A Model for Performance Evaluation of Emergency Department Physicians

By

Javier Fiallos

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfilment of the requirements for the degree of
Master of Science in Health Systems

uOttawa

The University of Ottawa

Abstract

Performance of Emergency Department (ED) physicians (MDs) is multi-faceted since it impacts multiple dimensions such as health outcomes of patients, utilization of resources, throughput of patients and timeliness of care. Therefore, the assessment of their performance demands the use of a tool that allows considering multiple evaluation criteria. However, commonly used multi-criteria evaluation methods often require assigning weights to dimensions in order to define their relative importance on a final performance score. This feature introduces subjectivity in the development of weights and has the potential to produce biased results.

The purpose of this thesis research is to develop a multi-dimensional evaluation tool for evaluating performance of ED MDs. The proposed evaluation tool relies on a mathematical programming model known as Data Envelopment Analysis (DEA). The use of DEA does not ask for subjective weighting assignments for each dimension that describe the ED MDs' performance. It is capable of considering multiple heterogeneous performance measures to identify benchmark practice and the individual improvements leading to best practice of each evaluated unit.

The DEA model described here was developed from real data to assess the performance of 20 PED MDs from the Children's Hospital of Eastern Ontario (CHEO).

Multiple evaluations were run on stratified data in order to identify benchmark practice in each of seven categories of patients' complaints and to determine the impact of accompanying MD trainees on PED MDs' performance.

For each PED MD, performance scores and improvements in each category of patients' complaints (i.e. respiratory, trauma, abdominal, fever, gastroenterology, allergy and Ear-Nose-Throat complaints) were determined. This helped identifying the required improvements that would lead PED MDs to achieve benchmark performance.

Regarding the influence of MD trainees on PED MDs' performance, results show that most PED MDs (15 out of 20) perform better when they are not accompanied by a trainee which motivates further research to assess trade-offs between teaching and clinical performance.

In summary, DEA proved to be an appropriate tool for performance evaluation of PED MDs because it helped to identify benchmark performers and provided information for performance improvements under a multi dimensional performance evaluation framework.

Acknowledgements

I would like to thank Professor Wojtek Michalowski for encouraging me to pursue this research topic. His contributions to the definition of research objectives, his close supervision and pertinent feedback have helped me to make of my research an enjoyable learning process.

I would like to express my gratitude to Professor Jonathan Patrick whose supervision, expertise and keen insight have inspired me to uncover the best of my analytical capabilities.

This research would not have been possible without the guidance and contextualization provided by Dr. Ken Farion. I owe him my deepest gratitude for contributing his medical and managerial expertise to the development of this research and for facilitating the collection of data.

Finally, I would like to thank my wife Orietta and my children Julián, Pablo and Emilia and my mother Maribel, for all their support and inspiration.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Motivation

The performance of an Emergency Department (ED) is of importance to patients, health system authorities, hospitals and physicians (MDs). An ED's effectiveness is related to patient outcomes (Mulley, 1990) and its efficiency impacts the budgeting decisions of hospital management since for a given throughput level of patients, prudent utilization of resources contribute to reduce operational costs. As the ED is often the entry point to the healthcare system, their performance is highly scrutinized. This has motivated several research initiatives primarily focused on improving patient flow and quality of care (Eitel et al., 2010), (Lowthian et al., 2012). Most of this research aims to establish efficient workflow processes, appropriate staffing and enhanced teamwork operations but does not assess the impact of the performance of each provider (MDs and nurses) on the ED's overall performance. This is a significant omission because almost all of the measures that describe the performance of a clinical unit such as an ED are controlled to a large extent by the performance of the MDs.

This research aims to fill this gap by developing a multi-criteria model to assess the performance of MDs working in an ED. We use data from a Pediatric Emergency Department (PED) to assess the appropriateness of our performance evaluation model.

1.2 Performance of ED MDs

Patients arrive to an ED with different medical complaints, some of which require more urgent care than others. To determine the priority of care required by each patient, a triage nurse makes a preliminary assessment to determine the level of severity of the patient's health status. For each patient, a CTAS (Canadian Triage and Acuity Scale) score and a CEDIS (Canadian Emergency Department Information System) label is assigned. The CTAS value provides a grading of the severity on a scale of 1 to 5, where 1 represents the highest severity, while the CEDIS label provides a short description of the presenting complaint of the patient.

After registration and triage, the patient is placed in the prioritized queue for an ED MD assessment. Typically, in the course of this process, an ED MD interviews a patient, assesses the patient's history, performs a physical examination, and if necessary orders laboratory or imaging tests prior to arriving at a diagnosis and identifying therapy. At the end of the process, the ED MD makes a disposition decision (admit or discharge).

ED MDs' actions and decisions impact the quality of a patient's experience and his/her health outcomes. Although, un-controllable factors such as timeliness of diagnostic tests' results, nursing staff fluctuations, shift handovers, and ED load influence the experience and health outcome of each patient, the clinical competence of the ED MD certainly plays an important role in the quality of care delivered. Generally, highly competent ED MDs assess patients correctly in shorter time and have a higher number of patients with positive health outcomes.

ED MDs' practice also has an impact on the utilization of resources. Highly competent ED MDs will order laboratory and imaging tests only when clinically

necessary. He/she will also make better use of time producing a higher throughput of patients and thus shorter wait times in the ED and possibly also shorter lengths of stay (LOS).

Considering all the above, it is natural to conclude that an ED MDs' performance is multi-faceted and its assessment requires the consideration of a number of heterogeneous factors (Dubinsky, 2010). In the case of an ED, the best performing ED MD will be one who *provides timely and high quality care to patients while utilizing the available resources as effectively as possible*.

Often the goals of using resources efficiently and providing quality care appear to be in conflict. For example, as an ED MD works toward arriving at a proper diagnosis by ordering additional tests, the patient's stay is prolonged, the timeliness of care is affected and the use of resources (laboratory tests, x-ray orders, imaging orders, etc) is increased.

These trade-offs highlight the need for a multi-criteria approach to the evaluation of a ED MD's performance. Limiting an assessment to a single criterion skews performance towards the chosen measure; quite often at the expense of others. For example, sole focus on reducing the rate of returns to the ED (commonly used as a measure of quality of care) might motivate ED MDs to over-treat (and over-admit) patients resulting in a higher cost per patient and a reduced throughput in the ED.

Thus, a good performance evaluation model should not only account for multiple dimensions, it should also reward ED MDs that perform well according to most measures more than it would reward ED MDs who show exceptionally good performance in a single measure but are poor in others. If an ED MD showing this kind of compensatory behavior is rewarded, he/she will not be motivated to improve his/her weaknesses.

Therefore, compensation of poor performance by attaining excellent performance in a single measure (criterion) should be discouraged. Most importantly, an evaluation model should properly discriminate between ED MDs' performances. Insufficient discrimination will not provide enough information about improvement opportunities nor a clear reference for benchmark practice.

In the context of ED MDs, defining the importance of each evaluation measure relative to the rest of the measures is a subjective assessment. Most multi-criteria evaluations found in the literature focus on developing composite scores, which require assigning a weight to each evaluation criterion to reflect its importance in a global score. Such an assignment of weights involves subjective judgments and introduces potential bias in the evaluations. Therefore, a proper performance evaluation tool should avoid any subjective definition of each measure's relative importance.

Another aspect influencing ED MDs' performance is the wide range of medical conditions treated in the ED. Since ED MDs manage different complaints by using different resources (e.g. bone fractures require X-ray orders while allergic reactions do not), performance measures related to resources used by the same ED MD, will differ considerably across the complaints. Also, depending on the work shifts patterns, different MDs working in the same ED may treat a different case mix of patients. These two aspects, complaint variation of visits and heterogeneous case mix potentially introduce bias into evaluation results if a global score is used to capture the overall performance of ED MDs.

When comparing MDs managing distinctly different case mixes using data that does not differentiate among presenting complaints, it becomes impossible to tell if the

differences in performance between two ED MDs is caused, for example, by one ED MD assessing a higher percentage of visits that demand specific resources and not due to different patterns of practice.

Another disadvantage of a global approach is the possible averaging effect of overall evaluations. Even if there were no differences in case mix across ED MDs', the performance of a particular ED MD might be the result of a levelling of high and low performances across presenting complaints.

These issues demand the development of a stratified performance evaluation model in which performance under each complaint or complaint category is measured separately, minimizing the negative effects of uneven case mixes of ED MDs and practice variation across complaints. Furthermore, a key advantage of a stratified approach is the possibility of providing focused feedback for each ED MD that will better describe actions to be taken in order to achieve benchmark performance in each complaint category.

In summary, performance of ED MDs must be assessed using a model that takes into account multiple performance measures (timeliness and effectiveness of care, utilization of resources and patient throughput), avoids the subjective establishment of the relative importance of these measures, penalizes compensatory behaviour and provides enough discrimination of performances in order to identify benchmark practice and improvement opportunities.

1.3 Contribution

Developing a multi-criteria evaluation framework presents a considerable challenge when selecting the type and number of performance measures. Weights-based methods

have the additional challenge of assigning a weight to each measure in order to capture its relative importance. Some methodologies assume equal weights while others assign different weights in an attempt to achieve a higher impact on a final score from the measures deemed most important. The main shortcoming of any weighting scheme is the subjectivity involved in the weight development process (Moers, 2004).

Other approaches, focused on determining a ranking of evaluated individuals do not require the assignment of weights but still require the user to define some performance dimensions as being more important than others (Saaty, 1994), (Cherchye and Vermeulen, 2006), (Van Ours and Vermeulen, 2006). In addition to the subjectivity in judging the importance of each measure, such methods do provide guidance as to how each individual's performance should be modified in order to achieve benchmark performance.

To the best of our knowledge, there is no research that has addressed the problem of MDs' assessment that simultaneously takes into account the measures of timeliness of care, effectiveness of care, resource utilization and throughput of patients. This thesis research aims to fill this void by applying a methodology that allows for the multi-dimensional assessment of the performance of this specific group of MDs - ED MDs, by capturing the quality and effectiveness of their services, the efficiency involved in the provision of these services and determining individual performance improvements to meet benchmark practice. The proposed approach relies on Data Envelopment Analysis (DEA) (Charnes et al., 1978), which is a mathematical programming methodology that produces a measure of relative efficiency (measure of proximity to benchmark practice) for each individual being evaluated (referred to as a Decision Making Unit (DMU))

(Cooper et al., 2007). The main strength of DEA is its ability to facilitate multi-criteria evaluations while avoiding the subjectivity involved in the assignment of weights to performance measures as well as its ability to identify the required improvements for each DMU to achieve benchmark performance.

In addition, this research presents a novel extension of a DEA model that mitigates shortcomings related to the low discriminatory power of DEA when the sample of evaluated individuals is small.

1.4 Research Question and Objectives

This research seeks to determine if it is possible to derive a realistic model for assessing an ED MD's performance based on multiple measures.

Answering this research question involves accomplishing the following objectives:

- The development of a DEA model that helps assess ED MDs' performance that meets the criteria mentioned earlier of a good assessment tool.
- Solving the model using data from the ED of the Children's Hospital of Eastern Ontario (CHEO).
- Evaluation of the results.

Chapter 2: Related Literature

This chapter presents a review of literature related to the general topics of the performance evaluation of MDs, main DEA modelling concepts and related DEA applications.

2.1 Performance evaluation of MDs

2.1.1 Evaluation criteria

The evaluation of MDs' performance reported in the literature focus primarily on assessing the quality of care in relation to selected therapies where measures for specific patient populations are examined. Glickman et al. (2008) discuss criteria to support the selection of clinical measures to evaluate quality of care provided by ED MDs. These measures include use of β-blocker therapy for acute myocardial infarction at the arrival to the ED and performing a diagnostic ECG for syncope in patients older than 60 years. Hess et al. (2010) proposes using performance measures such as completion of retinal and foot exams, blood pressure levels, etc in order to assess quality of care provided to diabetic patients. Pure clinical performance measures are rather disease-specific and as such are difficult to apply in the ED context.

General evaluations deal with the measures required to jointly assess clinical and behavioral competencies of the MDs. Evaluation of clinical competency usually involves such measures as the selection of appropriate diagnostic tests and treatment. It is assumed

that these measures represent a valid proxy for performance evaluation. Behavioral competency refers to humanistic aspects, patient and coworker communication and social skills (Dubinsky et al., 2010). The assessment of these competencies can be performed through various methods such as questionnaires, peer reviews (Smith et al., 2004), (Hall et al., 1999), or patient chart audits (Goulet et al., 2002). Most of these evaluations deal with qualitative constructs that are difficult to measure. In addition, the common practice of obtaining qualitative data through self-assessment questionnaires and peer reviews might introduce bias in the evaluation. More importantly, there is a belief that competency is not fully related to performance of MDs (Rethans et al., 1991). The same authors maintain that competency refers to a capability to do work, whereas performance relates to the actual doing of the work, so these should be considered as different constructs.

There are a number of publications where MDs' performance is assessed based on the effective operation of a clinical unit. These assessments are focused on measuring performance from a cost containment perspective. Ozcan (1998) used measures such as the number of visits to primary care MD, number of prescriptions, laboratory procedures, etc. when assessing the performance of 160 primary care MDs treating otitis media. The results allowed them to propose a model of practice that promotes savings in the cost of treatment while maintaining the same quality of care. Chilingerian and Sherman (1996) evaluated the performance of 326 primary care MDs using measures such as office visits, specialist referrals, diagnostic tests, and visits to the ED to capture the level of resource utilization. Their research highlights the savings that could be derived by adopting benchmark practices and suggests a reassessment of the pay for performance care

delivery model. These cost containment evaluations do not apply to the ED context because they assume cost differentiation as the main factor in a situation characterized by uniform health outcomes, whereas the purpose of ED MD evaluation is mainly to assess differences in the quality of care provided.

2.1.2 Evaluation methods

Most evaluation methods construct a global score for each MD by calculating the average of questionnaires' responses using a Likert scale (Smith et al., 2004) (Hall et al., 1999). The scores can also be constructed from the counts of deviations from threshold values (Weber et al., 2010) or categorizing Likert scale responses into satisfactory and unsatisfactory (Goulet et al., 2002). In some cases, to develop a global measure, composite scores are created by assigning weights to each performance measure to capture its level of importance (Hess et al., 2010). This can be problematic due to the risk of bias involved in the weight assignment (Moers, 2004). Moreover, it cannot be used when an evaluation involves the use of heterogeneous measures due to the impossibility of combining within the same score units measured in different and incompatible scales.

A multi-criteria evaluation method known as Data Envelopment Analysis (DEA) (Charnes et al., 1978) has been widely used to assess performance. DEA is a mathematical programming method that evaluates the performance of multiple units (called Decision Making Units or DMUs) such as branches of a bank, people, or army combat units. It categorizes evaluation measures into inputs and outputs, where inputs represent the resources used to create an outcome (investments, work-hours, etc.) and outputs represent outcomes of a process (revenues, satisfied customers, products, etc.). Unlike other methods, DEA does not assign weights to each performance measure in a

subjective manner. Instead, weights are determined for each input and output of each DMU by solving a Linear Programming (LP) model. These weights, derived from the data are the optimal set of weights each DMU can have in order to generate the best possible performance score. In addition to a performance score, DEA also identifies benchmark DMUs that serve as reference for each underperforming DMU, determining the practice modifications required for achieving benchmark performance.

DEA has been used to evaluate the performance of primary care MDs treating a specific condition such as otitis media (Ozcan, 1998), asthma (Ozcan 2007) and sinusitis (Pai et al., 2000). Chillingerian and Sherman (1996), Wagner et al, (2003) and Collier et al. (2006) have also used DEA to assess the performance of primary care MDs in a more general context. These are described in greater detail below.

DEA is a more suitable method for evaluating the performance of ED MDs than calculating composite scores because of the following characteristics:

- It does not require the subjective evaluation of the importance of individual measures;
- It can incorporate performance measures evaluated on different scales;
- It identifies sources of inefficient performance for each DMU and determines the necessary improvements under each evaluation measure required to overcome these inefficiencies.

2.2 Performance evaluations using Data Envelopment Analysis (DEA)

Most healthcare applications of DEA have involved performance evaluation of primary care MDs, controlling for the differences in patient mix solely by considering

either patients' severity (Ozcan, 1998), (Ozcan 2007) and (Pai et al., 2000) or the gender and age group of patients (Chillingerian and Sherman, 1996). However, none of these applications considered measures related to the quality of care provided by the MDs. Uniform patient outcomes for all MDs were assumed and focus was directed towards reduction of resource utilization. We argue that assuming uniform patient outcomes does not allow the modeler to account for the trade-offs between care provided and the resources.

Collier et al. (2006) assessed performance of 16 primary care MDs using one input and three outputs. Outputs selected for this evaluation were the total billable charges attributed to the MD and patient satisfaction reported by the MDs' patient panel. Although, patient satisfaction can be considered as a proxy for health services' outcomes, resource utilization was assumed uniform across MDs. The only input considered was the number of Relative Value Units (RVU) per MD - a measure used to capture standards of time duration, resource intensity, cost and malpractice risk for the set of services provided by the MD. The selection of this measure, which would assign the same value of RVUs to a given type of medical service regardless of which MD was responsible for the care, ignores the possibility of practice variation across MDs. Furthermore, since the total number of RVUs per MD depends on the number and type of visits rather than each MD's performance, identification of practice improvements was not considered. In evaluating the performance of ED MDs it should be assumed that there are practice variations in terms of time duration, resource intensity and cost for the same mix of patients' complaints and that an ED MD's performance has an impact on all the measures used for their evaluation.

Wagner et al. (2003) included 3 inputs and 6 outputs to assess performance of 21 primary care MDs. The inputs dealt with payments for admissions, payments for visits by patients of their practice and payments for visits by patients outside their practice. Some outputs were used as control variables (patient panel size, resource intensity of admissions and a health status indicator of patients) while others captured the quality of services' outcomes such as the percentage of readmissions to hospital within 15 days, the percentage of complications as a result of treatment during the course of hospitalization, and health indicators that capture patients' satisfaction and reenrollment with the same MD. This evaluation assumes that all outputs are non-controllable measures which imply that quality of outcomes is not influenced by the MDs' performance. Another shortcoming of this study was the high proportion of MDs deemed as representing benchmark performance (over 50%) which raises doubts about the discriminatory power of the model.

Although DEA has clear advantages over composite score methods due to its ability to consider simultaneously a number of measures without the need of pre-defining their relative importance, models reported in literature are not conducive to ED MDs performance evaluation.

To the best of our knowledge, DEA has never been applied in a comprehensive manner to assess the performance of a group of MDs as we are doing in this research.

Chapter 3: Methodology

This section describes the justification for selecting DEA as modeling method, the theoretical foundations of DEA, a graphical illustration of the main DEA concepts, a taxonomy of DEA models and the specific DEA model applied to ED MDs' performance assessment.

3.1 Justification for Selecting DEA as Performance Evaluation Method

Considering that this research is concerned with a multidimensional evaluation of ED MDs' performance, an evaluation method should meet the following requirements:

1. Ability to consider multiple evaluation measures.
2. Ability to assess performance without the need of human judgment in determining the relative importance of each measure.
3. Ability to establish benchmark performance.
4. Ability to identify improvement opportunities and determine practice modifications required to attain benchmark performance.
5. Ability to adequately discriminate between the performances of DMUs.
6. Ability to identify and penalize compensatory behavior (high performance in one or few measures compensating for low performance on the others).

For the selection of an appropriate assessment tool, not only methods reported in medical literature were considered but also methods that were applied in business, sports,

academics and other fields. However, none of the studies found outside the medical field reported a model that meets our requirements. The reviewed methods included the Analytic Hierarchy Process (AHP) (Saaty, 1994) and the ordinal approach for the development of robust rankings (Cherchye and Vermeulen, 2006), (Van Ours and Vermeulen, 2006). AHP is a technique that involves pairwise comparisons of the measures first (in order to asses their relative importance) and then each possible alternative (ED MD in our case) in order to derive a final ranking. It is important to note that weighting the evaluation criteria produces rankings that are sensitive to changes in these weights' values. To overcome this shortcoming, Cherchye and Vermeulen (2006), as well as Van Ours and Vermeulen, (2006) have proposed a ranking method that does not require expressing the relative importance of each criterion in terms of weights. Their method, defined as an ordinal approach, requires determining if a criterion is more important than another but does not require expressing these differences numerically. Although, this method is more robust and avoids the assignment of weights it still introduces subjectivity when establishing the importance of one criterion relative to another. Moreover, none of these methods provide information that helps establish the necessary performance revisions for an individual DMU that would lead to benchmark practice.

In contrast, careful examination of DEA characteristics allows us to confirm that this methodology has the desired features for our evaluation problem. Although meeting requirement 5 is not guaranteed by DEA, the method has the ability to satisfy it if the number of DMUs is large enough compared to the number of selected inputs and outputs. In regards to requirement 6, there are DEA models that apply restrictions to the eligible

values for weights thereby restricting compensatory behavior. Although the definition of these weights' limits involves a subjective judgment, an original contribution of this research includes a modification of an existing DEA model such that human intervention is reduced when accounting for compensatory behavior.

3.2 Foundations of Data Envelopment Analysis

DEA is a method that uses a linear programming model to carry out evaluations of Decision Making Units (DMUs) such as companies, hospitals, police stations, bank branches, stores, teams, individuals, etc. It applies a benchmarking approach where each DMU's performance is compared to DMUs that represent the benchmark, also known as the efficient DMUs under DEA terminology. It assumes that performance of a DMU can be characterized by a relation between the inputs and outputs. In practice this is translated into assessing a relation between the resources used to conduct a task and the outcome of the task. If a DMU has higher output values and lower input values, DEA will consider it as having better performance than a DMU with low output values and high input values.

A key concept in DEA analysis is the notion of efficiency. DEA considers a DMU efficient if it complies with the Pareto-Koopmans efficiency which states "a DMU is efficient if and only if it is not possible to improve any input or output without worsening some other input or output" (Cooper et al., 2007). Efficient DMUs define the efficient frontier that forms a reference of best performance. Inefficient DMUs are projected onto the efficient frontier to enable their comparison relative to efficient ones with similar profiles in terms of input-output mix.

For each DMU, DEA calculates a measure known as the virtual efficiency score. Depending on the type of DEA model, this score can be interpreted as the required modification (in %) of inputs or outputs to transform an inefficient DMU into an efficient one, or as a value that summarizes all the inefficiencies found in a DMU's inputs, outputs or both.

The basic DEA model calculates the virtual efficiency score as the ratio of a weighted sum of outputs to a weighted sum of inputs. The weights used in the virtual efficiency score are not determined by human judgment but by solving a linear programming model. Thus, they are derived from the data and not from expert opinion. Alternatively to most scoring schemes where a single structure of weights is used to evaluate each unit's performance, the weights in the DEA virtual efficiency score differ across DMUs. This is because for each individual DMU, a model is run to calculate a set of weights for each input and output. These can be interpreted as the set of most favorable weights for a specific DMU, in order to maximize its own virtual efficiency score (Cooper et al., 2007).

3.3 Graphical Illustration of Main DEA Concepts

Figure 1 depicts an example where three DMUs (for example representing three MDs) are being evaluated using one output (i.e. number of patients seen) and two inputs (i.e. number of ordered tests and number of specialist consults). Each point in the graph represents a DMU's performance defined by the values of his/her own inputs and outputs. Assume for the sake of illustration, that all the MDs have the same output value so that the evaluation need only be based on the comparison of their inputs. Thus, MD *D*, which had the lowest number of consults and MD *E*, which had the lowest number of tests, will be considered efficient since they operate with minimal use of resources in at least one of

the inputs. Those MDs define the efficient frontier (dotted line in Figure 1) that envelopes the MD *F*. MD *F* is inefficient since each of its inputs can still be reduced before reaching the efficient frontier. Solving the DEA model will provide information regarding the reduction in inputs required for MD *F* to become efficient.

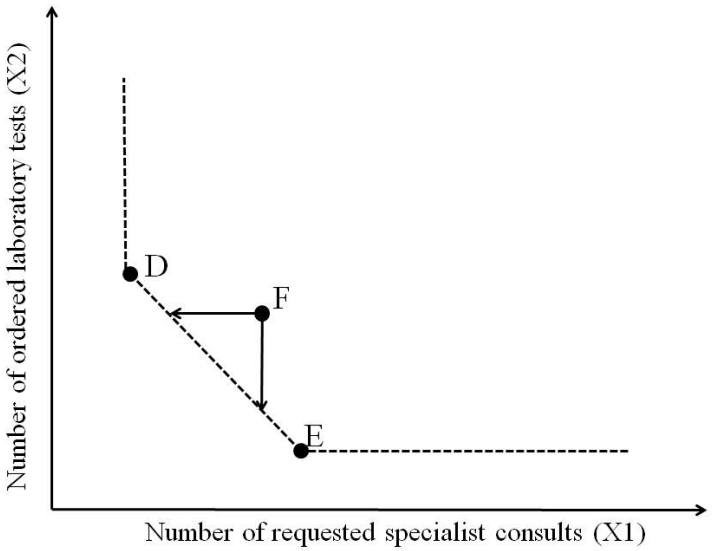

Figure 1. Plot of three MDs

The arrows originating at F, in Figure 1, represent two improvement opportunities available to MD *F*. The horizontal arrow represents the direction in which F must improve its performance by reducing the number of requested specialist consults until it reaches the efficient frontier. Similarly, F can also achieve full efficiency if it modifies its performance by reducing its number of ordered tests (vertical arrow). Simultaneous reduction of both inputs is also a possibility and would land DMU F somewhere on the line between D and E.

DEA's efficient frontier is not only defined by the performances of efficient MDs but also by their linear combinations. In Figure 2, MD *F'* can be represented as a linear

combination of the performances of MDs *D* and *E* and constitutes a projection of F (from Figure 1) onto the efficient frontier after removing the inefficient use of inputs.

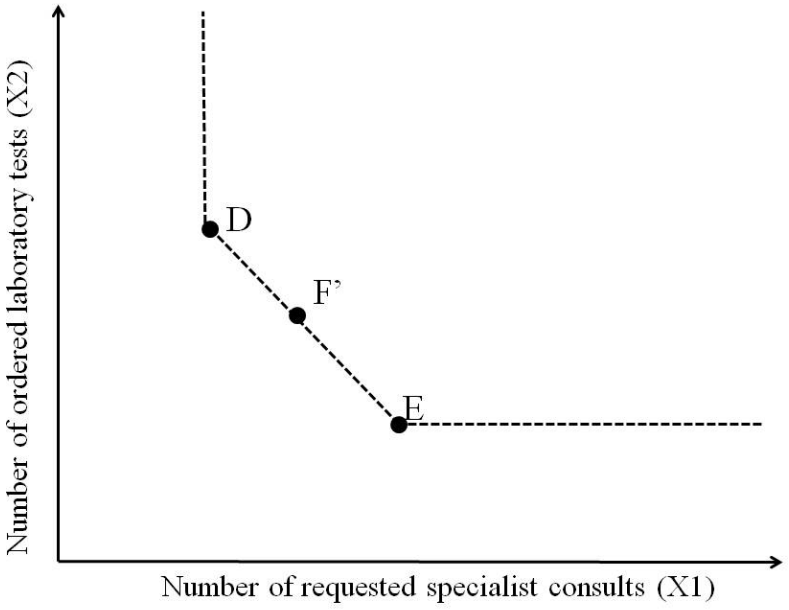

Figure 2. Plot of MD efficient performance

3.4 Taxonomy of DEA Models

All DEA models can be categorized according to model orientation (input-oriented, output-oriented or non-oriented), type of efficiency (technical, mix, others), and a return to scale assumption (constant or variable) (Cooper et al., 2007). In this section we discuss the taxonomy of DEA models.

3.4.1 Model Orientation

Model orientation defines the intended improvements derived from solving the DEA model. These intentions may be to reduce inputs, improve outputs, or both. Assuming that the problem of determining how to transform an inefficient DMU into an efficient one is narrowed just to the issue of improving inputs, then such a problem is modeled by

an *input-oriented DEA model*. Conversely, *an output-oriented model* determines how much the outputs of an inefficient DMU must be increased in order to achieve efficiency. A *non-oriented DEA model* is used to determine simultaneous improvement of inputs and outputs to attain full efficiency (Cooper et al., 2007).

3.4.2 Types of Efficiency

In the DEA literature several types of efficiency are considered but in this research we focus on two of the most common types: technical and mix efficiency. Under the Pareto-Koopmans definition of efficiency, a DMU is efficient only when technical and mix inefficiencies are absent.

A DMU is *technically inefficient* (also called radially inefficient) if it is possible for all its inputs to be reduced proportionally at the same time (Cooper et al., 2007). In an input-oriented model, technical inefficiency is captured by the score that represents the ratio of benchmark performance inputs to actual performance inputs. In an output-oriented model, technical inefficiency is captured by a similar score calculated for the outputs. If these scores equal to 1, the DMU is technically efficient.

In some cases, after the technical inefficiency is removed, there are still improvement opportunities. Figure 3 shows how DMU H is projected onto an efficient frontier defined by DMUs G and I. After proportionally reducing the inputs, H is projected onto point H' where it achieves technical efficiency. However, it can be noticed that for H', the number of specialist consults can still be reduced without worsening the number of laboratory tests, implying that H' does not comply with the Pareto-Koopmans definition of efficiency. This is due to the fact that H' still has *mix inefficiency*. If projection H'

reduces its X1 input until it coincides with DMU I, then it will be located at a point where it will be considered fully efficient.

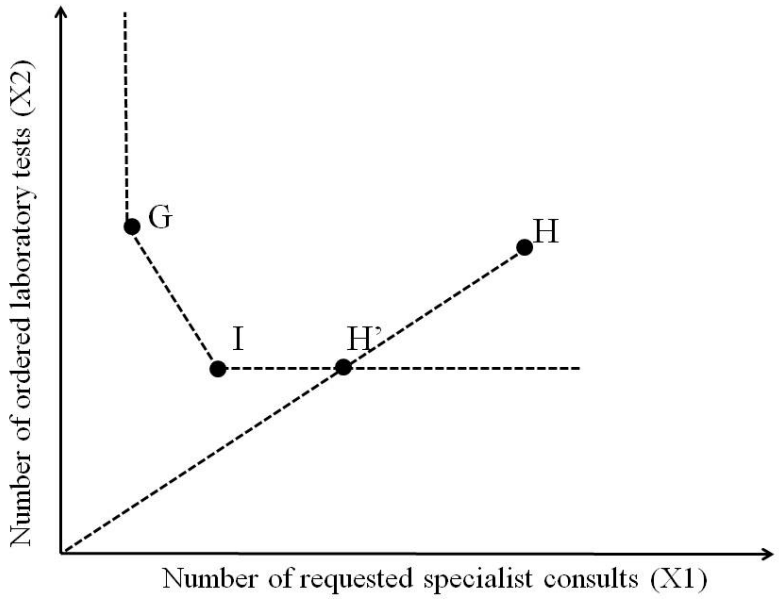

Figure 3. Plot of DMU performances (one output and two inputs) with projection unto inefficient frontier.

3.4.3 Return to Scale Assumptions

Returns to scale refer to the assumed effect of inputs on outputs of a DMU when performing in the efficient frontier. If a proportional change in inputs produces a more than proportional change in outputs, then increasing returns to scale prevails. Conversely, if the effect on outputs is less than proportional, then decreasing returns to scale is assumed. When the proportional change in inputs produces the same proportional change in the outputs, constant returns to scale (CRS) is assumed (Cooper et al., 2007). Some DEA model constraints are used to impose CRS or variable returns to scale (VRS) which allows for constant, increasing and decreasing returns to scale.

Figure 4 shows a graphical representation of the efficient frontier under the constant return to scale assumption, while Figure 5 illustrates the variable return to scale assumption. DMU B is considered inefficient in a constant return to scale (CRS) model while it is considered efficient in a variable return to scale (VRS) model.

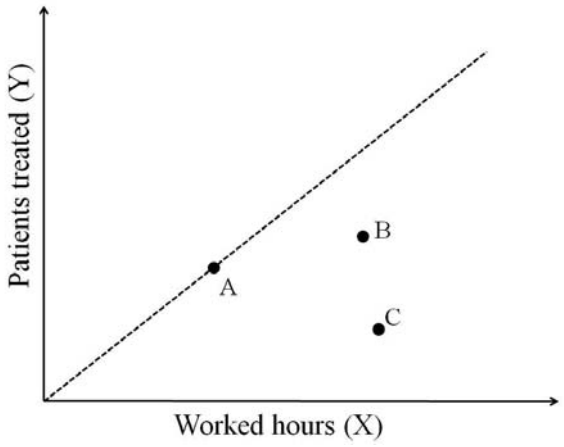

Figure 4. . Efficient frontier under a CRS model (one output and one input) (Cooper et al., 2007)

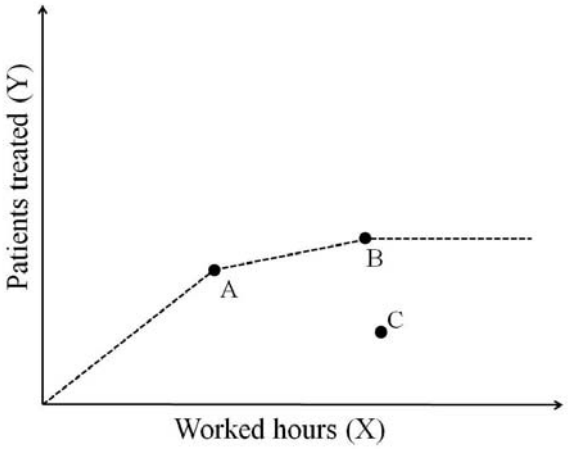

Figure 5. Efficient frontier under a VRS model (one output and one input) (Cooper et al., 2007)

3.5 Taxonomy

The most commonly used DEA models (Cooper et al., 2007) are:

- Charnes, Cooper, Rhodes or CCR input (output)-oriented model (CCR-I (O)) (proposed by Charnes et al., 1978)
- Banker, Charnes, Cooper or BCC input (output)-oriented (BCC-I(O)) (proposed by Banker et al., 1984)
- Additive Model with Constant Returns to Scale (ADD-CRS) (proposed by Charnes et al., 1985)
- Additive Model with Variable Returns to Scale (ADD-VRS) (proposed by Lovell and Pastor, 1995)
- Slack Based Measure with Constant Returns to Scale (SBM-CRS) (proposed by Tone, 2001)
- Slack Based Measure with Variable Returns to Scale (SBM-VRS) (proposed by Tone, 2001)

Below we describe each of these models in more detail.

3.5.1 CCR Model

The CCR model is the basic DEA model. Its formulation, presented in the multiplier form with input orientation is as follows:

CCR FP*o*

Max

$$\theta_o = \frac{u_1 y_{1o} + u_2 y_{2o} + \cdots + u_s y_{so}}{v_1 x_{1o} + v_2 x_{2o} + \cdots + v_m x_{mo}} \qquad (3.1)$$

Subject to:

$$\frac{u_1 y_{1j} + u_2 y_{2j} + \cdots + u_s y_{sj}}{v_1 x_{1j} + v_2 x_{2j} + \cdots + v_m x_{mj}} \leq 1 \quad \text{j} = 1,2,3,...,\text{n} \qquad (3.2)$$

$$v_1, v_2, \ldots, v_m \geq 0 \qquad (3.3)$$

$$u_1, u_2, \ldots, u_s \geq 0 \qquad (3.4)$$

Each DMU *j* has *s* output measures represented by parameters y_{rj} (*r=1, 2, ..., s* and *j=1,2, ...,n*) and *m* input measures represented by parameters x_{ij} (*i=1, 2, ..., m* and *j=1,2, ...,n*). This model has to be solved *n* times (once for each DMU) in order to derive an efficiency score θ for each of them. Values for *y* and *x* are obtained from the data for each DMU, while *u* and *v* weights are the decision variables in the model. These variables are interpreted as measuring the importance of each input and output in the evaluation of the current DMU and unlike typical weighted average methodologies, the set of weights is different for each DMU.

To avoid complications of non-linear expressions represented in (3.1), the CCR FP*o* model can be transformed into an LP model. The resulting CCR LP*o* can be expressed in its dual formulation CCR DLP*o*, known as the envelopment model, written below in a matrix form.

CCR DLP*o* (Phase I)

Min

$$\theta \qquad (3.5)$$

Subject to:

$$\theta \boldsymbol{x_o} \geq \boldsymbol{X\lambda} \qquad (3.6)$$

$$\boldsymbol{y_o} \leq \boldsymbol{Y\lambda} \qquad (3.7)$$

$$\boldsymbol{\lambda} \geq \boldsymbol{0} \qquad (3.8)$$

This input oriented CCR DLP*o* (Phase I) model solves for variables θ and **λ**. It minimizes the value of θ (3.5) that multiplies the actual input vector until this product reaches a limit defined by a linear combination of DMUs' inputs (3.6). In other words, θ

is the factor by which the inputs are reduced. Constraints described in (3.7) ensure that output vector is always less than or equal to a linear combination of DMUs' outputs. Each λ is associated with a DMU and optimal values greater than zero indicate that a given DMU is efficient. The constraint set in (3.8) ensures that variables are always non-negative.

After solving for θ and **λ** for each DMU, the improvements that would transform each of them into benchmark performers can be calculated. This projection of the DMU onto the efficient frontier is a linear combination of the performance of DMUs in its reference benchmark group. For CCR DLP*o*, the projections can be calculated by defining θ^* as the optimal score for the DMUo. Then, its improved performance will be defined as (θ^* ,) where $\theta^* \leq 1$. Notice that for $\theta^*=1$, no changes in the DMU's current performance are required since the DMU is technically efficient. However, even after these improvements, some inefficiency could still remain since this model deals only with technical efficiency. To attain full efficiency, any mix inefficiency present in a DMU's performance must be eliminated. This is accomplished by solving a Phase II model using the optimal solution of CCR DLP*o*.

CCR DLP*o* (Phase II)

Max

$$\boldsymbol{1}^t \boldsymbol{s}^- + \boldsymbol{1}^t \boldsymbol{s}^+ \qquad (3.9)$$

Subject to:

$$\theta^* \boldsymbol{x}_o - \boldsymbol{s}^- = \boldsymbol{X}\boldsymbol{\lambda} \qquad (3.10)$$

$$\boldsymbol{y}_o + \boldsymbol{s}^+ = \boldsymbol{Y}\boldsymbol{\lambda} \qquad (3.11)$$

$$\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{s}^- \geq \boldsymbol{0}, \boldsymbol{s}^+ \geq \boldsymbol{0} \qquad (3.12)$$

This model solves for s^-, s^+, λ and uses θ^* obtained in the Phase I as a parameter. After solving the above model and obtaining optimal values s^{-*}, s^{+*} and λ^* for each DMU, final improvements can be established and a fully efficient performance can be calculated as ($\theta^* \quad - s^{-*}, \quad + s^{+*}$).

The CCR DLP*o* (Phase I) formulation can be easily transformed to an output oriented model by changing the objective function (3.5) to maximize the value of the η score and associating η to the output constraints instead of the input constraints.

3.5.2 BCC Model

BCC model formulation is as follows:

BCC DLP*o*

Min

$$\theta \qquad (3.13)$$

Subject to:

$$\theta x_o \geq X\lambda \qquad (3.14)$$

$$y_o \leq Y\lambda \qquad (3.15)$$

$$1^t \lambda = 1 \qquad (3.16)$$

$$\lambda \geq 0 \qquad (3.17)$$

Since this model assumes a variable returns to scale (VRS), its efficient frontier is represented by convex combinations of efficient DMUs. In order to impose the convexity condition, the constraint (3.16) is added to impose convex combinations of λs by ensuring the sum of λs is equal to 1. The BCC formulation can be easily transformed into an output-oriented model.

3.5.3 ADD Model

The ADD model, also known as the Additive Model is a non-oriented model that simultaneously considers technical and mix inefficiencies and works under CRS or VRS assumptions. Solving the ADD model does not produce a score that measures the inefficiencies. Instead it determines the excess of inputs and shortfall of outputs for each DMU relative to its benchmark. Using this model, it is not possible to differentiate between technical and mix inefficiency.

The ADD-CRS model is formulated as follows:

ADD CRS DLP*o*

Max

$$\boldsymbol{1}^t \boldsymbol{s}^+ + \boldsymbol{1}^t \boldsymbol{s}^- \qquad (3.18)$$

Subject to:

$$\boldsymbol{x}_o - \boldsymbol{s}^- = \boldsymbol{X}\boldsymbol{\lambda} \qquad (3.19)$$

$$\boldsymbol{y}_o + \boldsymbol{s}^+ = \boldsymbol{Y}\boldsymbol{\lambda} \qquad (3.20)$$

$$\boldsymbol{\lambda} \geq \boldsymbol{0}, \boldsymbol{s}^- \geq \boldsymbol{0}, \boldsymbol{s}^+ \geq \boldsymbol{0} \qquad (3.21)$$

In the above formulation the variables $\boldsymbol{s}^-$ and $\boldsymbol{s}^+$ represent the gap between the actual performance and the benchmark. They are called slack variables regardless of whether they represent excess input or output shortfall (Cooper et al., 2007).

The ADD-CRS model can be easily transformed into the ADD-VRS by adding the convexity condition constraint similar to the one required for the BCC model.

3.5.4 SBM Model

The SBM model addresses the limitation of the ADD model in that it allows for the calculation of a score. This model produces a score ρ with a value between 0 and 1 that is

interpreted as "the ratio of mean input and output mix inefficiencies" (Cooper et al., 2007). A score of ρ = 1 indicates an efficient DMU.

The variables of the SBM model are **λ**, $^{-}$ and $^{+}$, and its formulation under the VRS assumption is the following:

SBM VRS FP*o*

Min

$$\frac{1-\frac{1}{\mathrm{m}}\sum_{\mathrm{i}=1}^{\mathrm{m}}\frac{s_i^-}{x_{io}}}{1+\frac{1}{\mathrm{s}}\sum_{\mathrm{r}=1}^{\mathrm{s}}\frac{s_r^+}{y_{ro}}} \qquad (3.22)$$

Subject to:

$$\boldsymbol{x_o - s^- = X\lambda} \qquad (3.23)$$

$$\boldsymbol{y_o + s^+ = Y\lambda} \qquad (3.24)$$

$$\boldsymbol{1^t \lambda = 1} \qquad (3.25)$$

$$\boldsymbol{\lambda \geq 0, s^- \geq 0, s^+ \geq 0} \qquad (3.26)$$

The above formulation takes into account all types of inefficiencies by maximizing the values of slack variables. Objective function (3.22) can be interpreted as the ratio of mean input and output mix inefficiencies (Cooper et al., 2007). The first constraint (3.23) ensures that the input reduction represented by the augmentation of slack variable $^{-}$does not go beyond a value represented by a linear combination of DMUs' inputs. The second constraint (3.24) restricts the increase of evaluated DMU's output, represented by the augmentation of slack variable $^{+}$, to a value represented by a linear combination of DMUs' outputs. The third constraint (3.25) adds the convexity condition by ensuring the

sum of variables is equal to 1. Improved performance of a DMU_o evaluated under this model is calculated as ($-\boldsymbol{s}^{-*}$, $+\boldsymbol{s}^{+*}$).

SBM VRS FP_O is presented as a fractional programming model, meaning its objective function is expressed as a ratio. This can be transformed to a tractable LP formulation as follows:

SBM VRS LP*o*

Min

$$t - \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \frac{S_i^-}{x_{io}} \qquad (3.27)$$

Subject to:

$$t + \frac{1}{s} \sum_{r=1}^{s} \frac{S_r^+}{y_{ro}} = \mathbf{1} \qquad (3.28)$$

$$t\boldsymbol{x_o} - \boldsymbol{S^-} = \boldsymbol{X\Lambda} \qquad (3.29)$$

$$t\boldsymbol{y_o} + \boldsymbol{S^+} = \boldsymbol{Y\Lambda} \qquad (3.30)$$

$$\boldsymbol{1}^t \boldsymbol{\Lambda} = \boldsymbol{1}^t t \qquad (3.31)$$

$$\boldsymbol{\Lambda} \geq \boldsymbol{0}, \boldsymbol{S^-} \geq \boldsymbol{0}, \boldsymbol{S^+} \geq \boldsymbol{0}, t > 0 \qquad (3.32)$$

The above formulation assumes $\boldsymbol{S^-} = t\boldsymbol{s^-}, \boldsymbol{S^+} = t\boldsymbol{s^+}$ and $\boldsymbol{\Lambda} = t\boldsymbol{\lambda}$ and solves for variable t and variable vectors $\boldsymbol{S^-}, \boldsymbol{S^+}$ and $\boldsymbol{\Lambda}$.

SBM VRS LP*o* can also be considered in its dual form, which will serve as the basis to illustrate the model modification proposed in a later section, to address the compensatory behavior:

SBM VRS DLP*o*

Max

$$\xi \qquad (3.33)$$

Subject to:

$$\xi + \boldsymbol{v}\boldsymbol{x}_o - \boldsymbol{u}\boldsymbol{y}_o - \pi_o = \mathbf{1} \quad (3.34)$$

$$-\boldsymbol{v}\boldsymbol{X} + \boldsymbol{u}\boldsymbol{Y} + \pi_o \leq \mathbf{0} \quad (3.35)$$

$$\boldsymbol{v} \geq \frac{1}{m}[1/\boldsymbol{x}_o] \quad (3.36)$$

$$\boldsymbol{u} \geq \frac{\xi}{s}[1/\boldsymbol{y}_o] \quad (3.37)$$

The above model solves for the variables $\boldsymbol{v},\boldsymbol{u}$, ξ and π. Variables represented by v and u are the weights to be calculated for each of the m inputs represented by x and s outputs represented by y respectively. Variable π is used to implement an assumption of variable returns to scale. The objective function (3.33) maximizes the values of variable ξ that represents the score. Constraints (3.34) and (3.35) set bounds on the variables to ensure that variable ξ always has a value less than or equal to 1. Constraints (3.36) and (3.37) ensure that $\boldsymbol{v}$ and $\boldsymbol{u}$ are positive.

3.6 Issues with DEA Models

Given the nature of DEA models, which assign the most favourable weights to inputs and outputs for each DMU in order to maximize its score, some DMUs will attain a high score because of a very good performance in a single input or output. In general it is possible to say that a DMU that has the highest ratio of one of the outputs to one of the inputs will be deemed efficient (Ali Emrouznejad's Data Envelopment Analysis webpage found at deazone.com). This is caused by a model assigning extreme values to weights (asymmetric weight assignment) in a way that promotes the contribution of high performing inputs/outputs to the score and devalues poorly performing inputs/outputs (Dimitrov and Sutton, 2010). Such compensatory behaviour reduces the discriminatory power of DEA models. This problem is further exacerbated as the number of inputs (m) and outputs (s) increases for a given number of DMUs as the number of efficient DMUs

will be approximately equal to the product *ms*. As *ms* grows, the number of efficient DMUs will approximate the number of total DMUs in the evaluation, hence, reducing the ability of DEA to identify differences among their performances.

To deal with the compensation issue, DEA researchers have proposed including constraints to restrict the values of the weights (Allen et al., 1997). One of the well-known methods for restricting weights is a DEA model with upper and lower limits on the ratios of each pair of input weights, each pair of output weights or pairs of input and output weights. Other approaches apply absolute weight restrictions which set limits for each individual weight.

The drawback of using weight restrictions is the subjectivity involved in the selection of limits. In some cases these limits are determined through expert opinion, knowledge of the cost and price of inputs and outputs or both (Allen et al., 1997). However, regardless of the method, the difficulty of justifying values for these limits remains a challenge.

3.7 Models with Symmetric Weight Assignment Technique (SWAT)

An alternative approach to defining weight restrictions is the SWAT (Symmetric Weight Assignment Technique) approach proposed by Dimitrov and Sutton (2010). Several DEA models can be modified in order to incorporate SWAT, by penalizing the weights' differences in the objective function. As an effect of this penalization, the score of a DMU is reduced as the difference between the weights of individual pairs (input-input, output-output and input-output) increases. Since higher compensation behaviour is directly related to higher differences between the weights of pairs (weights' asymmetry), SWAT is an effective approach to avoid the rewarding of DMUs exhibiting compensatory behaviour.

3.7.1 Preferred Model for Performance Evaluation of ED MDs

Since it is reasonable to assume that ED MDs' performance should influence their own results under each of the measures by which they are being evaluated, a non-oriented model is preferred. Only non-oriented models such as ADD and SBM are capable of identifying improvement opportunities in each input and output. However, the ADD model does not provide a performance score that can be used to identify how near or far an ED MD is from benchmark. On the other hand, solving the SBM model produces a score that captures all input and output inefficiencies of an ED MD's performance and provides information to determine the level of compliance with the benchmark.

Unfortunately, the absence of compensatory behavior in ED MDs performance cannot be guaranteed; therefore the SBM model should incorporate SWAT to deal with this issue. The use of SWAT in the model also ensures higher discrimination power even for a small sample size.

Finally, there is no information available to determine what type of returns to scale (increasing, constant or decreasing) prevails in the ED context; therefore all possibilities should be left open by assuming VRS conditions.

The only model that effectively addresses all the above points is the SBM-SWAT VRS DLP*o* model which justifies its selection for evaluating performance of ED MDs.

To the best of our knowledge, the SWAT has been used only in CCR and BCC models (Dimitrov and Sutton, 2010). In this research we use SWAT in the SBM models. The proposed formulation is as follows:

SBM-SWAT VRS LP*o*

Max

$$\xi - \beta z^{-} - \beta z^{+} - \beta z^{-+} \quad (3.38)$$

Subject to:

$$\xi + \boldsymbol{v}\boldsymbol{x}_o - \boldsymbol{u}\boldsymbol{y}_o - \pi_o = 1 \quad (3.39)$$

$$-\boldsymbol{v}\boldsymbol{X} + \boldsymbol{u}\boldsymbol{Y} + \pi_o \leq \boldsymbol{0} \quad (3.40)$$

$$\boldsymbol{v} \geq \frac{1}{\mathrm{m}}[1/\boldsymbol{x}_o] \quad (3.41)$$

$$\boldsymbol{u} \geq \frac{\xi}{\mathrm{s}}[1/\boldsymbol{y}_o] \quad (3.42)$$

$$\boldsymbol{v}_l - \boldsymbol{v}_k \leq z_{lk}^{-} \quad \text{for } \mathrm{l} \neq \mathrm{k} \quad (3.43)$$

$$\boldsymbol{v}_k - \boldsymbol{v}_l \leq z_{lk}^{-} \quad \text{for } \mathrm{l} \neq \mathrm{k} \quad (3.44)$$

$$\boldsymbol{u}_l - \boldsymbol{u}_k \leq z_{lk}^{+} \quad \text{for } \mathrm{l} \neq \mathrm{k} \quad (3.45)$$

$$\boldsymbol{u}_k - \boldsymbol{u}_l \leq z_{lk}^{+} \quad \text{for } \mathrm{l} \neq \mathrm{k} \quad (3.46)$$

$$\boldsymbol{v}_l - \boldsymbol{u}_k \leq z_{lk}^{-+} \quad \text{for all } \mathrm{l,k} \quad (3.47)$$

$$\boldsymbol{u}_l - \boldsymbol{v}_k \leq z_{lk}^{-+} \quad \text{for all } \mathrm{l,k} \quad (3.48)$$

This formulation is an extension of the SBM VRS DLP$_O$ formulation presented in (3.33) - (3.37). It solves for the variables $\boldsymbol{v},\boldsymbol{u},\boldsymbol{z}^{-},\boldsymbol{z}^{+}$, $\mathbf{z}^{-+}$, ξ and π. Constraints (3.39) - (3.40) are the same constraints as (3.34) - (3.35).

For the SBM-SWAT VRS LP*o* model, the objective function maximizes the difference between variable ξ and its penalization. This penalization comprises the sum of variables $\boldsymbol{z}$, each of them multiplied by a constant β, labelled as the symmetry factor (Dimitrov and Sutton, 2010). As β increases, a greater penalization of the score is obtained which leads to greater discriminatory power at the expense of producing lower variation of weights' values. Excessive limitation of weights' variability goes against one

of the main advantages of DEA models that allows for a different set of input/output weights to be calculated for each DMU.

Constraints (3.43) and (3.44) are used to ensure that variables z^- capture an absolute value of the difference between each pair of the input weights while (3.45) and (3.46) serve the same purpose for each pair of the output weights. Constraints (3.47) and (3.48) are used to ensure that variables z^{-+} captures an absolute value of the difference between each pair of the input-output weights.

The above formulation does not allow identifying slack variables that represent required modifications of practice and the reference DMUs that will serve as benchmarks for each evaluated DMU. The dual form of the SBM-SWAT VRS LP*o* overcomes these difficulties. The formulation is as follows:

SBM-SWAT VRS DLP*o*

Min

$$t - \frac{1}{\mathrm{m}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \frac{S_i^-}{x_{io}} \tag{3.49}$$

Subject to:

$$t + \frac{1}{s} \sum_{r=1}^{s} \frac{S_r^+}{y_{ro}} = \mathbf{1} \tag{3.50}$$

$$\boldsymbol{x}_o t - \boldsymbol{S}^- = \boldsymbol{X\Lambda} + \boldsymbol{\Psi\delta}^- - \boldsymbol{\Psi\omega}^- + \boldsymbol{\Gamma\delta}^{-+} - \boldsymbol{\Gamma\omega}^{-+} \tag{3.51}$$

$$\boldsymbol{y}_o t + \boldsymbol{S}^+ = \boldsymbol{Y\Lambda} + \boldsymbol{\Pi\delta}^+ - \boldsymbol{\Pi\omega}^+ + \boldsymbol{\Delta\delta}^{-+} - \boldsymbol{\Delta\omega}^{-+} \tag{3.52}$$

$$\boldsymbol{1}^t \boldsymbol{\Lambda} = \boldsymbol{1}^t t \tag{3.53}$$

$$\boldsymbol{\delta}^- + \boldsymbol{\omega}^- = \beta \tag{3.54}$$

$$\boldsymbol{\delta}^+ + \boldsymbol{\omega}^+ = \beta \tag{3.55}$$

$$\boldsymbol{\delta}^{-+} + \boldsymbol{\omega}^{-+} = \beta \tag{3.56}$$

$$\boldsymbol{S}^-, \boldsymbol{S}^+, \boldsymbol{\Lambda}, \boldsymbol{\delta}^-, \boldsymbol{\delta}^+, \boldsymbol{\delta}^{-+}, \boldsymbol{\omega}^-, \boldsymbol{\omega}^+, \boldsymbol{\omega}^{-+} \geq 0 \tag{3.57}$$

$$t > 0 \tag{3.58}$$

The above model solves for variables $t, \boldsymbol{S}^{-}, \boldsymbol{S}^{+}, \boldsymbol{\Lambda}, \boldsymbol{\delta}^{-}, \boldsymbol{\delta}^{+}, \boldsymbol{\delta}^{-+}, \boldsymbol{\omega}^{-}, \boldsymbol{\omega}^{+}$ and $\boldsymbol{\omega}^{-+}$, where $\boldsymbol{S}^{-} = t\boldsymbol{s}^{-}, \boldsymbol{S}^{+} = t\boldsymbol{s}^{+}$ and $\boldsymbol{\Lambda} = t\boldsymbol{\lambda}$. Matrixes $\boldsymbol{\Psi}, \boldsymbol{\Gamma}, \boldsymbol{\Pi}$ and $\boldsymbol{\Delta}$ are comprised of the values 1, 0 and -1 and are used to generate patterns that add, subtract and omit the $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ variables in each constraint. Variables $\boldsymbol{\delta}^{-}$ and $\boldsymbol{\omega}^{-}$ are dual variables related to constraints that capture the input-to-input weight differences in (3.43) and (3.44). Variables $\boldsymbol{\delta}^{+}$ and $\boldsymbol{\omega}^{+}$ relate to constraints that capture output-to-output weight differences in (3.45) and (3.46) and $\boldsymbol{\delta}^{-+}$ and $\boldsymbol{\omega}^{-+}$ relate to input to output weight differences in (3.47) and (3.48). Each constraint in a primal model, including a pair of input-input weights, output-output weights or input-output weights is related to a particular pair of $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ dual variables. Each $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ pair adds up to β as expressed in constraints (3.54), (3.55) and (3.56). The presence of the $\boldsymbol{\delta}$ and $\boldsymbol{\omega}$ variables in constraints (3.51) and (3.52) produce alterations to the efficient frontier making the achievement of full efficiency more difficult as β increases.

Improved performance of a DMU_o evaluated under this model is calculated as ($-\boldsymbol{s}^{-*}$, $+\boldsymbol{s}^{+*}$), where $\boldsymbol{s}^{-*}$ and $\boldsymbol{s}^{+*}$ are the optimal values of slack variables calculated from expressions $\boldsymbol{S}^{-}* = t * \boldsymbol{s}^{-}*$ and $\boldsymbol{S}^{+*} = t * \boldsymbol{s}^{+}*$ respectively. An improved DMU's performance (*, *) is efficient only if all inputs and outputs are simultaneously modified according to the optimal values of slack variables.

Chapter 4: Performance Evaluation Measures for ED MDs

Selected performance evaluation measures should represent ED MDs' practice in each of the dimensions considered in the evaluation. Inputs and outputs that are directly affected by ED MDs' actions and decisions and not by uncontrollable factors is prefered. To some extent, this allows attributing responsibility of individual assessment results to ED MDs' performance and aids setting achievable improvement goals.

In an ideal situation, only measures meeting the above-description should be included in a model; however in most situations data is neither reliable nor available.

Generally, more performance measures than the ones selected in the research can be considered, subject to data that meets criteria described in next section.

4.1 Selection of Inputs and Outputs

Although timeliness of care, quality of health outcomes, utilization of resources and patient throughput of an ED MD service could be captured in a number of possible performance measures, there are certain limiting factors that influence their selection:

- Data availability: there is a limited availability of data (or data is very difficult to access) that precludes the use of some measures.

- Data reliability: qualitative measures (e.g. communication skills, patient satisfaction) are prone to be biased since they require subjective judgments to be expressed in a numerical scale.
- Variation between MDs' performances: only measures with sufficient discriminatory power should be considered.
- Frequency of occurrence: some events that do not occur very often would require an extremely large data set in order to be representative of ED MD's pattern of practice.
- Degree of control the MD has over the selected measures: there are instances where performance depends on factors that are outside the ED MD's control. Such measures should be excluded from consideration.
- Number of DMUs in a model: when using the DEA model, the higher the number of inputs and outputs, the less the discriminatory power of the performance evaluation. It is recommended that the total number of DMUs under assessment be equal to or greater than max $\{ms , 3(m + s)\}$ where m is the number of inputs and s is the number of outputs (Cooper et al., 2007).

After defining the performance measures for the evaluation, the use of a DEA model requires the classification of these measures into inputs and outputs. Inputs should represent the utilization of resources and their reduction should be associated with efficiency improvement assuming the output level remains fixed. For example, if an ED MD reduces the number of laboratory orders while keeping the same level of health outcomes, he/she would be improving efficiency; therefore the number of laboratory orders should be classified as an input. Conversely, outputs should represent outcomes of

the service provision and their increase should be associated with more efficient performance assuming the level of inputs remains unchanged.

Table 1 presents the inputs and outputs proposed for this research (for each ED MD labeled j) :

Inputs	Outputs
Average encounter time per patient visit[1] ($AVG\_MDTIME\_PAT_j$)	Rate of non-return patient visits within 72 hours ($RATE\_NR72_j$)
Average number of laboratory tests per patient visit ($AVG\_LAB\_PAT_j$)	Rate of patient visits per hour worked ($RATE\_PAT\_WH_j$)
Average number of radiology orders per patient visit ($AVG\_RAD\_PAT_j$)	

Table 1. Inputs and outputs used in the research

Inputs $AVG\_MDTIME\_PAT_j$, $AVG\_LAB\_PAT_j$ and $AVG\_RAD\_PAT_j$ are calculated per patient visit, which controls for differences in the number of patients seen by each ED MD. $RATE\_NR72_j$ is a rate, which also controls for differences in the number of patients seen. $RATE\_PAT\_WH_j$ is calculated per worked hour, which accounts for differences in the number of worked hours for each ED MD.

Some of the potential inputs and outputs that were excluded from this evaluation were costs, number of specialist consults, hospital admissions, adverse events, patient satisfaction, length of stay (LOS) in the ED and the waiting time to be seen by ED MD.

[1] One patient might have more than one visit; therefore the number of patients is not necessarily equal to the number of patient visits.

Expressing resource utilization in terms of costs would have provided a direct link between the use of diagnostic tests and cost efficiency of each ED MD's performance, however, these data were not available. Each time an ED MD requests a specialist consult, he/she is using an expensive resource. However, registry of these consults was not reliable enough to be used in the research. Admission of patients to the hospital would represent an additional use of resources as a consequence of an ED MD decision; however, there was no significant variation in practice according to this measure. Adverse events attributable to the specific ED MD are important indicators of patient outcome. These were not included since at CHEO they are very rare and using them would require a very large sample of patient visits. Patient satisfaction is an important outcome but cannot be directly attributed to a specific ED MD but rather reflects overall ED experience. Since factors such as ED load, nursing staff availability and shift handovers impact the waiting time of patients, LOS and the waiting time to be seen were not included due to the lack of direct control of the ED MDs over these measures.

Average Encounter Time per Patient Visit (AVG_MDTIME_PAT$_j$)

An encounter time is defined as the number of minutes between the first ED MD contact with the patient until the time a disposition decision is made. This time interval is also known as "doctor to discharge time" or "doctor to decision to admit time" (Welch et al., 2006). The first contact with the patient is interpreted as the moment when the ED MD approaches the patient to evaluate his/her condition. The disposition decision refers to the moment when the ED MD records the disposition decision on a patient's chart (electronic or paper). Events outside this interval are not under the control of the ED MD so they are not included. Within the encounter time, the ED MD does not control the time

spent waiting for tests results etc. This portion of time may influence the total encounter time; however, it is reasonable to assume that this equally affects all ED MDs.

One of the reasons for using this measure is that it has an impact on the length of stay (LOS) of a patient, which is considered to be a good measure of a EDs' performance (Hung and Chalut, 2008). A patient's LOS in the ED depends on factors such as front-end processes (triage, registration), ED load, number of ED MDs on staff etc. All of these factors add to LOS but the ED MD does not control some of them so they should not be considered. Thus, instead of using LOS as an input, we use $MDTIME\_PAT_j$ since the level of control of ED MDs' over this measure is higher.

For each MD_j, $AVG\_MDTIME\_PAT_j$ is calculated as an average duration of the ED MD encounter per patient visit. This measure is selected as an input since the MD's time is a consumed resource.

Average of Number Laboratory Tests per Patient Visit ($AVG\_LAB\_PAT_j$) and Average Number of Radiology Orders per Patient Visit ($AVG\_RAD\_PAT_j$)

These input measures are intended to capture practice variations in diagnosing a patient (some ED MDs may diagnose patients from the same diagnostic group using a smaller number of tests and radiology orders). They are selected as inputs because they represent resources used in the process of managing a patient in the ED. Unlike other inputs, laboratory tests and radiology orders can be directly linked to the use of financial resources. Similar inputs have been used in almost all healthcare applications of DEA models (Ozcan, 1998), (Ozcan 2007), (Pai et al, 2000), (Chillingerian and Sherman, 1996).

It is expected that benchmark ED MDs will arrive at the correct diagnosis while ordering fewer tests and radiology orders than others.

To define these inputs for each MD_j, the average number of laboratory tests (radiology orders) per patient visit requested by MD_j during the study timeframe is calculated.

Rate of Non-Return Patient Visits within 72 hours ($RATE\_NR72_j$)

A return visit to the ED (for the same presentation) within 72 hours of discharge is considered to be a measure of care quality. According to a survey of PED medical directors of accredited Canadian PED programs (Hung and Chalut, 2008), the value of the $RATE\_RNR72_j$ was ranked as one of the most useful indicators of ED MDs' performance.

The $RATE\_NR72_j$ is an output because it is a proxy for the quality of care provided by the ED MD. Due to the nature of DEA models that consider higher values of outputs as being more desirable than lower values, the rate of non-return visits (more being better) is used instead of the normally used rate of return visits.

For each ED MD labeled *j*, the calculation of _ 72 is as follows:

$$\text{RATE\_NR72}_j = \mathbf{1} - \frac{SUM\_RV\mathbf{72}_j}{SUM\_PAT_j} \qquad (4.1)$$

In 4.1, SUM_PAT_j represents the sum of patient visits associated with ED MD *j* and SUM_RV72*j* represents the sum of return visits associated with ED MD *j*.

Rate of Patient Visits per Hour Worked ($RATE\_PAT\_WH_j$)

Throughput of patients is represented by the volume of patients that an ED MD is capable of managing in a given time interval. This measure is important in the assessment

of ED MDs' performance since it captures the productivity of each ED MD. It is influenced by the rate of patient arrivals, therefore it must be noted that if a low volume of patient arrivals is observed, this will negatively impact the throughput of the ED MD.

For each ED MD this output is calculated as the number of patient visits seen during the study period divided by the total number of working hours of this ED MD.

Unfortunately, due to the experimental design of our study (described in a later section) that requires the disaggregation of visits into disease categories, this output measure was eliminated from the evaluation since, the worked hours of each ED MD (a data item required for the calculation of this output) cannot be disaggregated into different categories of visits

Chapter 5: Experimental Design

5.1 Study Setting

The Children's Hospital of Eastern Ontario (CHEO) is a pediatric academic hospital located in Ottawa, Ontario. Its ED has more than 65,000 visits per year (about 190 per day) (http://www.cheo.on.ca/uploads/Welcome%20to%20ED.pdf).

CHEO offers medical services to patients from birth until 18 years of age. Presenting complaints of PED visits from patients of this age group are less diverse than what is observed in the general population. Also, resource utilization and length of stay in the ED is greater for older patients than it is for the pediatric population (Baum and Rubenstein, 1987). Nevertheless, reasons for visiting a PED are numerous and practice variations are expected across presenting complaints due to the difference in resource utilization patterns for each type of complaint. This fact motivates the use of a stratification approach, applied in this research, that produces performance evaluations per type of complaint.

ED MDs at CHEO are full-time or part-time physicians. Some of them have administrative responsibilities and since CHEO is a teaching hospital, full-time ED MDs are involved in teaching medical residents and ED fellows. It is hypothesized that the presence of trainees might have an influence on ED MDs' performance given that trainees often assist ED MDs in the medical assessment of patients. This motivates a

further stratification based on the level of the trainee. Identifying ED MDs who perform better/worse in the presence and absence of trainees, provides insights into the trade-offs between teaching and clinical work.

5.2 ED MDs Included in the Sample

To determine which ED MDs should be included in the evaluation, each ED MD needs to satisfy all of the following criteria:

- Must be a full-time member of the CHEO PED medical staff during the entire year 2011.
- Must have worked all types of shifts spread throughout 2011.

The first criterion ensures all ED MDs have worked during all seasons of 2011 and hence were exposed to similar patient arrival patterns and seasonal disease prevalence. The second criterion ensures that each ED MD included in the study has assessed similar mix of patients during the study period (did not work specific shifts only).

The above criteria narrowed our focus down to 20 ED MDs at CHEO.

Inputs and outputs of each ED MD were calculated using data gleaned from records of patient visits. To be considered for these calculations each patient visit had to meet the following criteria:

- Must be associated with ED MDs enrolled in the study.
- Must take place in 2011.
- Must be associated with the set of most frequent presenting complaints used in complaints' stratification (defined in Section 5.6.1).

5.3 Data

Data for this research were extracted from PED visits' records, laboratory tests and radiology orders for the year 2011.

Specifically data were extracted from the following information systems at CHEO:

- EDIS (Emergency Department Information System)
- LIS (Laboratory Information System)
- RIS (Radiology Information System)

Inputs and outputs were calculated from data items extracted from the information systems as described in Table 2 below.

Input / Output	**Source Information System**
$AVG\_MDTIME\_PAT_j$	EDIS
$AVG\_LAB\_PAT_j$	LIS
$AVG\_RAD\_PAT_j$	RIS
$RATE\_NR72_j$	EDIS

Table 2. Inputs / outputs and data sources

The ED MD encounter time of each visit is not stored in any information system; however EDIS contains data items of Physician Initial Assessment (PIA) time and disposition decision time. Prior to obtaining the $AVG\_MDTIME\_PAT_j$ input, the difference between these data items was calculated to obtain the duration of the ED MD encounter with each patient (MDTIME_PAT). For the patient visits attributable to each ED MD, the average of this difference ($MDTIME\_PAT_j$) was calculated to obtain $AVG\_MDTIME\_PAT_j$. It is worth mentioning that the ED MD who starts the patient assessment is not necessarily the same ED MD that makes the disposition decision, due

to shift handovers. However, we assumed that such a visit is attributed to the ED MD who started the assessment since he/she defines the main trajectory of the management that will be followed by the second ED MD.

To obtain AVG_LAB_PAT$_j$ input, data preprocessing was required. LIS contains multiple entries for a single patient's visit, each associated with a specific test parameter item. In order to collapse multiple test items into a single test, a correspondence table provided by the medical expert was used to determine the membership of these items in a single test. For example, entries corresponding to sodium, potassium and chlorine levels were aggregated as a single electrolytes test.

No preprocessing of data was required to obtain values for AVG_RAD_PAT$_j$ from RIS.

In regards to RATE_NR72$_j$, assessment by an independent medical expert was needed to determine if a return visit to the PED within 72 hours after discharge of the same patient should be associated with performance of the ED MD who assessed this patient during the first visit. This was done by comparing presenting complaints reported in consecutive visits. If similarities between complaints in these visits indicated an unresolved health problem of the patient, then the return visit was considered a consequence of inadequate management and was attributed to the previously attending ED MD.

Information about each visit's presenting complaint (CEDIS code) and the absence/presence and type of trainee, who assisted the ED MD during the visit, were obtained to enable data stratifications explained in later Section 5.6.2. These data items were extracted from EDIS.

5.4 Data Screening

Initial data set comprises 36,441 visits classified under the 25 highest volume complaints. These were assessed during 2011 by the 20 ED MDs under evaluation.

As a first step, visits associated with 7 of the 25 most frequent presenting complaints were excluded from further analysis. Reasons for the exclusion these complaints are described in Section 5.6.1. Data errors affecting the calculation of the values under $MD\_TIME\_PAT_j$ and consequently $AVG\_MDTIME\_PAT_j$, mostly resulted from incorrect entry of triage assessment, physician assessment or disposition time. Some of these errors produced negative values caused by entering disposition time as an event occurring prior to nursing triage or physician assessment time. Even after eliminating these errors, some visits had extremely high or low values that were assessed as errors or outliers through consultation with the independent medical expert. For the $MD\_TIME\_PAT_j$, visits with values less than one minute and values greater than six hours were eliminated from the data set.

After consultation with the medical expert, outliers affecting the calculation of the number laboratory tests were established as values greater than 25 laboratory tests per patient visit. These were removed from the dataset.

Visits with missing information regarding the trainee were also eliminated.

Table 3 summarizes the data screening process.

Stage	Initial Number of visits	Number of eliminated visits	Final number of visits
Elimination of visits from excluded presenting complaints	36441	10357	26084
Elimination of visits with outliers, errors or missing data in MD_TIME_PAT	26084	3139	22945
Elimination of visits with no trainee information	22945	47	22898
Elimination of visits with outliers, errors or missing data in laboratory tests	22898	8	22890

Table 3. Data screening process

The resulting final data set of 22,890 visits served as the master data set from which all stratified input/output data sets were constructed.

5.5 DEA Model Used in the Study

In section 3.7 we presented SBM-SWAT VRS DLP*o* model as the selected DEA model for assessing ED MDs' performance. However, DEA models with restricted weights or weights penalized for asymmetry alter the efficient frontiers. This can cause results that are difficult to interpret. For example, a required reduction of an input as proposed by the solution of the model should not exceed the current value of that same input to avoid $x_o - s^{-*}$ becoming negative. Also, in our case, the improved output performance $y_o + s^{+*}$ should not be greater than 1 (since it represents a percentage). To ensure these conditions are met, a set of constraints was added to the original model formulation of SBM-SWAT VRS DLP*o* described in Section 3.7.1.

To avoid that improved input values become negative it must be ensured that $\boldsymbol{x_o - s^- \geq 0}$. Since the formulation defines $\boldsymbol{s^- = \frac{S^-}{t}}$ and $\boldsymbol{t > 0}$, the input slack constraints were defined as $\boldsymbol{S^- \leq x_o t}$.

To avoid that the improved output value becomes greater than 1, it must be ensured that $\boldsymbol{y_o + s^{+*} \leq 1}$. Given that the formulation defines $\boldsymbol{s^+ = \frac{S^+}{t}}$ and $\boldsymbol{t > 0}$, the output slack constraint is defined as $\boldsymbol{S^+ \leq (1 - y_o)t}$.

These constraints can be modified to incorporate more strict rules such as ensuring that input reductions do not exceed a certain percentage of current inputs. However, as restrictions become more demanding the likeliness of infeasibility increases.

The final formulation of the model used in this research is the following:

SBM-SWAT VRS DLP*o*

Min

$$t - \frac{1}{\mathrm{m}} \sum_{\mathrm{i=1}}^{\mathrm{m}} \frac{S_i^-}{x_{io}} \qquad (5.1)$$

Subject to:

$$t + \frac{1}{s}\sum_{r=1}^{s}\frac{S_r^+}{y_{ro}} = \mathbf{1} \quad (5.2)$$

$$\boldsymbol{x}_o t - \boldsymbol{S}^- = \boldsymbol{X\Lambda} + \boldsymbol{\Psi\delta}^- - \boldsymbol{\Psi\omega}^- + \boldsymbol{\Gamma\delta}^{-+} - \boldsymbol{\Gamma\omega}^{-+} \quad (5.3)$$

$$\boldsymbol{y}_o t + \boldsymbol{S}^+ = \boldsymbol{Y\Lambda} + \boldsymbol{\Pi\delta}^+ - \boldsymbol{\Pi\omega}^+ + \boldsymbol{\Delta\delta}^{-+} - \boldsymbol{\Delta\omega}^{-+} \quad (5.4)$$

$$\boldsymbol{1}^t \boldsymbol{\Lambda} = \boldsymbol{1}^t t \quad (5.5)$$

$$\boldsymbol{\delta}^- + \boldsymbol{\omega}^- = \beta \quad (5.6)$$

$$\boldsymbol{\delta}^+ + \boldsymbol{\omega}^+ = \beta \quad (5.7)$$

$$\boldsymbol{\delta}^{-+} + \boldsymbol{\omega}^{-+} = \beta \quad (5.8)$$

$$\boldsymbol{S}^- \leq \boldsymbol{x}_o t \quad (5.9)$$

$$\boldsymbol{S}^+ \leq \mathbf{1}t - \boldsymbol{y}_o t \quad (5.10)$$

$$\boldsymbol{S}^-, \boldsymbol{S}^+, \boldsymbol{\Lambda}, \boldsymbol{\delta}^-, \boldsymbol{\delta}^+, \boldsymbol{\delta}^{-+}, \boldsymbol{\omega}^-, \boldsymbol{\omega}^+, \boldsymbol{\omega}^{-+} \geq 0 \quad (5.11)$$

$$t > 0 \quad (5.12)$$

Introduced modifications are captured through new constraint sets (5.9) and (5.10).

To solve the model, parameters $\boldsymbol{X}$ (inputs) and $\boldsymbol{Y}$ (outputs) are calculated from a data set and the parameter known as symmetry factor β is obtained from a calibration process.

The calibration process involves a sensitivity analysis where the scores produced by models with different values of β are analyzed. As the value of β approaches 0, the asymmetry of weights is higher and compensatory behaviour is not penalized. As the value of β grows, asymmetry of weights is reduced but at the expense of diminished flexibility of weights. Thus, the β value must be selected in such a way that compensation is sufficiently penalized while retaining the flexibility of weights.

Using the master input/output data set (for illustration purposes), Figure 6 illustrates the penalized DEA scores (ξ- β**z**) of the SBM-SWAT VRS LP*o* formulation on the vertical axis for each ED MD at different values of beta in the horizontal axis. ED MDs are represented by each line.

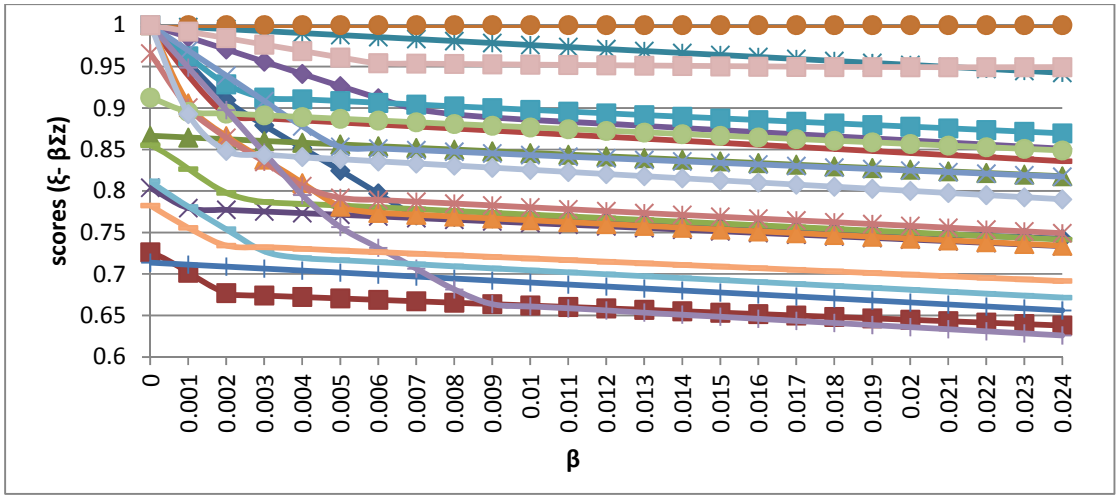

Figure 6. SBM-SWAT scores for each ED MD for different values of β

For some ED MDs, scores were abruptly reduced before reaching a small steady decrease; others do not show any abrupt change in their score at any point. By examining the sum of weight differences represented by $\Sigma\mathbf{z}$ observed at different β values, it was found that weights were modified at transition points where a significant change in slope occurs. After exceeding the β values where these significant changes in slope take place, it can be noticed that the change between consecutive scores is mostly due to changes in the β parameter and not due to weight modifications. In other words, weights stabilize after some point. Thus, the minimum β value at which the slopes of all curves are stabilized (approximately fixed) was defined as the criterion for selecting the most appropriate value of β.

As an example, Table 4 illustrates the score ξ, penalized score $\xi - \beta\Sigma\mathbf{z}$ and $\Sigma\mathbf{z}$ of a ED MD for a set of β values. The last column provides the percentage change relative to previous iteration score.

β	ξ	ξ-βΣz	Σz	% change in score
0.000	0.892	0.892	18.124	N/A
0.003	0.892	0.837	18.124	6.1%
0.006	0.892	0.783	18.124	6.5%
0.009	0.892	0.729	18.124	6.9%
0.012	0.737	0.715	1.874	1.9%
0.015	0.737	0.709	1.874	0.8%
0.018	0.737	0.704	1.874	0.8%
0.021	0.737	0.698	1.874	0.8%
0.024	0.737	0.692	1.874	0.8%

Table 4. Progression of scores (ξ-βΣz) per β for MD13

It can be noted that for β >= 0.012, **Σz** remains constant indicating that weights are unchanged across iterations after a significant drop in percentage change of the score (from 6.9% to 1.9%). Using the evaluation described above, β=0.012 was selected as the symmetry factor for the ED MD in this example. After repeating this procedure for all ED MDs, the maximum of all β values was selected as the model parameter.

5.6 Data Stratifications

As explained earlier, performance evaluation needs to be decomposed into more or less homogenous groups of presenting complaints in order to avoid averaging effect. This implied constructing an input/output data set per complaint group to enable the stratified performance evaluations. Figure 7 illustrates data stratification used in this research. The rest of this section describes how these data sets were constructed and presents how the models' parameters were calibrated.

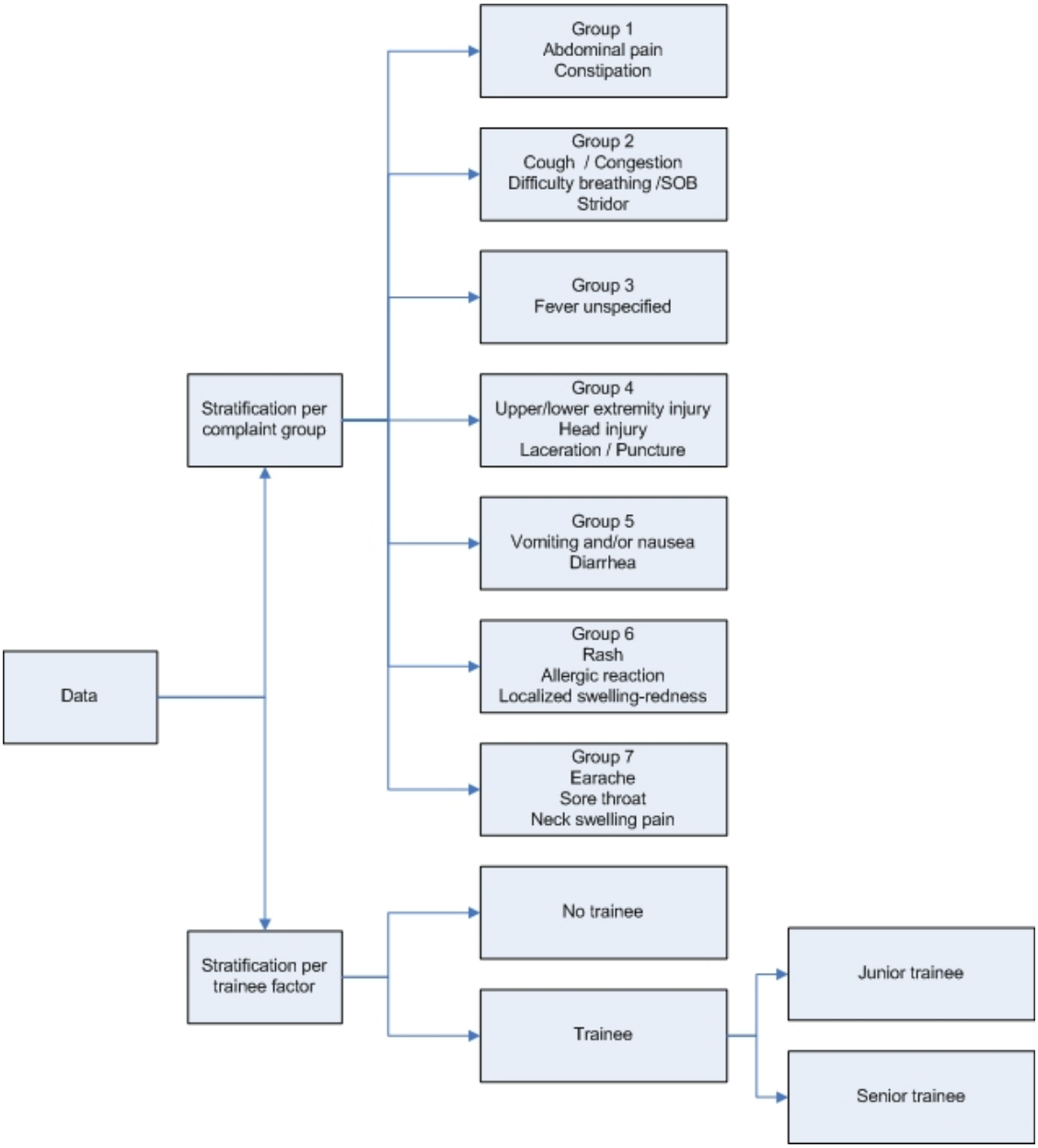

Figure 7. Data stratifications

5.6.1 Complaint Groups

Defining the stratification at the individual complaint level would be difficult and could introduce bias in inputs/outputs calculations due to the small number of visits under some presenting complaints for some ED MDs. This justifies grouping presenting

complaints based on clinical and diagnostic similarities. The complaint groups by which data are stratified comprise the majority of complaints presenting in the PED. Table 5 shows a list and distribution of the 25 most common complaints in CHEO's PED for 2011. Entries in this table are the total number of visits of initial data set (36,441 patient visits) which represent about 80% of all visits assessed by the evaluated ED MDs during 2011.

Complaint	Number of visits	Percent	Cummulative Percent
Fever Unspecified	5356	14.7%	14.7%
Difficulty Breathing/ Shortness of breath (SOB)	3775	10.4%	25.1%
Cough/congestion	2269	6.2%	31.3%
Abdominal Pain	2117	5.8%	37.1%
Upper Extremity Injury	2011	5.5%	42.6%
Vomiting and/or Nausea	1899	5.2%	47.8%
Head Injury	1491	4.1%	51.9%
Lower extremity injury	1330	3.6%	55.6%
Laceration / Puncture	1073	2.9%	58.5%
Rash	1070	2.9%	61.4%
Earache	743	2.0%	63.5%
Diarrhea	615	1.7%	65.2%
Headache	605	1.7%	66.8%
Seizures	604	1.7%	68.5%
Localized Swelling - Redness	517	1.4%	69.9%
Sore Throat	475	1.3%	71.2%
Allergic Reaction	438	1.2%	72.4%
Minor Complaints Unspecified	403	1.1%	73.5%
Urinary tract infection complaints / Symptoms	391	1.1%	74.6%
Stridor	387	1.1%	75.7%
Lower extremity pain	385	1.1%	76.7%
Chest Pain (Non Cardiac Features)	287	0.8%	77.5%
Constipation	278	0.8%	78.3%
Neck Swelling / Pain	240	0.7%	78.9%
Overdose ingestión	237	0.7%	79.6%

Table 5. Distribution of presenting complaints at CHEO's PED (N=36,441)

After consultation with an independent medical expert, 18 out of the 25 complaints with the highest volume of visits were stratified as presented in Table 6. The remaining 7 complaints were excluded from the evaluation. The grouped complaints account for 26,084 visits representing 72% of visits in initial data set and 57% of all visits assessed by the evaluated ED MDs during 2011.

Complaint Group ID	Presenting complaints included
Group 1	- Abdominal pain - Constipation
Group 2	- Cough / congestion - Difficulty breathing/(SOB) - Stridor
Group 3	- Fever unspecified
Group 4	- Upper extremity injury - Lower extremity injury - Head injury - Laceration/puncture
Group 5	- Vomiting and/or nausea - Diarrhea
Group 6	- Rash - Allergic reaction - Localized swelling-redness
Group 7	- Ear ache - Sore throat - Neck swelling/pain

Table 6. Stratification of presenting complaints into complaint groups

Stratification controls for variation in clinical practice since the use of resources for managing complaints within the same group should be similar. Using clinically justified

groupings also minimizes the impact of an incorrect classification of presenting complaint (CEDIS code) since both, correct and incorrect classifications ought to be part of the same complaint group. For example, if a presenting complaint labeled in triage as "cough / congestion" would have been more accurately labeled as "difficulty breathing / SOB", then the visit would be classified in the same complaint group regardless of this inaccuracy.

Table 7 shows the number of visits and their percentage of the total visits for each of the complaint groups.

Group ID	Included Presenting Complaints	Number of visits	Percentage	Cumulative Percentage
2	Cough/congestion, Difficulty breathing/SOB, stridor	6,431	18%	18%
4	Upper extremity injury, lower extremity injury, head injury, laceration/puncture	5,905	16%	34%
3	Fever unspecified	5,356	15%	49%
5	Vomiting and/or Nausea, diarrhea	2,514	7%	55%
1	Abdominal pain, constipation	2,395	7%	62%
6	Rash, allergic reaction, localized swelling – redness	2,025	6%	68%
7	Earache, sore throat, neck swelling/pain	1,458	4%	72%
	Others	10,357	28%	100%

Table 7. Distribution of complaint groups and others (N=36,441)

Before solving the models for each complaint group, a single value for parameter β was selected for use in all models. This allows us to compare the results between the

complaint groups. As a result of the calibration process described in Section 5.5, the value of β=0.021 was selected.

5.6.2 Trainee Factor

For the purpose of this research, trainees are classified into junior and senior categories. Junior trainee refers to medical students and first year residents while senior trainees are second year residents and fellows. In addition to comparing the performance of each ED MD based on the absence and presence of trainees, such classification helps verifying the hypothesis that junior and senior trainees have a different effect on ED MDs' performance.

Each patient visit is associated with a unique trainee factor which we define as a categorical variable representing the following situations:

- No Trainee (N): patient visits in which the ED MDs were not assisted by a trainee of any level.
- Trainee (T): patient visits in which the ED MDs were assisted by any category of trainee.
- Junior (J): patient visits in which the ED MDs were assisted by junior trainees.
- Senior (S): patient visits in which the ED MDs were assisted by senior trainees.

Parameters of the models assessing the trainee factor were calculated from a data set that included trainee information for the situations N, T, J, and S. Table 8 shows that more than half of the visits to CHEO PED (55%) involve assistance of some category of trainee.

Data set ID	Type of trainee assisting PED MD in patient visit	Number of visits	Percentage	Cumulative Percentage
J	Junior trainee	4,822	21%	21%
S	Senior trainee	7,655	33%	55%
N	No trainee	10,413	45%	100%

Table 8. Distribution of visits per trainee factor (N=22,890)

For this analysis, presenting complaint stratification into groups was not considered since we wanted to assess the impact of the trainee factor on overall performance of ED MDs. Disadvantages of such evaluations discussed in Section 1.2 do not apply here since variations in the percentage of visits under each presenting complaint across MDs and the averaging effect are both distributed across the trainee factor. In other words, it is assumed that all groups are similarly affected by these issues, hence reducing the bias in performance comparisons.

To determine if an ED MD works better/worse when working with/without trainees, his/her own performances should be compared by contrasting the scores obtained in each situation. However, considering that DEA scores reflect how near or far an ED MD's performance is from the efficient frontier, the scores of the same ED MD obtained for situation N and for situation T should not be compared. This is because no assumptions regarding the similarity of both efficient frontiers can be made. In order to make a valid comparison, DEA scores obtained under each trainee situation should reflect the distance

relative to the same efficient frontier. To calculate this, the following procedure was used:

1. Data sets N and T were pooled into a single data set resulting in a set of 40 ED MDs (20 assisted by trainees and 20 working without any trainee) treated separately (i.e. ED MD assisted by trainee was treated as different than the same ED MD without assistance).
2. The model was calibrated, new calculations for the β parameter were carried out and a final value was set to β=0.009.
3. The model was solved.
4. Resulting scores for each ED MD under each trainee factor were examined to identify ED MDs whose performance was positively/negatively affected or unaffected by the presence/absence of trainees.

To assess how each ED MD performance was affected by each type of trainee, a similar procedure as described above was followed. In such a case, the pooled set of data sets N, S and J comprised of 60 PED MDs was used with the same value of β parameter.

Chapter 6: Results

This chapter summarizes the main results of the research. The Sections 6.1 and 6.2 present results per ED MD and per complaint group while Section 6.3 presents the impact of the trainee factor.

6.1 Results per ED MD

Table 9 shows the scores derived from the model described in Section 5.5.

ED MDs	G1	G2	G3	G4	G5	G6	G7	GLOBAL
MD1	0.454	1.000	0.753	0.633	0.674	0.493	0.324	0.747
MD2	0.582	0.441	0.489	0.403	0.691	0.300	0.340	0.643
MD3	0.626	0.744	0.623	0.452	0.623	0.396	0.980	0.824
MD4	0.663	0.489	0.544	0.450	0.867	0.431	0.438	0.739
MD5	1.000	0.536	0.654	0.523	1.000	0.673	0.665	0.950
MD6	0.906	0.625	1.000	0.575	0.954	1.000	0.436	1.000
MD7	0.670	0.419	0.552	0.754	0.506	0.369	0.423	0.663
MD8	0.627	0.639	0.834	0.804	0.655	0.366	0.418	0.843
MD9	0.628	0.520	0.642	0.487	0.748	0.383	1.000	0.748
MD10	0.563	0.555	0.848	0.439	0.784	0.799	0.405	0.858
MD11	0.672	0.636	0.669	0.514	0.732	0.487	0.346	0.875
MD12	0.566	0.492	0.811	0.593	0.599	0.302	0.295	0.741
MD13	0.701	0.509	0.635	0.548	0.689	0.391	0.541	0.823
MD14	0.594	0.522	0.604	0.496	0.705	0.290	0.466	0.756
MD15	0.675	0.521	0.670	0.741	0.750	0.962	0.826	0.855
MD16	0.704	0.364	0.482	1.000	0.432	0.297	0.828	0.634
MD17	0.628	0.493	0.517	0.481	0.601	0.316	0.415	0.679
MD18	0.721	0.453	0.558	0.419	0.658	0.267	0.318	0.697
MD19	0.714	0.587	0.632	0.555	0.631	0.522	0.792	0.798
MD20	0.853	0.566	0.744	0.787	0.967	0.630	0.455	0.949

Table 9. Scores per ED MD under each complaint group

Each column represents a complaint group while the last column (labeled "Global") presents the score from a non-stratified global evaluation for comparison.

Results show that ED MDs 1, 5, 6, 9 and 16 are efficient (score equal to 1) for at least one complaint group, with MD5 and MD6 representing benchmark performance in more than one complaint group. Although MD5 and MD6 performances are the highest in a few complaint groups, the rest of their performance scores are mid-range. Thus the group of high performing MDs is not consistent across all groups.

More consistency is observed in the scores for MD2 and MD17 that do not to exceed 0.7 for any complaint group. This trend confirms the presence of overall "underperformers".

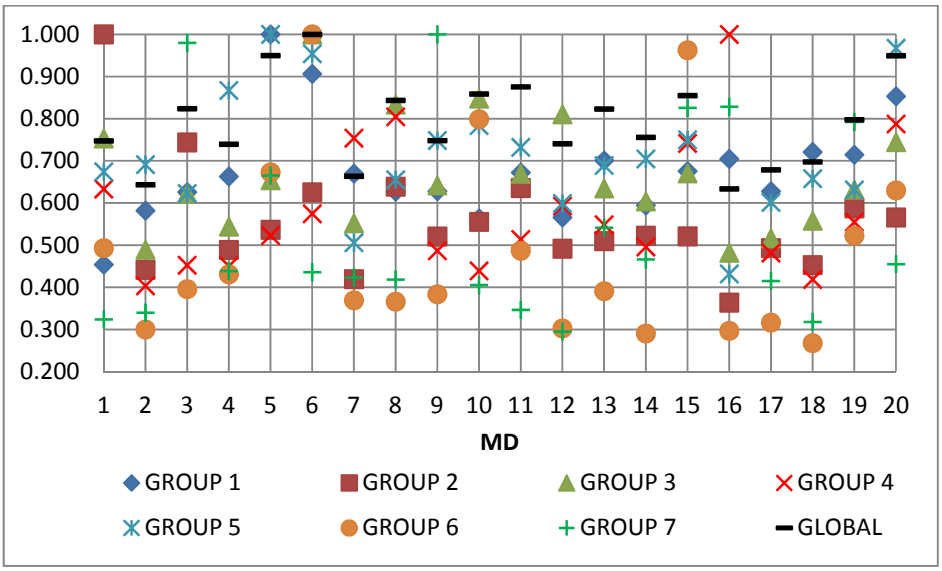

Figure 8. Scores for each ED MD stratified by complaint group

Figure 8 displays the scores for each ED MD for each complaint group, demonstrating that not only does performance differ across ED MDs, but also across groups for a single ED MD. It can also be noticed that some ED MDs show a higher

dispersion of scores across complaint groups than others. For example, MD19 has a low variation of scores compared to MD1 whose scores range from 0.324 to 1.

Table 10 displays descriptive statistics for the results in Table 9. A comparison of the averages per complaint group provides a general idea of the proximity of performances to benchmark. For example, the difference between averages from Group 5 and Group 6 could indicate that most ED MDs' performances in Group 5 are closer to benchmark than in Group 6.

	G1	G2	G3	G4	G5	G6	G7	GLOBAL
Min	0.454	0.364	0.482	0.403	0.432	0.267	0.295	0.634
Max	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Std. deviation	0.124	0.136	0.134	0.158	0.147	0.221	0.227	0.105
Average	0.677	0.556	0.663	0.583	0.713	0.484	0.536	0.791

Table 10. Descriptive statistics of scores per complain group

Presenting the results by complaint groups leads to the identification of groups that deserve more attention and should be prioritized in the development of any improvement plans.

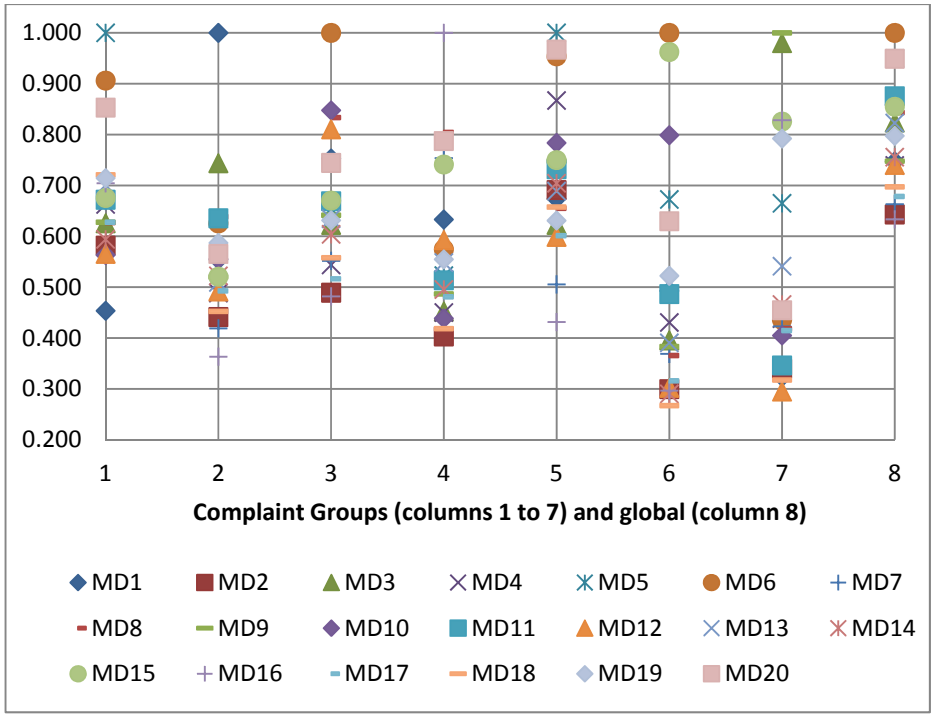

Figure 9. Scores under each complaint group stratified by ED MDs

Figure 9 illustrates from a complaint group perspective, the scores of the 20 MDs. Column 8 displays the scores of an overall evaluation. For some complaint groups, differences among MDs' scores are more evident than for others. For example, scores for Group 3 (fever unspecified: range = 0.518) are less scattered than scores for Group 6 (rash, allergic reaction, localized swelling-redness: range = 0.733).

Scores for Group 2 (Cough congestion, difficulty breathing/Shortness of breath (SOB), stridor) show that MD1 outperforms the rest of the MDs by a large margin. This might be evidence of "specialist" ED MDs that have excelled in managing a particular complaint group. The same is true for Group 4 (upper/lower extremity injury, head injury, laceration/puncture) with MD16 excelling in this group.

To determine if the stratification provided additional information, we compared the stratified evaluations to a global one. To enable this comparison, the model was solved

using the global dataset and the same β value was used as for the complaint group models (β=0.021). These global scores show MD6 as the best MD (Figure 9) but fail to reveal his/her weaknesses when managing complaint Groups 4 and 7. Also, global scores are less scattered than the scores from the stratified evaluations, which most likely is caused by the averaging effect.

To allow a better visualization of results, Figure 10 shows the rankings based on scores where the benchmark is represented by 1 and the worst is represented by 20.

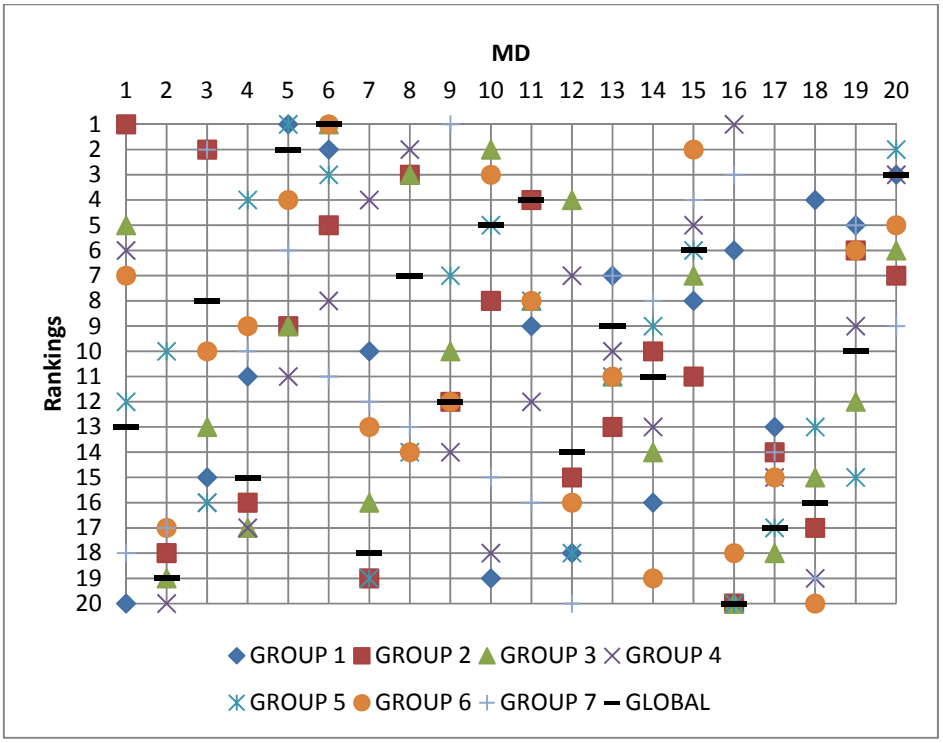

Figure 10. Rankings obtained by each ED MD under each complaint group and global (Best=1, Worst=20)

Figure 10 depicts the ED MDs' rankings calculated from scores with the complaint group and the ranking based on an overall evaluation. It can be noted that rankings from the overall evaluation are not substitutes for complaint group rankings since each ED MD's rankings show variation across groups. For example, MD16 would have been

considered the worst performer under an overall evaluation, yet his/her scores under groups 4, 7 and 1 are in the top six.

6.2 Improvements for each ED MD per Complaint Group

The improvements were calculated from the optimal values for slack variables after solving the model for each complaint group. The value of each slack variable represents the number of units by which the current input should be reduced or the number of units by which the output should be increased in order to achieve benchmark performance. These improvements can be also presented as percentages of current inputs and outputs to reduce/augment. They are calculated by dividing the optimal values of the slacks by their corresponding input/output value.

Improvement information for MD1 under complaint group 1 is shown in Table 11 where each required improvement is calculated by dividing the optimal value of slack variable by the current performance value. Benchmark performance can be calculated by simultaneously modifying all inputs and output by subtracting values of the slacks from their corresponding inputs and by adding the slack to the corresponding output. Another way of calculating the benchmark performance is by subtracting from the current input performance, the required improvement percentage of that current input. In the case of the output, the improvement percentage of the current output should be added to the output.

MD1 (Group 1)	AVG_MD_TIME_PAT	LAB_PAT	RAD_PAT	RATE_NR72
Slacks	0.494	2.733	0.607	0.000
Current Performance	2.026	3.471	1.000	1.000
Required Improvement %	24%	79%	61%	0%
Benchmark Performance	1.532	0.737	0.393	1.00

Table 11. Improvements report for MD1 for complaint group 1

Figure 11 and Figure 12 provide a visualization of the required improvements' represented as percentage for each input and output of MD2, on complaint groups 1 and 3 respectively.

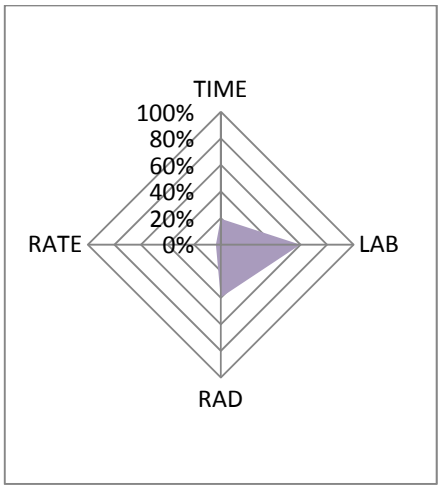

Figure 11. Radial graph of improvements' percentages for MD2 for Complaint Group 1

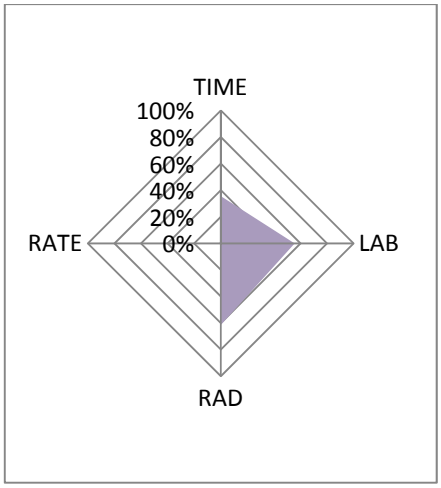

Figure 12. Radial graph of improvements' percentages for MD2 for Complaint Group 3

As seen on these graphs, the level and direction of improvements for an ED MD depends on the complaint group analyzed. For instance, MD2 needs to improve (lower) on the number of laboratory tests ordered (AVG_LAB_PAT labelled as "LAB" in the graph) for complaint group 1 while for complaint group 3 needed improvement involves the number of radiology orders (AVG_RAD_PAT labelled as "RAD" in the graph).

The improvements report can help an ED MD identify what inputs or output deserve the greatest attention which can help determine priorities in the design of personalized improvement plans.

6.3 Assessing the Impact of Trainee Factor on Performance

Figure 13 shows the scores for each ED MD under the presence and absence of trainees. Most ED MDs (16 out of 20) perform better working without any trainees. For some ED MDs such as MD5, MD6 and MD20 there are large differences in performance under each trainee factor while MD8 and MD10 seem to be unaffected by the presence of trainees. So while the impact of the presence of a trainee is generally negative, the magnitude of the impact is not uniform for all ED MDs.

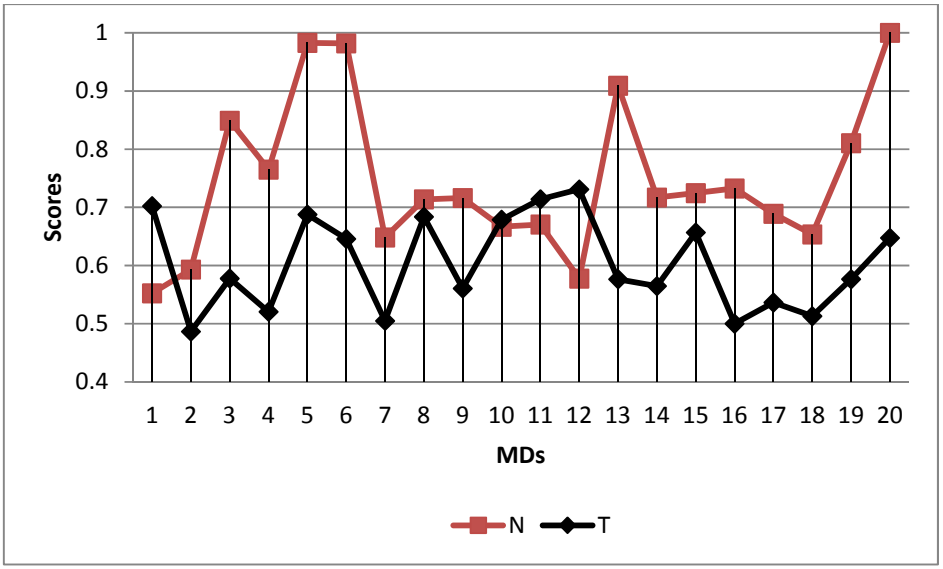

Figure 13. ED MDs' scores per trainee factor (Pooled data sets N and T)

Considering that the normality of the scores cannot be assumed, we used (as advocated by Brocket and Golany (1996)) a non-parametric statistical test to verify if differences in scores between "No trainee" and "Trainee" populations are significant. For two populations the Mann-Whitney test was used whereas in case of more than two populations, the Kruskal-Wallis rank sum test was used.

Results from the Mann-Whitney test for the "No Trainee" and "Trainee" populations give that the level of significance less than 0.001 confirming that the scores are higher for the "No Trainee" population.

To further explore these differences and to verify our hypothesis that assisting junior and senior trainees have different effects on ED MDs' performances, results from a pooled set of data sets N, S and J were plotted.

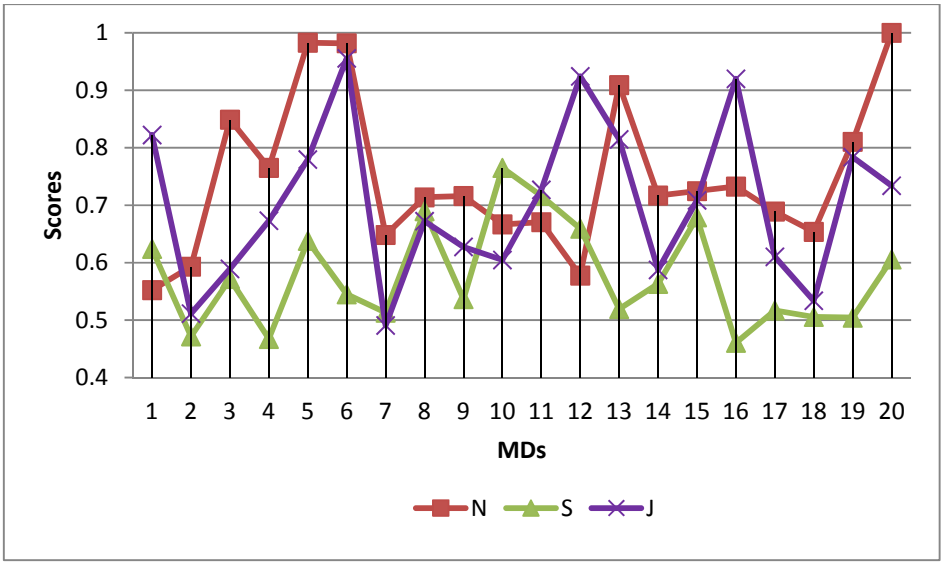

Figure 14. MDs' scores per trainee factor (Pooled data sets N,J and S)

Figure 14 further confirms that there are clear differences between the effects on performance associated with the assistance of junior and senior trainees. It can be noticed that 17 out of 20 ED MDs perform better when assisted by junior trainees than when assisted by senior trainees. Although this is true for most ED MDs, the magnitude of these differences varies across ED MDs. Interestingly, for some ED MDs the presence of junior trainees improves their performance in relation to a situation with no trainees. For example, MD16 was thought to present better performance in the absence of trainees as suggested by the interpretation of Figure 13. However, by examining the separate effects of trainee types in Figure 14, it can be concluded that the performance of MD16 is better when assisted by a junior trainee than when working alone.

Results from the Kruskal-Wallis rank sum test produced a level of significance less than 0.001, confirming score differences among the three populations. To identify the source of this difference, Mann- Whitney tests were run for the No Trainee- Junior

population pair and also for the Junior-Senior population pair. The level of significance were 0.414 and 0.003 for the No Trainee- Junior and the Junior-Senior population pairs respectively. This confirms that statistically performance of ED MDs worsens when senior trainees assist them in patients' visits while this is not a case when ED MDs are assisted by junior trainees.

Chapter 7: Discussion

7.1 Implications of Results

Interpretation of DEA scores should be done carefully to avoid definitive judgments about performances. Due to the benchmarking, low scores should not be necessarily interpreted as representing unsatisfactory performance. DEA scores are calculated in relation to the benchmark practice observed for the specific model setup (study setting, data set used, stratification, etc); therefore their interpretation should not lead to any conclusions regarding the level of compliance to some external standards. The reverse is also true where an efficient DMU is only efficient in comparison with his/her peers.

7.1.1 DEA Model

Generally, results from most DEA models have many DMUs qualified as benchmarks, especially when the sample of evaluated DMUs is relatively small and the compensatory behavior is not penalized. Results obtained from solving the revised DEA model used in this research show a single benchmark per evaluation and a high dispersion of scores; strong evidence of the high discriminatory power of the model and a testimony to the successful mitigation of the compensatory behavior. Appendix A presents an analysis comparing the SBM- SWAT model and the SBM model to support this argument.

7.1.2 ED MD Performance Perspective

Interpretation of results raises questions regarding the reasons behind the variation of scores both analyzed from the ED MD perspective and complaint groups' perspective. From an ED MD perspective, the high variation of scores obtained for each complaint group lead us to suppose that there are considerable differences in medical acumen across the groups. Regardless of obtaining benchmark scores in some complaint group evaluations, an ED MD might still show low scores in other complaint groups. This performance inconsistency across complaint group evaluations is observed in all the evaluated ED MDs, thus, all of them have opportunities to improve their performance.

7.1.3 Complaint Group Perspective

Analyzing ED MDs' scores from the complaint group perspective reveals significant variation of ED MDs' performances. Within a single complaint group, the presence of gaps between the benchmark ED MD and the rest helps identify "specialists" that are particularly adept in managing a given set of complaints. The comparison of averages across groups leads to the identification of "problematic" complaint groups where performance of most ED MDs is lower. Dispersion of the scores within a group might indicate a lack of standard of practice (or non-adherence to a standard if present).

7.1.4 Impact of Trainee Factor on Performance of ED MDs

The analysis of the trainee factor demonstrated that teaching often hinders the performance of ED MDs, especially when senior trainees are present. The most probable explanation for this phenomenon is the difference in the level of responsibility that ED MDs delegate to each category of trainee. It is hypothesized that when junior trainees are assisting the ED MDs they are not independent enough to make their own decisions as

the attending ED MD is responsible for patient assessment. On the other hand, senior trainees are more independent and are assigned tasks to do on their own. Due to their limited experience, they might use more resources and also their work may need to be more thoroughly evaluated by the attending ED MD, thus making patient management less efficient.

7.1.5 Stratification Approach

All the above insights could have only been inferred because the stratification was applied. A single overall evaluation would not have captured the inconsistency of performance across complaint groups and would not have facilitated the identification of complaint-specific improvements of inputs and outputs. The granularity of results from the stratified design aids the development of individualized improvement plans for each ED MD and for each complaint group.

7.1.6 Management Perspective

Assuming a performance improvement framework that allows confidential communication of evaluation results to individual ED MDs, the results could be used to identify the best performers for each complaint group to facilitate the development of a practice-oriented model. This can promote reduction of the performance gap between best and worst performers. High dispersion of scores for the same complaint group should prompt investigation as to whether the guidelines for standard practice for the complaint group are used as a tool to reduce variance in patient management.

Although modifying the clinical practice of ED MDs is a challenge, improvement information for each input/output and scores under each complaint group provide a

reference to establish priorities of performance improvement. These improvements could be defined as dimension-oriented (what input/output of the ED MD needs the greatest improvement?) or complaint-oriented (for what complaint group does the ED MD need the greatest improvement?).

In the case of dimension-oriented improvements, clinical managers must be aware of the possible interactions between inputs and outputs when setting improvement goals using the results from our model. For example, an unintended effect such as worsening the rate of return visits after 72 hours of discharge due to an improvement (decrease) of patient encounter time is a possibility. If improvement goals are set for more than one measure, it is not realistic to expect success for all measures due to inherent trade-offs between them. Thus, improvement goals should involve a single performance measure and once expected result is achieved, a model should be re-run in order to establish new scores and therefore new improvement targets.

Trade-offs between teaching and clinical tasks unveiled by the assessment of the trainee factor on ED MDs' performances should trigger investigation of the clinical practice causes. Once these factors are understood they should be considered in the design of a system by which trainees are assigned to ED MDs in order to align clinical and teaching performance objectives.

7.1.7 Summary of Implications

In summary, results of the SBM-SWAT model used in this research are useful to define benchmark, to define priorities for performance improvement and to determine the specific practice modifications leading to best practice for each ED MD for each of the complaint groups.

7.2 Assumptions/Limitations

Our model assumes un-controllable factors such as type of shifts, complaint mix (percentage of patient visits for each complaint) within the same complaint group, severity of visits and exposure to seasonal diseases are uniform across ED MDs. However, these factors were not examined in detail to verify if they would not introduce bias. In principle, there are no reasons to believe that there are considerable differences in patient mix and work conditions across ED MDs therefore this assumption should not compromise the validity of the results.

According to the recommendations presented in Section 4.1 regarding the number of desired inputs and outputs in relation to the number of evaluated DMUs, the sample size in our study would allowed for at most two more inputs/outputs. However, the availability of data precluded the use of more performance measures. In the case of the outputs, we recognize that the use of a single measure limits our ability to further discriminate between ED MDs' performances. Other measures such as guideline compliance, patient satisfaction and patient throughput would have provided additional insights. To assess how well each ED MD complied with practice guidelines when managing their patients, guidelines need to be in place for all presenting complaints considered in the study and patient chart audits would have been necessary. While standards of practice were not always available, the chart audits required additional resources that were not available. For the modelling purposes information about patient satisfaction would need to be associated with a specific patient, so it can be attributed to a specific encounter. However, such information was not available. While data about the

throughput of patients was available, it could not be decomposed per a complaint group and therefore this output measure could not be used with a stratified approach. .

The use of discharge diagnoses instead of presenting complaint at nursing triage as the attribute used to develop the complaint groups could have improved the results since each stratum will be composed of hopefully more homogeneous mix of patients.

Chapter 8: Conclusions

To the best of our knowledge our research is the first that evaluates the performance of ED MDs using measures of timeliness, resource utilization and quality of care. It is also the first to develop a performance assessment model that does not rely on subjective judgements of the relative importance of each measure.

A methodological contribution of this research is that it proposes an extension of a commonly used DEA model that controls for the compensatory behaviour linked to the asymmetric assignment of weights to inputs/outputs. Although the SWAT method has been used in oriented DEA models (input-oriented and output-oriented), our proposed model is the first non-oriented DEA model that deals with the problem of asymmetric weight assignment without the need for defining weight limits.

The results were evaluated by an independent medical expert who provided potential explanations for some of the findings and confirmed that the results are reasonable. However, the validity of the DEA model used in the study is yet to be assessed due to the absence of a known standard of best practice for ED MDs. Strategies to consider when validating the model include:

- Conducting chart audits to verify if there are differences in the management of patients between “high performing” ED MDs and those with lower rated performance.

- Requesting a feedback from ED MDs to contrast self-assessments of their performance against the results produced by our model.

Once validated, the model could be used for continuing performance evaluations to assess ED MDs' improvement progress. Due to the large sample of patient visits required to generate a representative input/output dataset, it is not advisable to run more than one evaluation per year. However, improvement goals set in the evaluation should be frequently monitored to provide timely feedback to ED MDs in order to guide their efforts.

Future research should focus on the follow up of improvement goals to verify if performance targets derived from model's results represent achievable changes in medical practice.

Appendix A: Comparing SBM and SBM-SWAT models

To carry out an analysis on how SBM and SBM-SWAT models deal with low discrimination and compensation issues, input/output based rankings and scores for each MD under each model were compared. Table 12 shows these rankings where lower values equate to better performance.

Comparing performance discrimination between models, it can be noted that 11 out of 20 MDs were deemed efficient by the SBM model. This would lead one to think that all the efficient DMUs have a similar performance, however an examination of the rankings show great differences among SBM efficient MDs. For example, judging by the rankings, MD 6 has clearly a better performance than MD 16, yet their scores under the SBM model are equal.

In the SBM- SWAT model, only one out of the 20 MDs was deemed efficient and the difference in performance between MD6 and MD16 became evident. In terms of performance discrimination it can be concluded that the SBM-SWA model outperforms the SBM model.

	I1	I2	I3	O1	SBM Score	SBM-SWAT Score	% Change
MD1	8	13	18	2	1	0.747	-25.30%
MD2	19	17	19	11	0.7261	0.643	-11.44%
MD3	13	7	8	20	0.8665	0.824	-4.90%
MD4	15	10	15	14	0.804	0.739	-8.08%
MD5	10	1	1	16	1	0.95	-5.00%
MD6	2	3	2	17	1	1	0.00%
MD7	18	19	5	18	0.7137	0.663	-7.10%
MD8	5	9	4	9	1	0.843	-15.70%
MD9	9	15	9	8	0.8553	0.748	-12.55%
MD10	3	2	20	12	1	0.858	-14.20%
MD11	4	4	14	7	1	0.875	-12.50%
MD12	12	16	10	3	1	0.741	-25.90%
MD13	7	8	13	4	1	0.823	-17.70%
MD14	14	11	7	5	0.9658	0.756	-21.72%
MD15	6	6	16	15	0.9127	0.855	-6.32%
MD16	20	20	11	1	1	0.634	-36.60%
MD17	16	18	6	6	0.8118	0.679	-16.36%
MD18	17	14	17	13	0.7825	0.697	-10.93%
MD19	11	12	3	10	1	0.798	-20.20%
MD20	1	5	12	19	1	0.949	-5.10%

Table 12. Ranking of MDs for each Input and Output (Best = 1, Worst = 20)

To assess how each model deals with the compensation issue, we identified MDs who achieved a score of one on the SBM model while compensating low performances with a single high performance input / output. MD12 and MD16 are good examples of compensatory behavior since both of them have input values between the 10^{th} and 20^{th} rank and their respective output is within the top three. Their SBM-SWA scores were considerably reduced (in comparison to their SBM scores) due to the penalization function in the SBM-SWA model. It can be further noted that the scores of MD12 and MD16 have been reduced by 25.9% and 36.6% respectively due to the higher degree of compensation practiced by MD16. While having similar high performing output, inputs

for MD16 are closer to the worst rankings possible compared to inputs for MD12, which can be interpreted as a higher compensation for MD16.

This analysis shows that the SBM-SWA performs better than the SBM model since it deals satisfactorily with the compensation and performance discrimination issues that are known to plague DEA models.

Appendix B: Inputs and Outputs per ED MD

This appendix presents inputs and outputs for each evaluated ED MD_j on each complaint group, trainee factor group and global. The inputs displayed are the average encounter time per patient visit in hours ($AVG\_MDTIME\_PAT_j$), average number of laboratory tests per patient visit ($AVG\_LAB\_PAT_j$) and average number of radiology orders per patient visit ($AVG\_RAD\_PAT_j$). The output is the rate of non-return patient visits within 72 hours ($RATE\_NR72_j$).

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	2.026	2.760	0.920	1.000
MD2	1.959	2.381	0.774	0.961
MD3	2.223	2.333	0.643	0.905
MD4	1.884	1.823	0.661	0.952
MD5	1.511	0.857	0.487	0.952
MD6	1.456	1.330	0.648	0.978
MD7	1.903	1.877	0.596	0.956
MD8	1.704	1.730	0.678	0.939
MD9	1.708	1.927	0.657	0.968
MD10	1.979	1.508	0.820	0.922
MD11	1.652	1.618	0.592	0.981
MD12	2.169	1.863	0.608	0.961
MD13	1.634	1.538	0.786	0.979
MD14	1.745	2.117	0.738	0.942
MD15	1.594	1.548	0.602	0.957
MD16	2.311	1.538	0.462	0.974
MD17	1.962	1.748	0.557	0.948
MD18	1.804	1.590	0.723	0.977
MD19	1.567	1.487	0.601	0.937
MD20	1.435	1.198	0.568	0.969

Min	1.435	0.857	0.462	0.905
Max	2.311	2.760	0.920	1.000
Mean	1.811	1.739	0.656	0.958
Std. Deviation	0.254	0.431	0.112	0.022

Table 13. Data and descriptive statistics for inputs and output on complaint group 1 (abdominal pain and constipation)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.151	0.000	0.340	0.980
MD2	2.241	0.336	0.388	0.966
MD3	1.729	0.067	0.258	0.942
MD4	2.046	0.098	0.390	0.976
MD5	1.890	0.243	0.306	0.957
MD6	1.544	0.157	0.258	0.938
MD7	2.205	0.411	0.384	0.949
MD8	1.491	0.197	0.269	0.977
MD9	1.947	0.191	0.322	0.959
MD10	1.458	0.327	0.392	0.965
MD11	1.362	0.034	0.316	0.959
MD12	1.845	0.206	0.364	0.935
MD13	1.720	0.205	0.363	0.963
MD14	1.854	0.311	0.330	0.966
MD15	1.838	0.135	0.329	0.956
MD16	2.387	0.625	0.500	0.982
MD17	2.040	0.264	0.331	0.951
MD18	2.081	0.286	0.412	0.953
MD19	1.896	0.317	0.261	0.966
MD20	1.568	0.196	0.299	0.935

Min	1.151	0.000	0.258	0.935
Max	2.387	0.625	0.500	0.982
Mean	1.815	0.230	0.341	0.959
Std. Deviation	0.316	0.142	0.061	0.014

Table 14. Data and descriptive statistics for inputs and output on complaint group 2 (cough / congestion, difficulty breathing/(SOB), stridor)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.639	0.604	0.333	1.000
MD2	1.682	1.031	0.374	0.969
MD3	1.386	0.551	0.318	0.907
MD4	1.482	0.600	0.419	0.943
MD5	1.362	0.561	0.305	0.952
MD6	1.017	0.496	0.207	0.953
MD7	1.457	0.934	0.316	0.969
MD8	1.084	0.632	0.212	0.964
MD9	1.223	0.751	0.279	0.959
MD10	1.140	0.357	0.260	0.959
MD11	1.538	0.384	0.299	0.943
MD12	1.061	0.407	0.407	0.966
MD13	1.255	0.730	0.340	0.977
MD14	1.473	0.659	0.388	0.976
MD15	1.265	0.581	0.372	0.977
MD16	1.752	0.912	0.412	0.985
MD17	1.571	1.101	0.314	0.977
MD18	1.597	0.772	0.308	0.965
MD19	1.306	0.743	0.273	0.970
MD20	1.044	0.549	0.302	0.941

Min	1.017	0.357	0.207	0.907
Max	1.752	1.101	0.419	1.000
Mean	1.367	0.668	0.322	0.963
Std. Deviation	0.227	0.206	0.061	0.020

Table 15. Data and descriptive statistics for inputs and output on complaint group 3 (fever unspecified)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.293	0.000	0.699	0.957
MD2	1.287	0.166	0.847	0.983
MD3	1.123	0.030	0.723	0.970
MD4	1.122	0.115	0.803	1.000
MD5	1.050	0.021	0.609	0.979
MD6	0.914	0.021	0.689	0.992
MD7	1.056	0.108	0.652	0.990
MD8	0.950	0.000	0.728	0.981
MD9	1.027	0.090	0.754	0.983
MD10	1.173	0.024	0.778	0.986
MD11	1.046	0.020	0.654	0.986
MD12	0.943	0.074	0.595	0.992
MD13	0.995	0.052	0.617	0.991
MD14	1.139	0.176	0.617	0.991
MD15	0.852	0.090	0.639	0.976
MD16	0.988	0.000	0.478	1.000
MD17	1.092	0.127	0.756	0.991
MD18	1.264	0.110	0.793	0.984
MD19	1.010	0.109	0.592	0.990
MD20	0.836	0.085	0.667	0.977

Min	0.836	0.000	0.478	0.957
Max	1.293	0.176	0.847	1.000
Mean	1.058	0.071	0.684	0.985
Std. Deviation	0.132	0.055	0.090	0.010

Table 16. Data and descriptive statistics for inputs and output on complaint group 4 (upper extremity injury, lower extremity injury, head injury, laceration/puncture)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	2.210	1.292	0.375	1.000
MD2	1.891	1.104	0.327	0.964
MD3	1.923	1.139	0.162	0.892
MD4	1.751	1.446	0.231	1.000
MD5	1.596	0.616	0.232	0.949
MD6	1.396	0.993	0.151	0.957
MD7	2.233	2.150	0.188	0.941
MD8	1.631	1.397	0.198	0.983
MD9	1.450	1.512	0.289	0.980
MD10	1.486	0.808	0.158	0.933
MD11	1.830	0.709	0.183	0.983
MD12	1.960	1.377	0.242	0.968
MD13	1.646	1.160	0.393	0.982
MD14	1.780	1.307	0.246	0.974
MD15	1.989	1.071	0.246	0.947
MD16	2.459	2.152	0.242	0.970
MD17	1.854	1.370	0.309	0.973
MD18	1.907	1.400	0.221	0.964
MD19	1.769	1.125	0.270	0.971
MD20	1.404	0.743	0.233	0.961

Min	1.396	0.616	0.151	0.892
Max	2.459	2.152	0.393	1.000
Mean	1.808	1.243	0.245	0.965
Std. Deviation	0.284	0.405	0.067	0.025

Table 17. Data and descriptive statistics for inputs and output on complaint group 5 (vomiting and/or nausea, diarrhea)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.118	0.128	0.179	1.000
MD2	1.613	0.443	0.180	0.959
MD3	1.161	0.216	0.243	0.973
MD4	1.044	0.560	0.080	0.960
MD5	1.292	0.151	0.053	0.987
MD6	0.987	0.000	0.167	0.958
MD7	1.464	0.378	0.061	0.959
MD8	1.347	0.494	0.130	0.974
MD9	1.422	0.534	0.073	0.990
MD10	0.915	0.275	0.108	0.983
MD11	1.008	0.246	0.067	0.959
MD12	1.647	0.528	0.057	1.000
MD13	1.219	0.468	0.090	1.000
MD14	1.751	0.477	0.080	0.955
MD15	0.911	0.161	0.046	0.977
MD16	1.701	0.900	0.100	1.000
MD17	1.612	0.452	0.075	0.989
MD18	1.707	0.564	0.105	0.955
MD19	1.174	0.117	0.117	0.975
MD20	1.081	0.123	0.090	0.987

Min	0.911	0.000	0.046	0.955
Max	1.751	0.900	0.243	1.000
Mean	1.309	0.361	0.105	0.977
Std. Deviation	0.287	0.219	0.052	0.017

Table 18. Data and descriptive statistics for inputs and output on complaint group 6 (rash, allergic reaction, localized swelling-redness)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	0.890	0.931	0.345	1.000
MD2	1.015	0.407	0.174	0.977
MD3	0.640	0.000	0.185	1.000
MD4	0.881	0.174	0.130	0.957
MD5	0.776	0.126	0.081	0.973
MD6	0.728	0.508	0.175	0.984
MD7	0.765	0.541	0.180	1.000
MD8	0.828	0.227	0.227	0.970
MD9	0.602	0.248	0.120	1.000
MD10	0.880	0.178	0.218	0.970
MD11	0.902	0.664	0.248	0.976
MD12	1.121	0.972	0.194	1.000
MD13	0.681	0.253	0.115	0.989
MD14	0.837	0.433	0.090	1.000
MD15	0.650	0.000	0.250	0.942
MD16	0.856	0.000	0.190	1.000
MD17	0.800	0.464	0.214	1.000
MD18	1.073	0.564	0.209	0.991
MD19	0.683	0.165	0.224	0.965
MD20	0.741	0.331	0.181	0.984

Min	0.602	0.000	0.081	0.942
Max	1.121	0.972	0.345	1.000
Mean	0.817	0.359	0.188	0.984
Std. Deviation	0.141	0.281	0.061	0.017

Table 19. Data and descriptive statistics for inputs and output on complaint group 7 (earache, sore throat, neck swelling/pain)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.376	0.664	0.528	0.984
MD2	1.397	0.488	0.480	0.973
MD3	1.092	0.246	0.377	0.958
MD4	1.242	0.265	0.402	0.974
MD5	1.151	0.196	0.293	0.982
MD6	1.007	0.283	0.308	0.968
MD7	1.377	0.571	0.333	0.959
MD8	1.203	0.496	0.348	0.984
MD9	1.195	0.397	0.382	0.971
MD10	1.220	0.341	0.507	0.965
MD11	1.285	0.371	0.442	0.971
MD12	1.321	0.901	0.428	0.986
MD13	1.009	0.276	0.376	0.991
MD14	1.341	0.346	0.369	0.978
MD15	1.116	0.357	0.444	0.961
MD16	1.621	0.231	0.410	0.991
MD17	1.398	0.385	0.364	0.976
MD18	1.351	0.386	0.438	0.975
MD19	1.197	0.338	0.300	0.976
MD20	0.884	0.209	0.434	0.969

Min	0.884	0.196	0.293	0.958
Max	1.621	0.901	0.528	0.991
Mean	1.239	0.387	0.398	0.975
Std. Deviation	0.170	0.171	0.066	0.010

Table 20. Data and descriptive statistics for inputs and output from visits without any trainee assistance (No Trainee)

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.335	0.225	0.380	1.000
MD2	1.970	0.554	0.437	0.967
MD3	1.625	0.505	0.358	0.937
MD4	1.811	0.278	0.351	0.959
MD5	1.582	0.249	0.297	0.962
MD6	1.309	0.168	0.279	0.966
MD7	1.786	0.790	0.409	0.932
MD8	1.465	0.445	0.324	0.960
MD9	1.628	0.434	0.362	0.970
MD10	1.466	0.422	0.456	0.966
MD11	1.412	0.322	0.344	0.971
MD12	1.181	0.188	0.344	0.990
MD13	1.424	0.241	0.327	0.988
MD14	1.904	0.440	0.356	0.958
MD15	1.773	0.160	0.469	0.926
MD16	2.033	0.368	0.184	0.974
MD17	1.798	0.458	0.366	0.991
MD18	2.000	0.696	0.333	0.946
MD19	1.462	0.323	0.270	0.965
MD20	1.308	0.333	0.343	0.958

Min	1.181	0.160	0.184	0.926
Max	2.033	0.790	0.469	1.000
Mean	1.614	0.380	0.349	0.964
Std. Deviation	0.258	0.169	0.065	0.019

Table 21. Data and descriptive statistics for inputs and output from visits assisted by Junior trainees

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.517	0.526	0.351	0.965
MD2	2.044	0.716	0.451	0.967
MD3	1.755	0.456	0.381	0.939
MD4	1.919	0.960	0.427	0.960
MD5	1.507	0.422	0.363	0.952
MD6	1.524	0.740	0.405	0.950
MD7	2.004	0.761	0.381	0.981
MD8	1.404	0.435	0.316	0.955
MD9	1.667	0.844	0.381	0.969
MD10	1.220	0.233	0.481	0.981
MD11	1.161	0.419	0.388	0.969
MD12	1.773	0.346	0.333	0.962
MD13	1.657	0.672	0.448	0.960
MD14	1.631	0.729	0.381	0.984
MD15	1.605	0.465	0.302	0.982
MD16	2.034	1.317	0.413	0.992
MD17	1.749	0.941	0.376	0.962
MD18	1.954	0.643	0.425	0.972
MD19	1.743	0.856	0.414	0.964
MD20	1.421	0.600	0.362	0.947

Min	1.161	0.233	0.302	0.939
Max	2.044	1.317	0.481	0.992
Mean	1.664	0.654	0.389	0.966
Std. Deviation	0.254	0.256	0.046	0.014

Table 22. Data and descriptive statistics for inputs and output from visits assisted by Senior trainees

MDS	AVG_MDTIME_PAT$_j$	AVG_LAB_PAT$_j$	AVG_RAD_PAT$_j$	RATE_NR72$_j$
MD1	1.416	0.359	0.367	0.984
MD2	2.014	0.650	0.445	0.967
MD3	1.704	0.475	0.372	0.938
MD4	1.872	0.661	0.394	0.959
MD5	1.540	0.347	0.334	0.957
MD6	1.429	0.487	0.349	0.957
MD7	1.932	0.771	0.390	0.965
MD8	1.424	0.438	0.319	0.956
MD9	1.653	0.692	0.374	0.970
MD10	1.324	0.313	0.470	0.975
MD11	1.295	0.367	0.364	0.970
MD12	1.548	0.286	0.337	0.972
MD13	1.585	0.539	0.411	0.969
MD14	1.723	0.631	0.372	0.975
MD15	1.639	0.404	0.335	0.970
MD16	2.033	1.098	0.360	0.988
MD17	1.765	0.785	0.372	0.971
MD18	1.971	0.662	0.392	0.963
MD19	1.634	0.649	0.358	0.964
MD20	1.374	0.491	0.354	0.951

Min	1.295	0.286	0.319	0.938
Max	2.033	1.098	0.470	0.988
Mean	1.644	0.555	0.374	0.966
Std. Deviation	0.232	0.200	0.037	0.011

Table 23. Data and descriptive statistics for inputs and output from visits assisted by Junior or Senior trainees (Trainee)

MDS	AVG_MDTIME_PAT$_i$	AVG_LAB_PAT$_i$	AVG_RAD_PAT$_i$	RATE_NR72$_i$
MD1	1.439	0.524	0.473	0.988
MD2	1.726	1.013	0.471	0.973
MD3	1.451	0.479	0.353	0.945
MD4	1.590	0.854	0.391	0.973
MD5	1.471	0.494	0.315	0.964
MD6	1.214	0.517	0.315	0.964
MD7	1.623	0.838	0.368	0.965
MD8	1.378	0.716	0.328	0.973
MD9	1.486	0.804	0.382	0.971
MD10	1.285	0.498	0.470	0.971
MD11	1.293	0.398	0.377	0.974
MD12	1.479	0.659	0.352	0.975
MD13	1.395	0.640	0.408	0.978
MD14	1.543	0.846	0.369	0.978
MD15	1.342	0.554	0.405	0.969
MD16	1.937	0.899	0.398	0.985
MD17	1.663	0.911	0.375	0.976
MD18	1.675	0.816	0.428	0.971
MD19	1.463	0.708	0.338	0.973
MD20	1.195	0.532	0.356	0.967

Min	1.195	0.398	0.315	0.945
Max	1.937	1.013	0.473	0.988
Mean	1.482	0.685	0.384	0.972
Std. Deviation	0.184	0.179	0.048	0.009

Table 24. Data and descriptive statistics for inputs and output from all visits (Global)

Bibliography

Allen, R., Athanassopoulos, A., Dyson, R. G., & Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: evolution, development and future directions. *Annals of Operations Research*,*73*, 13-34.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science (Pre-1986), 30(9), 1078-1078

Baum, S. A., & Rubenstein, L. Z. (1987). Old people in the emergency room: age-related differences in emergency department use and care. *J Am Geriatr Soc*, *35*(5), 398-404.

Brockett, P. L., & Golany, B. (1996). Using rank statistics for determining programmatic efficiency differences in data envelopment analysis. *Management Science*, *42*(3), 466-472.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research, 2*(6), 429-444.

Charnes, A., Cooper, W. W., Golany, B., Seiford, L., & Stutz, J. (1985). Foundations of data envelopment analysis for pareto-koopmans efficient empirical production functions. *Journal of Econometrics, 30*(1-2), 91-107.

Cherchye, L., & Vermeulen, F. (2006). Robust Rankings of Multidimensional Performances An Application to Tour de France Racing Cyclists. *Journal of Sports Economics*, *7*(4), 359-373.

Chilingerian, J. A., & Sherman, H. D. (1996). Benchmarking physician practice patterns with DEA: A multi-stage approach for cost containment. *Annals of Operations Research, 67*, 83-116.

Collier, D. A., Collier, C. E., & Kelly, T. M. (2006). Benchmarking physician performance, part 1. *Journal of Medical Practice Management, 21*(4), 185-189.

Collier, D. A., Collier, C. E., & Kelly, T. M. (2006). Benchmarking physician performance, part 2. *The Journal of Medical Practice Management : MPM., 21*(5), 273-279.

Cooper, W. W., Seiford, L. M., & Tone, K. (2007). Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. Springer.

Dimitrov, S., & Sutton, W. (2010). Promoting symmetric weight selection in data envelopment analysis: A penalty function approach. *European Journal of Operational Research*, *200*(1), 281-288.

Dubinsky, I., Jennings, K., Greengarten, M., & Brans, A. (2010). 360-degree physician performance assessment. *Healthcare Quarterly*, *13*(2), 71-76.

Eitel, D. R., Rudkin, S. E., Malvehy, M. A., Killeen, J. P., & Pines, J. M. (2010). Improving service quality by understanding emergency department flow: a White Paper and position statement prepared for the American Academy of Emergency Medicine. *The Journal of emergency medicine*, *38*(1), 70-79.

Glickman, S. W., Schulman, K. A., Peterson, E. D., Hocker, M. B., & Cairns, C. B. (2008). Evidence-based perspectives on pay for performance and quality of patient care and outcomes in emergency medicine. *Annals of Emergency Medicine, 51*(5), 622-631.

Goulet, F., Jacques, A., Gagnon, R., Bourbeau, D., Laberge, D., Melanson, J., ... & Rivest, R. (2002). Performance assessment. Family physicians in Montreal meet the mark!. *Canadian family physician*, *48*(8), 1337-1344.

Hall, W., Violato, C., Lewkonia, R., Lockyer, J., Fidler, H., Toews, J., ... & Moores, D. (1999). Assessment of physician performance in Alberta the physician achievement review. *Canadian Medical Association Journal*, *161*(1), 52-57.

Hess, B. J., Weng, W., Lynn, L. A., Holmboe, E. S., & Lipner, R. S. (2011). Setting a fair performance standard for physicians' quality of patient care. *Journal of General Internal Medicine, 26*(5), 467-473.

Hung, G. R., & Chalut, D. (2008). A consensus-established set of important indicators of pediatric emergency department performance. *Pediatric Emergency Care, 24*(1), 9-15.

Knox Lovell, C. A., & Pastor, J. T. (1995). Units invariant and translation invariant DEA models. *Operations Research Letters*, *18*(3), 147-151.

Lowthian, J., & Cameron, P. (2012). Improving timeliness while improving the quality of emergency department care. *Emergency Medicine Australasia*, *24*(3), 219-221.

Moers, F. (2005). Discretion and bias in performance evaluation: the impact of diversity and subjectivity. *Accounting, Organizations and Society*, *30*(1), 67-80.

Mulley, A. J. (1990). Methodological issues in the application of effectiveness and outcomes research to clinical practice. *Effectiveness and outcomes in health care*.

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

Ozcan, Y. A. (2007). Health care benchmarking and performance evaluation: an assessment using Data Envelopment Analysis (DEA) (Vol. 120). Springer Verlag.

Pai, C. W., Ozcan, Y. A., & Jiang, H. J. (2000). Regional variation in physician practice pattern: an examination of technical and cost efficiency for treating sinusitis. *Journal of Medical Systems*, *24*(2), 103-117.

Rethans, J. J., Sturmans, F., Drop, R., Van Der Vleuten, C., & Hobus, P. (1991). Does competence of general practitioners predict their performance? Comparison between examination setting and actual practice. *BMJ: British Medical Journal*, *303*(6814), 1377.

Saaty, T. L. (1994). How to make a decision: the analytic hierarchy process.*Interfaces*, *24*(6), 19-43.

Smith, C. A., Varkey, A. B., Evans, A. T., & Reilly, B. M. (2004). Evaluating the performance of inpatient attending physicians: A new instrument for today's teaching hospitals. *Journal of General Internal Medicine, 19*(7), 766-771.

Tone, K. (2001). Slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research, 130*(3), 498-509.

Van Ours, J. C., & Vermeulen, F. (2007). Ranking Dutch economists. *De Economist*, *155*(4), 469-487.

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

Weber, R. S., Lewis, C. M., Eastman, S. D., Hanna, E. Y., Akiwumi, O., Hessel, A. C., et al. (2010). Quality and performance indicators in an academic department of head and neck surgery. *Archives of Otolaryngology - Head and Neck Surgery, 136*(12), 1212-1218.

Welch, S. J., Asplin, B. R., Stone-Griffith, S., Davidson, S. J., Augustine, J., & Schuur, J. (2011). Emergency department operational metrics, measures and definitions: Results of the second performance measures and benchmarking summit. *Annals of Emergency Medicine, 58*(1), 33-40