implemented. Another concern is that these EDOs might highly associated with market environments that operate competitive strategies informed by the “new health arms race”.

The roll-out of NHI makes provision for a Virtual Risk Pool Fund (VRPF). Risk equalization is about monetary transfers between medical schemes, for more socially optimal equity outcomes. The function of risk equalisation funds is to achieve horizontal and vertical justice. By way of subsidising: i) equal health for equal health needs (horizontal equity); ii) income cross-subsidisation (vertical equity).

The VRPF could be used to support income cross-subsidies, by making it easier for healthcare consumers in underserved geographic regions to demand healthcare services. Relocation subsidies to incentivise healthcare providers to operate in underserved areas will reduce disparities in healthcare utilisation inequality indices (see tables 1-4 in the results section). That said; risk equalisation schemes are prone to being abused. Re-imbursement subsidies for the relocation of healthcare providers, should be implemented with caution. VRPF money flows that are to incentivise needed healthcare resources to relocate to underserved geographic healthcare markets, should not
create perverse outcomes. The sustainability of such system would require a reimbursement system that induces the risk preferences of scheme outsourced service providers, and contracted healthcare providers (The Royal Swedish Academy of Sciences, 2014).

This means that re-imbursement contracts will have to be based on a god knowledge about the true efficiency of healthcare providers (Smith, 1976) and (Coase, 1960). Specifically, the healthcare providers who are being subsidised to relocate. The reason is that, fixed-price contracts and cost-reimbursement introduce adverse incentives, that increase the social cost of interventions such as risk equalization funds. For example, efficient healthcare providers may pretend to be inefficient. Thus; demanding re-imbursement on a cost-reimbursement basis, which provides payoffs that are far above the healthcare provider's true costs. These profits result in social costs that are higher than the benefits of equalizing risk profiles across schemes (The Royal Swedish Academy of Sciences, 2014).

If risk-equalization is implemented to improve market competition, it should be enhanced by other interventions to enable socially optimal outcomes. It is important that the appropriate types of contractual structures are used for service providers of differing scale efficiencies.
5. RECOMMENDATIONS

5.1 Redistribution of Healthcare Providers

i. *The issue of resolving unequal access to healthcare:*

vi. The issue should be resolved through regulatory coordinating actions. The regular operation of the market cannot resolve the unequal distribution of healthcare providers.

vii. In resolving issues on the technical allocation of healthcare resources, Data Envelopment Analysis (DEA) models should be used in evaluating the economic interaction of market agents and resulting healthcare market prices. Relevant market identification models measure market concentration, but the market forces nor market agent interactions behind high utilisation costs or efficiencies in healthcare production. These are merely assumed from the concentration indices. An economic evaluation model based on DEA analysis has been used to decompose technical efficiency into its pure efficiency and residual efficiency (technical efficiency components).

viii. DEA models should identify areas where there are technical inefficiencies in healthcare delivery and redirect these resources where there are positive returns to scale.

ix. The Virtual Risk Pooling Fund could be used to incentivise needed healthcare resources to relocate to underserved areas. This could also act as an indirect income cross-subsidization mechanism for deprived communities.

x. Re-imbursement contracts for healthcare providers who are subsidized for relocating, should be based on a good knowledge of the providers true efficiency. Providers could pretend to be less efficient than what they are, and benefit from fixed price contract arrangements like highly priced capitation agreements.

ii. *The issue of non-existing markets creating market failure in certain geographic markets:*

iv. The Health Market Inquiry (HMI) has been receiving recommendations on how to resolve the issues of arm in the private healthcare market.
v. The lack of market access, or lack of an existing market, in certain areas is a market failure. This is because the normal operation of competition has not allocated resources where there is demand, for lack of incentives.

vi. The analysis findings and methods in this document could assist in making HMI recommendations where relevant market methodology has not assisted.

iii. The issue of impediments to efficient operation of competition in geographic markets, as an "issue of harm":

v. In some instances, merger evaluations in healthcare markets have under-estimated the anticompetitive effects of price increases. This very particular to healthcare market.

vi. This is because market price mechanisms have been very difficult assess with market delineation methods.

vii. Understanding the nature of existing technical efficiencies in a geographical healthcare market, provides a useful inference on a driving force average costs resulting from healthcare facility production functions.

viii. The DEA approach to economic evaluation may help in monitoring market behaviour.

iv. The issue of where to intervene and how to intervene through provider networks and integrated healthcare delivery models:

iii. Integrated healthcare networks and networks in benefit designs should prioritise underserved areas

iv. Integrated healthcare networks and networks in benefit design should not be allowed in geographic markets with residual efficiencies (technical inefficiency). If Efficiency Discount Option (EDO) and low-cost integrated care networks are registered in these areas. Policy objectives are not met.
5.2 Considerations for a Mixed Bag of Policy Interventions

5.2.1 The Market Delineation Issue

i. The issue on whether to accredit healthcare facilities based on geographic markets or a certificate of needs:

- It may be easier to delineate and standardise product markets for primary and secondary healthcare than it may be for tertiary and quaternary services. As the latter may prove to be more complex product offerings.

- It may be difficult to delineate markets for certain complex healthcare interventions by geographic, regional or district level. Market forces will respond to factors that are external to markets that are delineated at sub-national or provincial level. Therefore, the policy intervention or regulated equilibrium will not hold.

ii. The issue on tariff regulation or bargaining councils/chambers:

- Standardising benefit designs could go a long way to making contracting more transparent in bargaining chambers. Consumer groups could ensure that consumers are protected as they are passive participants in the primary healthcare market.

- Standardised benefit options could remove the complexity of delineating geographic healthcare markets.

5.2.2 The new medical arms race issue

The issue on resolving ‘issues of harm’ that create high costs in geographic healthcare markets:

- Due to potential over servicing in resource rich geographical markets, and underservicing in resource deprived geographic markets, market forces are not able rein in on waste and inefficiency.

- Healthcare resource redistribution initiatives need to be based on distinguishing between pure efficiency and residual efficiency between peers across geographic healthcare markets.

- This is important as market delineation models and merger guidelines in other jurisdictions have not been able to manage the cause of small but significant price increases in healthcare markets. These models have only determined whether these price increases can be abated by substitute markets, and if not, have rejected mergers in geographic healthcare markets.
• Envelopment analysis, a technique that has been used in this analysis, are able to decompose overall efficiencies into pure and residual efficiency components. This decomposition informs efforts that are exploited in markets with diminishing returns to scale, for the betterment of equal access to healthcare resources in underserved geographic markets.
6. LIST OF REFERENCES


HMI, 2017. *Invitation to participate in a seminar on tariff determination*, Pretoria: Competition Commission South Africa.


7. TECHNICAL NOTE

7.1 Gini Co-efficient

The Gini-coefficient (Gini) describes the level of inequality in terms of GP access for medical scheme beneficiaries. The Gini is scores over a range of 0 to 1, it can also be reflected as a percentage. The higher the Gini, the higher the level of inequality in accessing GPs.

![Figure 18: Illustration of elements used for calculating the Hoover Index.](image)

$Gini \ Coefficient = \frac{A}{A + B}$

$A + B = 0.5$

$A = 0.5 - B$

$B = \frac{1}{2} \sum_{i=1}^{n} (x_i - x_{i-1})(y_i + y_{i-1})$

7.2 Hoover Index

The Hoover Index (HI) represents the proportion of redistribution required to make GP access equal for all beneficiaries in a province.
Figure 19: Illustration of elements used for calculating the Hoover Index.

\[
Hoover \ Index = \frac{1}{2} \sum_{i=1}^{N} \left| \frac{E_i}{E_{total}} - \frac{A_i}{A_{total}} \right|
\]

7.3 Decile Ratios

Figure 20: Illustration of elements used for calculating the Decile Ratios.

\[
Decile \ Ratio \ 80:20 = \frac{\mu_{80}}{\mu_{20}}
\]

\[
Decile \ Ratio \ 40:60 = \frac{\mu_{40}}{\mu_{40}}
\]
7.4 Efficiency Frontiers

![Figure 21: Production frontiers, efficiency and pareto allocation](image)

*Note: Adapted from (Industry Commission, 1997, p. 7)*

7.5 Constant Returns to Scale Method

The CRS Efficiency Score (ES), is a relative score of efficiency of GPs in producing healthcare visits that are paid by medical schemes. The CRS ES is a relative score of efficiency. This is because it compares provinces with each other, as though they have similar characteristics and have the same number of GPs. The higher the CRS ES score the higher the input-output productivity of a province, relative to, other provinces.

**Data Envelopment optimization problem equations (Input Method)**

\[
\min E_n = \frac{\sum_{i=1}^{N} \omega_i y_i}{\sum_{i=1}^{N} \omega_i x_i}
\]

Subject to:

\[
\sum_{i=1}^{N} \omega_i y_i - y^*_n \geq 0
\]

\[
\sum_{i=1}^{N} \omega_i x_i - E^*_n x^*_n \leq 0
\]

\[
i = 1, \ldots, N
\]

\[
\omega_i \geq 0
\]
Where:

\[ \omega_i = \text{efficiency weight for province } i \]

\[ E_n^* = \text{CRS efficiency score for the i' th province} \]

\[ y_i = \text{number of healthcare provider visits paid for in province } i \]

\[ x_i = \text{number of healthcare providers in the province } i \]

\[ y_n^* = \text{number of healthcare provider visits paid for in the n' th province} \]

\[ x_n^* = \text{number of healthcare providers in the n' th province} \]

\[ I = \{ i \in I \mid \text{is the set of all provinces} \} \]

Implementation of optimization model using Excel Solver

Figure 22: Implementation of CRS model in excel solver (input method)
7.6 Variable Returns to Scale Method

Variable Returns to Scale (VRS) efficiency level

The VRS Efficiency Score (ES), adjusts for differences in the availability of GPs across provinces. Thus, VRS ES assesses a province’s efficiency in terms of its own scale size. The higher the CRS ES score the higher the input-output productivity of a province, relative to, other provinces.

Data Envelopment optimization problem equations (Input Method)

\[
\min S_n = \frac{\sum_{i=1}^{N} \omega_i y_i}{\sum_{i=1}^{N} \omega_i x_i}
\]

Subject to:

\[
\sum_{i=1}^{N} \omega_i y_i - y_n^* \geq 0
\]

\[
\sum_{i=1}^{N} \omega_i x_i - S_n^* x_n^* \leq 0
\]

\[
\sum_{i=1}^{N} \omega_i = 1
\]

\[
i = 1, ..., N
\]

\[
\omega_i \geq 0
\]

Where:

\(\omega_i = \text{efficiency weight for province } i\)

\(S_n^* = \text{VRS efficiency score for the } i^{th} \text{ province}\)

\(y_i = \text{number of healthcare provider visits paid for in province } i\)

\(x_i = \text{number of healthcare providers in the province } i\)

\(y_n^* = \text{number of healthcare provider visits paid for in the } n^{th} \text{ province}\)

\(x_n^* = \text{number of healthcare providers in the } n^{th} \text{ province}\)

\(I = \{i \in I | is \ the \ set \ of \ all \ provinces \}\)
Implementation of optimization model using Excel Solver

7.7 Scale Efficiency

The Scale Efficiency (SE) efficiency score is the component of total efficiency that can be explained operational scale.

\[ SE = \frac{E_{n}^*}{S_{n}^*} \]

Where:

- \( SE = \text{Scale Efficiency} \)
- \( E_{n}^* = \text{CRS efficiency score for the } l^{th} \text{ province} \)
- \( S_{n}^* = \text{VRS efficiency score for the } i^{th} \text{ province} \)
7.8 Non-increasing Returns to Scale

Type of scale efficiency describes whether the province is operating at increasing or decreasing returns to scale. If a province is operating at increasing returns to scale, increasing the province's resource capacity should will benefit scheme beneficiaries.

**Data Envelopment optimization problem equations (Input Method)**

\[
\min R_n = \frac{\sum_{i=1}^{N} \omega_i y_i}{\sum_{i=1}^{N} \omega_i x_i}
\]

**Subject to:**

\[
\sum_{i=1}^{N} \omega_i y_i - y_n^* \geq 0
\]

\[
\sum_{i=1}^{N} \omega_i x_i - R_n^* x_n^* \leq 0
\]

\[
\sum_{i=1}^{N} \omega_i \leq 1
\]

\(i = 1, \ldots, N\)

\(\omega_i \geq 0\)

**Where:**

\(\omega_i = \text{efficiency weight for province } i\)

\(R_n^* = \text{efficiency score for the } i^{th} \text{ province}\)

\(y_i = \text{number of healthcare provider visits paid for in province } i\)

\(x_i = \text{number of healthcare providers in the province } i\)

\(y_n^* = \text{number of healthcare provider visits paid for in the } n^{th} \text{ province}\)

\(x_n^* = \text{number of healthcare providers in the } n^{th} \text{ province}\)

\(I = \{i \in I | \text{is the set of all provinces}\}\)
7.9 Increasing & Decreasing Returns to Scale

Returns to Scale \(= \frac{S^*_n}{R^*_n} \)

If:

\[ \frac{S^*_n}{R^*_n} > 1 \equiv IRS \]

\[ \frac{S^*_n}{R^*_n} < 1 \equiv DRS \]

Where:

IRS = Increasing Returns to Scale

DRS = Decreasing Returns to Scale
7.10 Trade-Offs

Inequality indicators are used as equity measures; i.e. the Gini-coefficient (Gini). The trade-offs are adjustments to resource allocation (GPs) and outcomes (GP visits), to reach an optimal allocation of resources without reducing GP visits to a sub-optimal level.

These trade-offs result in the “first-best” solution for resource allocation (Pareto optimality). In adjusting to the Pareto solution, the Gini-Coefficient Index (Gini) starts improving (CIHI, 2016: 15). Equity should improve, meaning that resources should become more accessible within a province.

Data Envelopment optimization problem equations (Additive Method)

\[
\text{minmax } z = \sum_{i=1}^{N} y_i^+ s_i^+ + \sum_{i=1}^{N} x_i^- s_i^-
\]

Subject to:

\[
\sum_{i=1}^{N} \omega_i y_i - s_i^+ \geq S_n^* y_i \\
\sum_{i=1}^{N} \omega_i x_i + s_i^- \leq S_n^* x_i^*
\]

\[
\sum_{i=1}^{N} \omega_i = 1 \
\]

\[
i = 1, ..., N \
\omega_i \geq 0 \
s_i^+, s_i^- \geq 0
\]

Where:

\[
\omega_i = \text{efficiency weight for province } i \\
S_n^* = \text{efficiency score for the } i^{th} \text{ province} \\
S_n^* = \text{number of healthcare provider visits paid for in province } i \\
x_i = \text{number of healthcare providers in the province } i \\
l = \{i \in l | \text{is the set of all provinces}\} \\
s_i^+, s_i^- = \text{positive and negative slack}
\]
Implementation of optimization model using Excel Solver

Figure 25: Implementation of Mini-max model in excel solver (additive method)