Quality-adjusted Efficiency Measures for Public Hospitals

Chun Lok K. Li

Submitted in total fulfillment of the requirements of the degree of Doctor of Philosophy

January 2014

Faculty of Business and Economics

The University of Melbourne
Abstract

Public hospital expenditures are a major cause of rising healthcare costs. Many remedies have been proposed, but a widespread concern is that some of these ideas may lower quality of care. In the literature there are many studies that discuss this very issue, but considerably less attention has been devoted to the way hospital quality should be estimated. This is important because inaccurate estimates of hospital quality may distort results for empirical studies that incorporate quality in the analysis.

This thesis contributes to the literature on estimating hospital quality by focusing on three specific cases. The first part pertains to the use of the standardised incidence ratio (SIR), commonly used in health economics as a risk-adjusted measure of performance. Existing analytical methods tend to capture the statistical dispersion of the SIR in simplistic ways. Bootstrapping or simulations are obvious answers, but they tend to be overly time-consuming. This thesis adopts a number of analytical approximations from medicine and applied statistics, and compares their performance against each other as well as the bootstrapped results. The main findings highlight the importance of recognising the stochastic element in the SIR; methods that explicitly account for this produce more accurate confidence intervals than those that did not do so.

The second part of this study investigates the significance of how hospital quality is defined and incorporated into empirical studies. Two issues relating to hospital quality are examined: the stochastic nature of quality statistics, and the multifaceted nature of quality of care. The first issue is tackled using bootstrap regression, adjusting for the distribution of the quality statistic as measured by the SIR. It is found that with bootstrapping, the statistical significance of the quality parameter increased noticeably. The second issue is addressed by introducing a second quality measure into the production function. Results for the two measures were significantly different, highlighting the importance of carefully
defining hospital quality.

Part three of the thesis examines the implications of disease aggregation, a little discussed issue in the literature. This study investigates the importance of disease aggregation via an empirical investigation of the relationship between hospital quality and efficiency. Results indicate an overall inverse relationship, but the magnitude and policy implications change significantly as disease aggregation becomes more refined.
Declaration

This is to certify that:

(i) the thesis comprises only my original work towards the PhD except where indicated in the Preface,

(ii) due acknowledgment has been made in the text to all other material used,

(iii) the thesis is fewer than 100000 words in length, exclusive of table, maps, bibliographies and appendices.

Chun Lok Kris Li
Acknowledgments

(blank)
# Contents

1 Hospital quality and its applications to cost and efficiency 1
   1.1 Motivation ................................................................. 1
   1.2 Theoretical background ................................................. 2
   1.3 Problem statement ....................................................... 5
   1.4 Thesis outline ............................................................ 6
      1.4.1 Chapter 2 ............................................................ 6
      1.4.2 Chapter 3 ............................................................ 7
      1.4.3 Chapter 4 ............................................................ 8

2 On the dispersion of expected incidents when estimating the Standardised Incidence Ratio 9
   2.1 Introduction .............................................................. 9
   2.2 Literature Review ....................................................... 10
   2.3 Method of Analysis ..................................................... 14
      2.3.1 Computing the SIR statistic ...................................... 16
      2.3.2 Estimating CIs: analytic methods ................................. 18
      2.3.3 Evaluating analytic CIs: coverage and location .................. 19
      2.3.4 Evaluating CIs: simulating a benchmark ......................... 21
   2.4 Data Description ....................................................... 25
   2.5 Results ................................................................... 28
      2.5.1 Coverage: comparing analytical CIs with the benchmark .... 28
      2.5.2 Coverage: hospital size and performance of methods .......... 33
      2.5.3 Location: comparison of analytic methods ....................... 35
      2.5.4 Location: analytic methods versus benchmark .................. 36
   2.6 Discussion .................................................................. 37

3 Estimating the cost of improving hospital performance using quality
measures with statistical error

3.1 Introduction .................................................. 42
3.2 Literature Review ............................................ 43
3.3 Method of Analysis .......................................... 47
  3.3.1 Quality measure and its statistical properties .... 47
  3.3.2 The hospital production function ................. 51
  3.3.3 Functional form specification ..................... 53
  3.3.4 Estimating the cost function ...................... 57
  3.3.5 Interpreting estimated results .................... 59
3.4 Data Description ............................................ 61
3.5 Results ..................................................... 63
  3.5.1 Cost function estimates ......................... 63
  3.5.2 Static comparison exercise ....................... 65
  3.5.3 Robustness checks ................................... 69
3.6 Discussion ................................................... 72
  3.6.1 Main findings ....................................... 73
  3.6.2 Limitations .......................................... 74

4 Hospital Quality, Disease Aggregation and Risk Adjustment 79
  4.1 Introduction ................................................ 79
  4.2 Literature Review ......................................... 80
  4.3 Method of Analysis ........................................ 83
    4.3.1 Risk-adjusted quality and disease aggregation .... 83
    4.3.2 Empirical setting for risk-adjusted quality measure ... 86
  4.4 Data Description ............................................ 92
  4.5 Results ..................................................... 95
    4.5.1 Estimates of hospital quality by level of disease aggregation .... 95
    4.5.2 Effect of disease aggregation on empirical results ... 97
    4.5.3 Robustness check: risk-adjustment by fixed effects .... 98
    4.5.4 Robustness check: using differences as quality statistic 100
  4.6 Discussion ................................................... 101
    4.6.1 Summary of findings ................................ 101
    4.6.2 Limitations and extensions ........................ 102

5 Conclusion ..................................................... 106

A Appendix to Chapter 2 ......................................... 121
A.1 Detailed derivation of analytical CI estimation methods ................. 121
A.2 Estimating the sufficient number of repetitions for convergence in CI coverage statistics .................................................. 127
A.3 Comparing the location of analytical CIs versus benchmark results .... 135
A.4 Kernel density plots of the location of analytical CIs ....................... 138
A.5 Comparison of CI positioning between analytic methods using observed hospital data .................................................. 148

B  Appendix to Chapter 3 ............................................................. 152

C  Appendix to Chapter 4 ............................................................. 157
   C.1 Parameter estimates of second-stage DEA/Tobit regression ............ 157
# List of Tables

2.1 Parameters for the hypothetical example hospital .............................................. 15  
2.2 Analytic CIs: a summary .................................................................................. 20  
2.3 Pairwise comparison of CI location: diagram representation ........................... 22  
2.4 Patient count and mortality rate of selected hospitals in the VAED, 1999-2004 25  
2.5 Average coverage and 95% CI width of hospital resamples, low-variance \(\text{(MDC)}\) case .......................................................... 29  
2.6 Average coverage and 95% CI width of hospital resamples, high-variance \(\text{(MDC)}\) case .......................................................... 30  
2.7 The effect of variance on analytical CIs: a breakdown ....................................... 30  
2.8 Patient count of hospitals whose coverage is closer to 95% in a pairwise \(\text{comparison exercise, low-variance (MDC)}\) method ............................. 33  
2.9 Patient count of hospitals whose coverage is closer to 95% in a pairwise \(\text{comparison exercise, high-variance (DRG)}\) method ............................. 34  
2.10 Approximate profile of five selected hospitals for simulating benchmark CI \(\text{locations}................................................................. 36  
2.11 Evaluation of analytical methods: a summary .................................................... 39  
3.1 Overview of distribution of length-of-stay for hospitals ................................. 54  
3.2 Selected hospitals in the VAED, 2002-2004: an overview ............................... 62  
3.3 Summary of patient mortality rates in VAED by year and Major Diagnostic \(\text{Categories}................................................................. 64  
3.4 Elasticities and marginal effects estimated using SUR, with simulated errors 66  
3.5 Elasticities and marginal effects estimated using bootstrap regression, with \(\text{bootstrapped errors}................................................................. 67  
3.6 Static comparison: number of hospitals requiring intervention .................... 68  
3.7 Static comparison: estimated cost of intervention ............................................ 69
3.8 Robustness check: more output types, SUR estimates  
3.9 Robustness check: underlying hospital types by FMM  
3.10 Robustness check: interaction terms between quality and other variables  

4.1 Three tiers of disease aggregation in the VAED dataset  
4.2 Summary statistics of hospital finances and types of patients  
4.3 Mean and variance of hospital SIR values  
4.4 Cluster analysis comparing the MDC and DRG aggregation levels  
4.5 Cluster analysis comparing the service category and DRG aggregation levels  
4.6 DEA estimates of hospital efficiency, by aggregation levels  
4.7 Comparing the 95% CI of quality parameters in the second-stage Tobit regression  
4.8 Robustness check: using hospital fixed-effects as a quality measure  
4.9 Robustness check: estimates of hospital efficiency using alternative quality measure  
4.10 Robustness check: comparing CIs of the quality parameters estimated using alternative quality measure  

A.1 Number of resamples, D1 MDC  
A.2 Number of resamples, D1 DRG  
A.3 Number of resamples, D2 MDC  
A.4 Number of resamples, D2 DRG  
A.5 Number of resamples, D3 MDC  
A.6 Number of resamples, D3 DRG  
A.7 Number of resamples, N1 MDC  
A.8 Number of resamples, N1 DRG  
A.9 Number of resamples, N2 MDC  
A.10 Number of resamples, N2 DRG  
A.11 Number of resamples, S1 MDC  
A.12 Number of resamples, S1 DRG  
A.13 Number of resamples, S2 MDC  
A.14 Number of resamples, S2 DRG  
A.15 Number of resamples, S3 MDC  
A.16 Number of resamples, S3 DRG  
A.17 Number of resamples, S4 MDC  
A.18 Number of resamples, S4 DRG
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A.19 Location of analytical CIs relative to benchmark simulation, D1</td>
<td>135</td>
</tr>
<tr>
<td>A.20 Location of analytical CIs relative to benchmark simulation, D2</td>
<td>135</td>
</tr>
<tr>
<td>A.21 Location of analytical CIs relative to benchmark simulation, D3</td>
<td>135</td>
</tr>
<tr>
<td>A.22 Location of analytical CIs relative to benchmark simulation, N1</td>
<td>136</td>
</tr>
<tr>
<td>A.23 Location of analytical CIs relative to benchmark simulation, N2</td>
<td>136</td>
</tr>
<tr>
<td>A.24 Location of analytical CIs relative to benchmark simulation, S1</td>
<td>136</td>
</tr>
<tr>
<td>A.25 Location of analytical CIs relative to benchmark simulation, S2</td>
<td>136</td>
</tr>
<tr>
<td>A.26 Location of analytical CIs relative to benchmark simulation, S3</td>
<td>136</td>
</tr>
<tr>
<td>A.27 Location of analytical CIs relative to benchmark simulation, S4</td>
<td>137</td>
</tr>
<tr>
<td>A.28 Kernel density plots of the location of analytical CI bounds: D1</td>
<td>139</td>
</tr>
<tr>
<td>A.29 Kernel density plots of the location of analytical CI bounds: D2</td>
<td>140</td>
</tr>
<tr>
<td>A.30 Kernel density plots of the location of analytical CI bounds: D3</td>
<td>141</td>
</tr>
<tr>
<td>A.31 Kernel density plots of the location of analytical CI bounds: N1</td>
<td>142</td>
</tr>
<tr>
<td>A.32 Kernel density plots of the location of analytical CI bounds: N2</td>
<td>143</td>
</tr>
<tr>
<td>A.33 Kernel density plots of the location of analytical CI bounds: S1</td>
<td>144</td>
</tr>
<tr>
<td>A.34 Kernel density plots of the location of analytical CI bounds: S2</td>
<td>145</td>
</tr>
<tr>
<td>A.35 Kernel density plots of the location of analytical CI bounds: S3</td>
<td>146</td>
</tr>
<tr>
<td>A.36 Kernel density plots of the location of analytical CI bounds: S4</td>
<td>147</td>
</tr>
<tr>
<td>A.37 Pairwise comparison between analytic methods, D1, low-variance (MDC) case</td>
<td>148</td>
</tr>
<tr>
<td>A.38 Pairwise comparison between analytic methods, D2, low-variance (MDC) case</td>
<td>148</td>
</tr>
<tr>
<td>A.39 Pairwise comparison between analytic methods, D3, low-variance (MDC) case</td>
<td>148</td>
</tr>
<tr>
<td>A.40 Pairwise comparison between analytic methods, N1, low-variance (MDC) case</td>
<td>149</td>
</tr>
<tr>
<td>A.41 Pairwise comparison between analytic methods, N2, low-variance (MDC) case</td>
<td>149</td>
</tr>
<tr>
<td>A.42 Pairwise comparison between analytic methods, S1, low-variance (MDC) case</td>
<td>149</td>
</tr>
<tr>
<td>A.43 Pairwise comparison between analytic methods, S2, low-variance (MDC) case</td>
<td>149</td>
</tr>
<tr>
<td>A.44 Pairwise comparison between analytic methods, S3, low-variance (MDC) case</td>
<td>149</td>
</tr>
<tr>
<td>A.45 Pairwise comparison between analytic methods, S4, low-variance (MDC) case</td>
<td>150</td>
</tr>
<tr>
<td>A.46 Pairwise comparison between analytic methods, D1, low-variance (DRG) case</td>
<td>150</td>
</tr>
<tr>
<td>A.47 Pairwise comparison between analytic methods, D2, low-variance (DRG) case</td>
<td>150</td>
</tr>
<tr>
<td>A.48 Pairwise comparison between analytic methods, D3, low-variance (DRG) case</td>
<td>150</td>
</tr>
<tr>
<td>A.49 Pairwise comparison between analytic methods, N1, low-variance (DRG) case</td>
<td>150</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Hospital patient count: kernel density plot . . . . . . . . . . . . . . . . . . . 26
2.2 Kernel density plot comparison of $\sigma_Y^2$ for low (MDC) and high (DRG) variance aggregations, by hospitals . . . . . . . . . . . . . . . . . . . 31
2.3 Kernel density plot comparison of $\sigma_Y^2$ for low (MDC) and high (DRG) variance aggregations, by disease types . . . . . . . . . . . . . . . . . . . 32
4.1 Frontier concept of hospital efficiency in DEA . . . . . . . . . . . . . . . . . 87
Chapter 1

Hospital quality and its applications to cost and efficiency

1.1 Motivation

In the fiscal year of 2011/12, Australia spent around 9.4% of their entire national GDP on healthcare (AIHW 2011), out of which 2.8 percentage points of total GDP were spent on public hospital expenses. Other OECD’s have similar proportions of their GDP on healthcare as well, ranging from 6.1% to the special outlier of 17.6% from the United States.\(^1\) Furthermore, historical statistics suggest that this percentage will keep increasing: from 1960 to 2010, OECD total health expenditures grew by around 2% a year.

Healthcare spending was not always a major part of the economy. Rather, it was a result of two general historical developments. First, the advent of social welfare and the public hospital system provided access to healthcare for many more patients than before. This included patients that did not have the necessary resources to pay for the treatments by themselves, and introduced an extra item on the public purse. Second, advances in medical technology expanded the number of treatment options available, increasing the average number of treatments per patient case. As a side-effect, it also raised the per-unit cost of each treatment, as newer treatments also tend to be more expensive. Together,

\(^1\)http://stats.oecd.org/Index.aspx?DataSetCode=SHA
these two factors caused medical expenses to increase at a higher rate than the economy as a whole.

The response to these developments was to focus government efforts towards improving the efficiency and healthcare quality of their public hospitals. For example, the US has implemented a number of pay-for-performance schemes to improve quality while controlling costs. Some, like the Premier Hospital Quality Incentive Demonstration, enjoyed mild improvements in outcomes (Maynard 2012), while others such as P4P under the Affordable Care Act failed to make a noticeable impact (Kruse et al 2012). Australia is in the process of developing similar schemes.

As a complement to direct policy interventions, another way to motivate hospitals to improve their performance is to create open-access databases that contain key statistical measures of performance. For example, Medicare contains a self-service website called Hospital Compare, while the USNews agency releases annual reports of the top-ranking hospitals. There are also a number of easily accessible sources in Australia; MyHospitals is perhaps the well-known and well-publicised one.

This study enhances existing methods of assessing the relationship between hospital quality, cost and efficiency. This is achieved by better capturing the statistical properties of quality measures and by examining in detail the process of disease aggregation during risk-adjustment. It is important because, in order for hospital-improvement schemes to be effective, the methods used for assessing hospital performance needs to be accurate and statistically valid. With good assessment techniques, the effectiveness of policy interventions can be evaluated precisely, and hospitals with the most need for improvement can be correctly identified. The aim is to maximise the quality and accessibility of healthcare from our public hospital system under limited resources.

1.2 Theoretical background

The production of goods and services are usually modeled using identical units, enabling the construction of useful economic models. In most cases this is an acceptable simpli-

---

2http://www.medicare.gov/hospitalcompare/search.html
3http://health.usnews.com/best-hospitals/rankings
4http://www.myhospitals.gov.au
fication to make; each container of orange juice presumably tastes slightly different, but they are still traded in the open market as the same commodities.\footnote{With market rates readily available online: \url{http://futures.tradingcharts.com/marketquotes/OJ.html}} There are also many industries that produce outputs with significant qualitative differences, traded at various prices, even though sometimes the only difference is one of branding. However, economic theory does not need to give special consideration for these industries because each of them individually constitutes a very small percentage of the total economy.

Healthcare is special because it constitutes a very large proportion of the economy and that its outputs differ significantly in quality. As a result, medical price indices such as the MCPI become both important to policymakers but also difficult to estimate at the same time. Some of these difficulties are caused by differences between quoted and actual prices (Berndt et al 2001), which are expected for insurance-based health systems. However, the more serious issue is the shortage of adequate tools for incorporating qualitative differences into units of output (Newhouse 2001).

Obviously there are many others reasons as to why controlling for qualitative differences in healthcare services is important, all of which are widely acknowledged. But until the more important issues pertaining to capturing service quality are addressed, the usefulness of subsequently-derived health-related statistics will remain limited. This reality is curtly summarised from a quote by NG Mankiw (2010):

One of the classic hypothetical questions economists ask when referring to healthcare costs is, “Would you rather go back to 1950s medical care and 1950s prices?” If that option were offered at your place of work, my guess is that you would not take it. What that means is, in some real sense, healthcare is cheaper today if you adjust prices properly to account for quality improvements. A dollar of healthcare today has more value than a dollar of healthcare in 1950.

Unfortunately, there are also numerous practical issues in adjusting for medical care quality, and two of them are arguably more important than the others. First, quality is an abstract and multi-faceted concept that is difficult to define. Second, smoothing out the statistical complications that arise from incorporating quality into existing economic models is non-trivial. Both of these are current research topics.

For the first issue, economists are admittedly less experienced than the medical profession.
Two definitions that are probably the most familiar to both disciplines are quality-adjusted life years (QALY) or its derivative the healthy-years-equivalent (HYE) (Mehrez, Gafni 1989). Both are conceptually sound but difficult to implement.

Perhaps because of this reason, later studies tend to use more tangible proxy measures for quality and directly measure that instead. An example of the latter approach is a study in the health literature that measures the relationship between treatment intensity in terms of dollars charged per treatment (Picone et al 2003). Sometimes more than one variable is used to construct a composite quality measure. The expression used to combine the values into one quantity is called the hedonic aggregator function, also used in a number of examples within health literature (Dudinski et al 1998). More examples are available later in the study.

For the second issue, containing the rising cost of public healthcare is the primary motivation. Since public hospitals are the most prominent providers of healthcare and often the carer of last resort, they are a major contributor of costs. Empirical economic studies on public hospitals usually treat hospitals as firms employing medical staff, capital and equipment to provide services for patients. The two traditional empirical methods for dealing with firms are production functions and efficiency analysis.

Production functions are traditionally presented in algebraic form, but it would be inaccurate to denote them as literal depictions of the underlying production process, except in the case of extremely simple production processes. Economic researchers understand that they normally do not have the data that an industrial engineer possesses, and that the actual production process is too complicated to capture in a closed, parametric form. Instead, the value of an algebraic form is judged by its ability to approximate the underlying process. Using this principle, the most popular production functions are probably free-form specifications such as the translog, Generalised Box-Cox (Khaled 1978) and generalised McFadden (Kumbhakar 1994). They are usually estimated using regression analysis and their properties are widely available in textbooks.

When the word ‘efficiency’ is mentioned in empirical economics, it usually means allocative and technical efficiency. In simple terms, allocative efficiency is the ratio between the

\[ \text{Allocative Efficiency} = \frac{\text{Value of Output}}{\text{Value of Input}} \]

For an introduction to basic definitions of different types of efficiency measurements, see Eilon et al 1976
quantities of outputs a firm actually produced, versus its potential quantity given the same budget, if they have purchased inputs optimally. Technical efficiency is analogous except it compares the output of its current production process to the optimal production method, given the same inputs. Overall efficiency is then the product of the two, or the ratio of what the firm is currently producing against the maximum that the firm could have reached given the same technology and endowments. Efficiency is usually estimated using non-parametric methods and common examples are discussed in more detail in later chapters.

1.3 Problem statement

Measuring hospital performance with adjustments for quality of care is an established research topic in health economics. While the relationships between hospital quality, cost and efficiency are relevant across most settings, there are also differences between public and private healthcare systems and these differences may require different analytical approaches.

Most existing literature on hospital studies were conducted in the US, a predominantly private-sector system funded by a mixture of personal insurance and limited public assistance. Examples of topics that are of interest to US researchers might be the regulation of competitive behaviour or the use of market power to set prices. This may take the form of an empirical study measuring the relationship between price inflation and market concentration, possibly measured by statistics such as the HHI index (Connor et al 1998). If a statistically significant relationship is found, the government might be suspicious of future proposals for hospital mergers (Dranove et al 1992), even if the grounds for efficiency improvements are substantial (Dranove, Lindrooth 2003).

However, for countries with a mostly public healthcare system, the relevance of the type of research as described above is limited. Here, prices are mostly inflexible and hospitals have to take any patients that arrive at the door, subject to capacity constraints. This has a

\[ HHI = \sum_{i=1}^{n} \pi_i^2 \]

, where \( \pi_i \) is the market share of firm i. HHI=0 at the limits indicate a perfectly competitive market and 1 would mean a pure monopoly.

7The Herfindahl-Hirschman Index is a measure of market concentration. The formula is given by
number of statistical and economic implications, some of which will be discussed at length in various parts of this study. But even without delving into the technical details, one can already see that the change in role required for policymakers, moving from a market regulator to a benevolent social planner. Re-shaping the structure of the firm exogenously to improve performance is a very real possibility, and the objective function shifts from simple Pareto-optimality to a more multidimensional set of goals.

This study relaxes a number of simplifying assumptions used to capture the variability of estimated quality measures, in the context of public healthcare systems. It is acknowledged that many of these simplifications were introduced out of necessity, in the face of practical difficulties that are not easily circumvented, and that there are too many of them to be dealt with in a single study. Nevertheless, it is still worthwhile to narrow the scope of future research by identifying promising lines of inquiry. The next section provides an outline of this study and its findings.

1.4 Thesis outline

Healthcare quality belongs to the broader area of health economics and medical research. This study focuses on three topics pertaining to the estimation and application of healthcare quality, with an emphasis on applications to public hospital evaluation. Each problem is studied in a separate chapter, with its own literature review, detailed methodology and results. The last chapter will summarise general findings across the study.

1.4.1 Chapter 2

Hospital quality is an estimated value with its own statistical properties. This is because quality is an abstract concept rather than something directly observable. Empirical economists are aware of this. However, most existing literature either treats hospital quality as known quantities or only makes small adjustments and its variance. The question is whether such treatments of variance in estimated quality measures are adequate.

This chapter analyses the statistical dispersion of the Standardised Incidence Ratio (SIR), a frequently used statistic in health economics, using common methods of estimating confidence intervals. Estimates are then compared against each other and against simulated
values. Results indicate that the SIR distribution is highly skewed and sensitive to both small hospital sizes and higher variances.

These findings provide an explicit demonstration of the pitfalls of treating measures of hospital quality as deterministic values. Additionally, a number of shared properties across approximation methods that capture the variability of the SIR and other similar statistics are identified. These results are important because evaluations of public hospital performance often form the basis of policy decisions. If the reliability of the point estimates are misrepresented, decisions based on these biased estimates could be misguided.

1.4.2 Chapter 3

The question of whether quality improvements raise or lower costs, and by how much, is a controversial topic in health economics that is still debated. Comparing results across studies is difficult because the data and methods vary. This study examines the cost-quality relationship by focusing on two empirical issues that are not well-discussed in the literature. The first issue is the stochastic properties of the quality statistics. The second issue is the multidimensional nature of the notion of quality itself.

To examine the impact of these issues, a comparison exercise is conducted where the data and cost function are kept constant and only the factor of interest is varied. For the first issue, parameter estimates from the estimation procedure with and without adjustments are compared. The estimated parameters using the more rigorous estimation method was negative and with higher statistical significance, while estimates without the adjustment were close to zero and not statistically significant. As for the second issue, the estimated marginal effects of quality were found to vary noticeably between different measures of quality.

These differences in parameter estimates serve to emphasise two empirical details that future studies need to be aware of. The first is the potential for estimation bias if estimated quality measures are directly added into cost functions without adjustment for its statistical properties. The second is the impact that the variable used to measure quality has on results. Both of these findings are applicable towards the goal of improving hospital quality because the financial impact of doing so is a common concern across countries.
1.4.3 Chapter 4

Some hospitals experience more adverse patient incidents than others, but it might be a direct result of serving sicker patients rather than poor performance. Capturing these differences in patient casemix in a fair way is essential for accurate risk-adjustment and for measuring the true performance of hospitals. This makes risk-adjustment an important part of estimating hospital quality. Part of the risk-adjustment process is disease aggregation, a necessary simplification because there are too many patients to individually account for. But there is little discussion in the literature about how various disease aggregation decisions can affect subsequent statistical analysis.

This chapter examines one detail of disease aggregation: whether patients should be aggregated into larger and hence fewer groups or smaller (and more specific) ones. The empirical setting is an estimation of the relationship between hospital quality and efficiency. The analysis is conducted using hospital quality measures computed from different disease aggregation levels and then compared. Results suggest that variations in disease aggregation can significantly change the relationship between quality and efficiency. This demonstrates the necessity of repeating the step of estimating quality measures using multiple levels of disease aggregation when evaluating hospitals.

The findings of this study are closely related to recent improvements in public hospital datasets. As more countries face the dual goals of improving the quality of public healthcare and of containing rising health expenditures, more efforts have been directed towards finding ways to improve the operation of public hospitals. Part of this process involves assembling detailed datasets together for research purposes. This includes setting model specifications for analysis, such as the disease aggregation mechanism for risk-adjustment. The findings of this chapter suggest that disease aggregation is an important detail that needs to be carefully considered when using hospital quality.
Chapter 2

On the dispersion of expected incidents when estimating the Standardised Incidence Ratio

2.1 Introduction

Minimising hospital incidence rate is an important topic in public policy debate. A common statistic used to measure hospital performance is the Standardised Incidence Ratio (SIR), defined as the ratio of observed incidents to the number of expected incidents. The SIR is an estimated statistic rather than an exact measure, with a statistical dispersion that needs to be captured. This is commonly done by computing confidence intervals (CI).

The defining property of the SIR is the expected incidence count in the denominator, which is a probability statement. As such, the integrity of the SIR depends on accurate estimates of its CI, since they are often used for subsequent analysis such as for hypothesis testing, as computed regressors or as instrumental variables. If the statistical dispersion is not estimated accurately, then conclusions about the statistical significance of particular findings may be inaccurate.

The problem with existing methods for estimating the CIs of ratio statistics such as the SIR is that they may not adequately account for the dispersion in expected incidence counts.
The variability in the denominator is usually assumed constant or approximated using simple methods. Additionally, there is little in the existing literature that explicitly tested these methods under actual empirical applications to justify the use of these simplifications. Bootstrapping and non-parametric methods are usually more accurate, but they are time consuming and difficult to use in practice.

To address this problem, we compute the SIR of a set of hospitals. We then compute the CIs of these SIRs using a number of methods adapted from the health and econometrics literature. The accuracy of these methods are measured by their coverage rates and by their similarity to benchmark CIs generated by parametric simulation. The aim is to generate promising lines of enquiry for future research on simple-to-use CI approximations by identifying analytical methods that are more accurate.

Results indicate that the most effective methods for approximating CIs are the ones that capture the higher moments of the SIR distribution. Not all of these approximations are asymptotically consistent, but they work well in an empirical setting. In contrast, more traditional methods that ignore the variance of risk-adjustment in the denominator produced CIs that are too narrow, while normal approximations of one or both random processes in the fraction resulted in CIs that are too wide.

The plan of this chapter is as follows: section 2.2 provides a literature review; section 2.3 describes the data; section 2.4 explains the methodology; section 2.5 reports the findings; section 2.6 summarises.

### 2.2 Literature Review

Concepts similar to the SIR were used for a long time before it was formalised as such. Existing literature commonly refers to Breslow and Day (1987) as the first biostatisticians to name the statistic for evaluating hospitals. Their most widely known contribution is the Breslow-Day homogeneity test for the equivalence of two odds ratios (OR). Some versions require results to be rescaled to an average of one, with values above one indicating below-average performance and vice versa. Historically, the SIR was also known as the Standardised Mortality Ratio (SMR) because for a long time it was used as such.
The SIR in the wider context belongs to the family of hospital quality measures used in health economics and health services research. The relative merits of different quality measures is a current research topic (Ryan et al 2012). Two other common examples of measures within the subgroup of those that form a ratio are the cost-effectiveness ratio (Polsky et al 1997) and the location-quotient ratio (Beyene, Moineddin 2005).

Values of computed quality measures such as the SIR contain statistical variance. This uncertainty is usually captured using CI. The earliest method is Wilson’s Score Interval (Wilson 1927), a continuous approximation based on the binomial distribution deemed suitable for modeling patient mortality. The interval is still in current use (Holman et al 2011) and its statistical properties are still being studied (Tang et al 2010) today. The SIR can also be used to model non-lethal incidents that can occur more than once. For example, Byar’s Method (Rothman, Boice 1979) is derived using an algebraic relationship between the Poisson and Chi-square distributions.

In recent years, most methods similar to the ones mentioned were largely replaced by various non-parametric or bootstrap methods. The benefit of bootstrap methods (Efron 1981) is it avoids the risk of mis-specifying the analytical function. This represents a quantum leap over analytical approaches. The increase in computational power is also an important contributing factor for the rise of bootstrapping. However, computational resources are still not unlimited and bootstrapping cannot be universally applied to all empirical problems just yet. Faster bootstrapping is an active research topic in econometrics (Kneip et al 2011).

A common feature of most existing methods is the assumption that the denominator of the SIR is a known constant or a single random variable. In reality, the incidence rate of each patient is a separate random variable that needs to be accounted for. This is especially important for hospitals that deal with a diversity of patient types. However, accounting for patient heterogeneity by risk-adjusting them individually is also very difficult because it involves computing convolutions (i.e. the distribution of the sum of two or more random variables), most of which do not have analytical solutions.

Indeed, even the simplest examples create surprisingly complicated results. For example, take two uniform distributions $F_1 \sim \text{Uniform}[0, K_{F1}]$ and $F_2 \sim \text{Uniform}[0, K_{F2}]$, with
$K_{F_2} > K_{F_1}$. Let $F_3 = F_1 + F_2$. The distribution of $F_3$ is as follows:

$$f(F_3) = \begin{cases} 
\frac{F_3}{K_{F_1}K_{F_2}} & \text{if } 0 \geq F_3 \geq K_{F_1} \\
\frac{1}{K_{F_2}} & \text{if } K_{F_1} \geq F_3 \geq K_{F_2} \\
-\frac{F_3+K_{F_1}+K_{F_2}}{K_{F_1}K_{F_2}} & \text{if } K_{F_2} \geq F_3 \geq K_{F_1} + K_{F_2} \\
0 & \text{otherwise}
\end{cases}$$

Studying hospital incidents face a similar problem, but to a much greater extent. That is, the distribution of its incidence count is made up of the individual distribution of each of a hospital’s patients. Solutions to a small number of simple distributions were rapidly discovered, although this limited set of available options may not be sufficiently versatile for the specific needs of most studies.\(^2\) As for more complicated distributions, some believe that their analytical solutions would ultimately be left undiscovered. The reason is that, even if the variance of each parameter is assumed normal and iid, a nonlinear function of multiple normal parameters is even more involved than convolutions of known distributions (Davidian, Carroll 1987).

An illustrative example is the convolution of the lognormal distribution, commonly used in finance for pricing assets.\(^3\) Since there is no closed-form solution to the integral (Dufresne 2004, pp. 747), commercial applications typically resort to numerical approximations. At the same time, having an analytical expression does not necessarily mean the problem is solved either. For example, although the solution exists for the convolution of iid logistic distributions (George, Mudholkar 1983), it is not easily computed because it involves factorials of large numbers.

Without a simple exact solution, the alternative is to use approximation methods. The question is which existing methods should be applied in various empirical applications. This study will approach this problem by starting with several well-known methods for approximating ratio statistics from the medical literature. Some of these can be directly applied to the SIR, while others require some modification. The CIs of the SIR point

\(^1\)The solution to multiple heterogeneous uniform distributions bounded below by zero was solved by Olds (1952). Further generalisations appeared after a few more decades.

\(^2\)A textbook summary is available at (Nadarajah 2008).

\(^3\)See Sundaresan (2000) for a general literature review of lognormal distributions.
estimates will be computed using each of the chosen methods. They are then compared against a simulated CI to see which ones line up closest in actual application. The rest of this section provides the background for these methods, with the actual implementation explained in the methods section.

Whenever a distribution seems too complicated to handle, the first instinct is usually to resort to the normal distribution. Fieller’s Interval (Fieller 1932) is probably the earliest to do this, providing a confidence interval for the ratio of the means of two normally distributed variables. The problem is that Fieller’s Interval does not perform well in small sample sizes (Hole 2007). It’s most common use in the economics literature is for deriving confidence intervals for other statistical constructs, such as ratios of regression parameters (Hirschberg, Lye 2010).

Alternatively, a stronger assumption of asymptotic normality of the variables themselves leads to the Gaussian Ratio distribution (Hinkley 1969), which produces a narrower CI than Fieller’s Interval. There are empirical studies that use the Gaussian Ratio directly (Jackson et al 2008), but Hinkley’s result is more commonly used as a lemma to derive other tools in the health literature.

However, there are also two main challenges in using approximations based on the normal distribution. The first is that the assumption of symmetric CIs may be inaccurate in the case of the SIR. This can happen with direct applications of the normal distribution, or with linearised versions such as the Delta-Method. Admittedly this problem does not commonly occur, as most current applications would introduce some adjustments. The second is that methods based on the normal distribution span across the entire real line, while the SIR only spans the non-negative part. This may lead to absurd situations where the estimated lower bound of the SIR is negative, or it may not even exist. This latter is particularly relevant for the SIR because in health economics applications, both halves of the fraction can be very close to zero.

Another approach that has been taken to relax the fixed denominator expectation assumption is to focus on the empirical properties of ratio distributions. An early example involves an approximation for the ratio of two Poisson variables (Ederer, Mantel 1974).

\[\text{In theory, random variables always have a theoretically correct CDF that spans } [0,1] \text{ for any point estimate, so a CI must exist. However, a function that approximates a CDF, such as the Fieller's Interval, may not possess this property.}\]
A refinement of this method is done in a later paper (Silcocks 1994) by transforming the standard normal into a Chi-squared distribution. Recently this was applied to a case where expected incidence is estimated from a small population (Morton et al 2010). Like Byar’s method, Poisson approximations in the algebra may also explain why CIs are used rather than the more obvious standard deviation: a positive mass at 0 in the denominator results in a second moment with a non-converging integral, so variance cannot be computed. A variance for the true distribution may actually exist but not for the approximating function.

Asymptotic properties are sometimes overlooked if the resulting approximation performs better in practice. For the SIR, skewness and kurtosis may be the main barriers preventing a direct normal approximation based on various central limit theorems. If so, adjustments to the normal symmetric CIs can be made by taking exponents of the CI (Hall 1992). Suggestions include cube-roots and kurtosis adjustment based on Edgeworth expansion (Ng et al 2008), as well as a natural log (Sherman et al 2011) method that takes advantage of certain algebraic relationships. This author is not aware of any major studies in the health literature aimed at evaluating these methods.

2.3 Method of Analysis

Before describing the methodology, a hypothetical example is presented to motivate the research question.

There are H hospitals in the dataset, labeled h = 1,...,H. Hospital h has $P_h$ number of patients. For notational simplicity, the subscript h is omitted in this chapter (it is used in subsequent chapters). We select one hypothetical hospital from the dataset at random and it contains 10000 patients. Each patient in the selected hospital is either discharged alive (0) or dead (1), with 0.01 probability of death. These parameters are summarised in Table 2.1.

For computing a 95 percent CI for this example, a binomial distribution obviously suffices. However, for argument’s sake we conduct a naive bootstrap with replacement. To do this, patients are drawn at random with replacement from the sample until a resample of the
Table 2.1: Parameters for the hypothetical example hospital

<table>
<thead>
<tr>
<th>Patients</th>
<th>Death</th>
<th>E(Death)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0.01</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>10000</td>
<td>0</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Each data point, or 'hospital', is constructed by counting the number of patients visiting a physical location within one year. Every patient that enters a hospital has a chance of leaving alive or dead.

To form the SIR, the sum total of the “Death” values of each patient goes to the numerator and the sum total of the “E(Death)” values goes to the denominator of the SIR. A total of 2100 resamples are generated, with each resample indexed using $n = 1,...,N$.

Using the 2100 resamples, the following statistics are computed:

$$SIR_n = \frac{\sum_{p=1}^{P} Death_{pn}}{\sum_{p=1}^{P} E(Death_{pn})}$$

$$SIR_{sorted} = [SIR_1, SIR_2, ..., SIR_N]$$

Lower bound = $SIR_{0.025\times N}$

Upper bound = $SIR_{0.975\times N}$

Using the process as described above, the CI of the SIR was [0.720, 1.060] for our hypothetical hospital. We can also check this bootstrap because for this example, the Data-Generating Process (DGP) is known. Specifically, the number of deaths follows a binomial distribution with many observations. It can be estimated using a normal approximation with mean of 100 and variance of 99. This gives a CI of [0.805, 1.195], which is similar to the bootstrapped one.

Now assume it is later discovered that the true specification of the expected mortality

---

5 Generated matrix available upon request.

6 2000 is a reasonable round number for this study, with 2100 used to avoid approximate continuity correction. The 53rd value of the sorted SIR estimates span the CDF ranges of 0.02476 to 0.02524 to form the lower bound, exactly averaging to a CDF value of 0.025. Similar logic applies for the 2047th value for the upper bound.
rate follows a normal distribution with mean of 0.01 and variance of 0.005. This changes the underlying DGP into a Gaussian Ratio distribution. The distribution of the SIR then becomes:

\[
\text{Hospital deaths} \sim \text{Normal ( Mean = 100, Variance = 10000(0.01*0.99+0.005) )}
\]

\[
\text{Expected hospital deaths} \sim \text{Normal ( Mean = 100, Variance = 99 )}
\]

Using the 2.5% and 97.5% points of the gaussian ratio CDF, the new CI interval becomes [0.7214, 1.3584]. This is wider than the normal approximation and also the bootstrapped value, caused by the variance in expected incidence probabilities that was previously ignored. For common procedures, the variance of incidence probabilities is typically low and the resulting inaccuracy of bootstrap CIs that do not account for this is not large. However, sometimes a hospital may also treat rare diseases, in which case the variance can be higher. As an example, when the variance for the same hypothetical data is increased to 0.02, the underlying CI interval becomes [0.6378, 1.4421]. In the latter case, the inaccuracy of the simple bootstrap CI becomes greater.

As contrived as the example above seems, there are two important messages. The first one is that the uncertainty with regards to expected incidence needs to be directly incorporated as part of any CI estimation method for the SIR. Better analytical or non-parametric methods may provide similarly poor approximations for the underlying DGP if the root cause is not addressed. The second one is that whether the uncertainty of the SIR is a significant problem or not depends very much on the parameters themselves. For this example, the problem is minor with a variance of 0.005 and more serious at 0.02. However, neither of these two sets of results might have any relevance to reality. Therefore, it is important for this study to evaluate the SIR and its CI approximation using an actual dataset of hospitals and patients rather than purely relying on simulated data.

### 2.3.1 Computing the SIR statistic

The SIR is modeled as two separate processes, the number of observed incidents \( X \) and the number of expected incidents \( Y \):

\footnote{Covariance between the numerator and denominator are assumed to be approximately 0.}
Let $\mu_X$ and $\sigma^2_X$ be the mean and variance of $X$ respectively. The distribution of $X$ depends on the type of incidence measured. This study employs mortality as the simplest example. In accordance to previous literature on measuring relative health risks such as Morris and Gardner (1988), mortality is modeled as a binary variable, the sum of which forms the binomial distribution. Let $\hat{\mu}_X$ be the estimate of $\mu_X$ and $\hat{\sigma}^2_X$ the estimate of $\sigma^2_X$, computed as follows:

$$
\hat{\mu}_X = \sum_{p=1}^{P} Death_p \\
\hat{\sigma}^2_X = \mu_X (1 - \frac{\mu_X}{P})
$$

Similarly, let $\mu_Y$ and $\sigma^2_Y$ be the mean and variance of $Y$. The variable $Y$ in the denominator is formed by the sum of the number of expected incidents of each patient. Since mortality is a binary variable, this study estimates its probability using Logistic regression, where patient demographic information as is used control variables. Obviously, binary variables can also be estimated using more complicated regression methods, but the gain in accuracy using those methods were limited in similar existing applications (Glynn, Bernard 1994).

Let $\delta$ be the vector of demographic variables, where each element in the vector is indexed by $u = 1,... U$. Let $\lambda$ be the corresponding vector of parameter estimates. $\hat{\mu}_Y$, the estimate of the mean of $Y$, is computed as follows:

$$
\hat{\mu}_Y = \sum_{p=1}^{P} \frac{exp(\lambda'\delta)}{1 + exp(\lambda'\delta)}
$$

The estimate of $Y$ also has a variance, written as $\hat{\sigma}^2_Y$. The source of the variability in $Y$ comes from the elements in $\lambda$, each of which are normally distributed. This is a problem
because there is no known closed-form analytical solution to capture the variance of the probability of death of each patient as specified. Therefore, we approximate it using the delta method.

Let \( \text{Var}(\lambda) \) be the variance-covariance matrix of the parameter estimates from the Logistic regression. The approximation for \( \sigma_Y^2 \) is given as follows:

\[
\hat{\sigma}_Y^2 = (\hat{\mu}_Y)^2 \delta' \text{Var}(\lambda) \delta
\]

Finally, there is also covariance between mortality outcomes and the probability of mortality events, denoted as \( \sigma_{XY} \). This is estimated using the equation for sample correlation coefficient. Denote the mean of X among all hospitals as \( \bar{X} \) and the mean of Y as \( \bar{Y} \). The equation is given as follows:

\[
\hat{\rho}_{XY} = \frac{\sum_{p=1}^{P} (X_p - \bar{X})(Y_p - \bar{Y})}{\left( \sum_{p=1}^{P} (X_p - \bar{X})^2 \right)^{1/2} \left( \sum_{p=1}^{P} (Y_p - \bar{Y})^2 \right)^{1/2}}
\]

### 2.3.2 Estimating CIs: analytic methods

The following is a list of methods that are evaluated. They are divided into three categories: deterministic (D), normal-based stochastic (N) and other stochastic (S):

- (D1) Wilson’s score interval for binomial distributions (Wilson 1927).
- (D2) Byar’s Method for exact CI assuming Poisson distribution (Sahai, Khurshid 1993).\(^8\)
- (D3) Bootstrap with replacement assuming fixed expected incidence.
- (N1) Fieller’s approximate CI for ratios of variables (Fieller 1932).
- (N2) Gaussian Ratio assuming two normal distributions (Hinkley 1969).

\(^{8}\)Methods are often rehashed in subsequent papers using slightly different algebra. The following citations refer to the one used by this study, which might not be the first paper.
• (S1) Chi-squared ratio approximation (Silcock 1994).
• (S2) Log-adjusted confidence intervals against skewness (Sherman et al 2011).
• (S3) Cube-root adjusted confidence intervals (Hall 1992).
• (S4) Edgeworth adjustment for skewness and kurtosis (Ng et al 2008).

For convenience, these analytical methods are presented in terms of the five variables $\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2$ and $\sigma_{XY}$ as defined above, summarised in Table 2.2. Detailed algebraic derivations are in Appendix A.

### 2.3.3 Evaluating analytic CIs: coverage and location

The quality of an estimated CI can be assessed using two statistics known as coverage and location. The coverage $\psi$ of an analytical CI method is defined as the probability that the CI contains the true value. If a CI approximation method is well-specified, the coverage value of the CI should be very similar to the significance level over a large sample. This probability can be estimated empirically by drawing a large number of samples and computing their CIs.

Let $\mathbb{1}$ be the indicator function. Coverage is estimated as follows:

$$\psi = \sum_{n=1}^{N} \mathbb{1}((lb_n < SIR < ub_n))/N$$

For most functions, the sum of covered and non-covered cases is exhaustive. However, this is not the case for the SIR because approximations to its CI do not always exist for analytical methods. Instead, there are multiple possible outcomes, listed as follows:

- CI bound(s) equal zero.
- Solution failure for CI.
- Negative lower bound for CI.
- CI covers true SIR [coverage].
- CI does not cover true SIR.
Table 2.2: Analytic CIs: a summary

<table>
<thead>
<tr>
<th>Method</th>
<th>Lower bound</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Two solutions to $\frac{\mu_X + z_{1-\frac{\alpha}{2}}/2P \pm z_{1-\frac{\alpha}{2}} \sqrt{\mu_X (1-\mu_X)/P + z_{1-\frac{\alpha}{2}}^2/4P^2}}{1 + z_{1-\frac{\alpha}{2}}^2/2}$</td>
<td>$\frac{\mu_Y}{P}$</td>
</tr>
<tr>
<td>D2</td>
<td>$\frac{\chi_2^2}{2\mu_Y}$</td>
<td>$\frac{\chi_2^2}{2\mu_Y}$</td>
</tr>
<tr>
<td>D3</td>
<td>53rd value out of 2100 bootstraps</td>
<td>2047th value out of 2100 bootstraps.</td>
</tr>
<tr>
<td>N1</td>
<td>Two solutions to $\frac{\mu_X \mu_Y - q^2 \sigma_{XY}}{\mu^2_X - q^2 \sigma_X^2}$</td>
<td>$\frac{\mu_X \mu_Y - q^2 \sigma_{XY}}{\mu^2_Y - q^2 \sigma_Y^2}$</td>
</tr>
<tr>
<td>N2</td>
<td>SIR value s.t. $G(SIR) = \frac{\alpha}{2}$</td>
<td>SIR value s.t. $G(SIR) = 1 - \frac{\alpha}{2}$</td>
</tr>
<tr>
<td>S1</td>
<td>Two solutions to $SIR^2 \mu^2_Y - \sigma^2_Y \chi_{0,1}^{-1} + SIR(2\sigma_{XY} \chi_{0,1} - 2\mu_X \mu_Y) + \mu^2_X - \sigma^2_X \chi_{0,1}^{-1} = 0$</td>
<td>Two solutions to $SIR \cdot \exp \left( \pm \frac{t_{\alpha/2}}{2} \sqrt{V(SIR)} \right)$</td>
</tr>
<tr>
<td>S2</td>
<td>Two solutions to $SIR \cdot \exp \left( \pm \frac{t_{\alpha/2}}{2} \sqrt{V(SIR)} \right)$</td>
<td>Two solutions to $SIR \cdot \exp \left( \pm \frac{t_{\alpha/2}}{2} \sqrt{V(SIR)} \right)$</td>
</tr>
<tr>
<td>S3</td>
<td>Two solutions to $SIR - \left( z_{1-\frac{\alpha}{2}} + \zeta(1 - \frac{\alpha}{2}) \right) \sqrt{\text{Var}(SIR)}$</td>
<td>$SIR + \left( z_{1-\frac{\alpha}{2}} + \zeta(\frac{\alpha}{2}) \right) \sqrt{\text{Var}(SIR)}$</td>
</tr>
<tr>
<td>S4</td>
<td>$SIR - \left( z_{1-\frac{\alpha}{2}} + \zeta(1 - \frac{\alpha}{2}) \right) \sqrt{\text{Var}(SIR)}$</td>
<td>$SIR + \left( z_{1-\frac{\alpha}{2}} + \zeta(\frac{\alpha}{2}) \right) \sqrt{\text{Var}(SIR)}$</td>
</tr>
</tbody>
</table>

Reference table for algebraic expressions of computing analytical CIs. See Appendix A for detailed derivation and usage.

*With conditions; see details in text*
Each CI is classified according to one of these categories, in descending order. The categories together are exhaustive but not mutually exclusive; the order is arranged in this way because a CI has to exist and be positive before it is compared to see if it covers the point estimate. CI bounds can equal zero when the resample has very few patients that died, making one or both CI bounds equal to zero, which is invalid. Some analytical methods may fail to produce a solution with certain values, such as the non-existence of real solutions to a quadratic equation. Negative lower bounds can happen with some analytical methods, especially normal approximations. CIs that pass these conditions are then checked to see whether the true SIR is contained in the CI.

Location acts as a complementary criterion to coverage when CI approximation methods have similar coverage. This means comparing the lower/upper bound values of analytical CIs. Two comparisons are made: 1) analytical methods against each other and; 2) each analytical method against the benchmark CI as implied by the underlying SIR distribution. The possibilities are as follows:

1. Both bounds within benchmarks.
2. Lower bound below lower benchmark, upper bound within benchmarks.
3. Upper bound above upper benchmark, lower bound within benchmarks.
4. Both bounds below lower benchmark.
5. Both bounds above upper benchmark.
6. Both bounds outside benchmarks.
7. Solution fail (coverage case 1-3).

Diagram representation of these possibilities are contained in Figure 2.3.

The estimation of the benchmark CI will be explained below.

2.3.4 Evaluating CIs: simulating a benchmark

Estimating the coverage and location of CIs computed analytically requires many observed samples from each hospital. Obtaining enough patient samples from each hospital is not always possible, so the alternative is to simulate them. This process starts in the same way as the one described for method D3. That is, for each sample from 1 to N, draw M
Table 2.3: Pairwise comparison of CI location: diagram representation

<table>
<thead>
<tr>
<th>Category 1</th>
<th>Category 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category 3</th>
<th>Category 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category 5</th>
<th>Category 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Possible outcomes when comparing CIs produced by one method against another. The subject CI is marked by the horizontal line. The yardstick, which could be the benchmark or the analytical method being compared against, is marked by the bold vertical lines.
resamples with replacement, indexed using \( m = 1_n, \ldots, M_n \). Compute the SIR for each of the \( M \) resamples drawn from sample \( n \), and use them to estimate its CI. The coverage statistic is defined as the percentage of the \( N \) estimated CIs that include the true SIR.

The problem with method D3 is it does not recognise that the probability of death for each patient is a random variable. That is, during the bootstrap stage where \( M \) bootstrap resamples are drawn from each of the \( N \) samples, the expected probability of death for each patient is taken as the real probability. In fact, the probability of death for each patient is computed from a vector of parameters \( \lambda \), with variance-covariance matrix \( \epsilon \). Hence, elements of \( \epsilon \) need to be simulated after each sample is drawn, before the process of bootstrapping resamples begins.

We simulate the error structure of the parameters using the diagonal elements of \( \epsilon \), labeled from \( \epsilon_1 \) to \( \epsilon_U \). This assumes that the correlation between the parameters are zero, a necessary simplification for reducing the simulation time to a manageable length.\(^9\) Let \( \epsilon_n \) be the vector the simulated error terms for each parameter associated with each patient in sample \( n \). To reduce simulation bias, a single error term \( \hat{\epsilon}_n \) is used; this is possible because the sum of each normal error term on the diagonal is also normal. Therefore, the simulated error term for a patient’s probability of death is as follows:

\[
\hat{\epsilon}_n \sim \text{Normal}(0, \sum_{u=1}^{U} \text{Var}(\lambda_u))
\]

Since \( \hat{\epsilon}_n \) varies for each sample from 1 to \( N \), the probability of death for the same patient will be different across samples. The result is a set of SIR estimates with a level of variation that is greater and more in-line with the underlying SIR distribution.

The next step is to adjust for the simulation bias of the estimated expected incidence count of each patient. The problem comes from the error structure of the sum of Logistic functions, summarised as follows:

\(^9\)As a check, the correlation matrix of each disease group was computed. On average, the correlation scores between the parameter estimates for age, gender and the charlson index were less than 0.1.
\[
E\left(\sum_{p=1}^{P} \frac{\exp(\lambda^\prime \delta)}{1 + \exp(\lambda^\prime \delta)}\right) \neq \sum_{p=1}^{P} E\left(\frac{\exp(\lambda^\prime \delta)}{1 + \exp(\lambda^\prime \delta)}\right)
\]

For this function, the bias as defined by the difference between these two expressions cannot be solved algebraically. A viable alternative is to estimate the bias empirically for each hospital.

Generate 1000 bias-adjustment resamples for each sample, indexed by \(s = 1, \ldots, S\). For each bias-adjustment resample, estimate the expected number of mortality incidents \(\mu_Y\) in the resample. The bias-adjustment factor, \(b(\mu_Y)\) is defined as follows:

\[
b(\mu_Y) = \mu_Y - \frac{1}{1000} \sum_{s=1}^{1000} \mu_{Y,s}
\]

Now there is enough information to compute the five variables \(\mu_X, \sigma_X^2, \mu_Y, \sigma_Y^2\) and \(\sigma_{XY}\) for each resample from 1 to \(N\). The entire process produces the following for each resample:

1. The SIR value, defined by \(\frac{\mu_X}{\mu_Y}\).
2. CI estimates using analytical methods, defined using the five parameters, for estimating coverage.
3. Benchmark CIs of each hospital, defined by the corresponding percentile values of the SIR estimates in each resample, for estimating location. Similar to D3, but with variable expected incidents.

There are two weaknesses to this bias-adjustment process. The first is that some bias-adjusted values may be negative as a result of over-adjustment for a small numbers of resamples. This produces negative CI lower bounds for some analytical methods that otherwise should never become negative. It does not happen asymptotically, but empirically it might take an impractical number of repetitions to fully remove the negative bounds. The second is it does not address possible bias for higher moments.
2.4 Data Description

The list of patients admitted to each hospital comes from the Victorian Admitted Episodes Database (VAED), a record of in-patient episodes treated by hospitals in the state of Victoria. Six years of VAED hospital data spanning 1999/00 to 2004/05 are chosen for this study, years during which variable definitions have remained relatively similar. This selection from the VAED results in patients from about 250 hospital campuses a year, bringing the total to about 1500 data points. Of this selection, patients from 1123 campuses are chosen on the basis that the hospital had incurred at least one mortality case. Names of each patient and each hospital are de-identified with random codes.

The means and standard deviations of the number of patients in each hospital for each of the six years are listed in Table 2.4, with a kernel density plot of these values provided in Figure 2.1.

Table 2.4: Patient count and mortality rate of selected hospitals in the VAED, 1999-2004

<table>
<thead>
<tr>
<th>Year</th>
<th>Hospitals</th>
<th>Average hospital size (Std Dev)</th>
<th>Average Death rate (Std Dev)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999/00</td>
<td>195</td>
<td>5793 (8233)</td>
<td>2.99% (6.40%)</td>
</tr>
<tr>
<td>2000/01</td>
<td>195</td>
<td>5992 (8445)</td>
<td>2.90% (6.36%)</td>
</tr>
<tr>
<td>2001/02</td>
<td>187</td>
<td>6290 (8892)</td>
<td>3.39% (7.87%)</td>
</tr>
<tr>
<td>2002/03</td>
<td>185</td>
<td>6623 (9264)</td>
<td>3.45% (8.13%)</td>
</tr>
<tr>
<td>2003/04</td>
<td>178</td>
<td>7067 (9664)</td>
<td>3.48% (8.56%)</td>
</tr>
<tr>
<td>2004/05</td>
<td>183</td>
<td>7009 (9896)</td>
<td>3.26% (8.18%)</td>
</tr>
</tbody>
</table>

This table lists the average number of patients in a hospital, for each year. The number in brackets is the standard deviation. Mortality rates are listed similarly.

Variables for each VAED record include the standard ones such as demographics, medical diagnosis and procedures undertaken. Most of them are typical for patient datasets everywhere and do not demand detailed discussion. The ones used for patient risk-adjustment include age, gender and a computed Charlson comorbidity index (Charlson et al 1987). The Charlson index is a score-based system used in health services for risk-adjustment, based on observed patient characteristics. This study uses the ICD-10-AM version, the Australian version of the tenth modification since the original system (Sundararajan et al 2007). It is also the classification system used in the VAED.

Additionally, each episode in the VAED are organised according to a hierarchy of casemix

\[\text{For more details of the dataset, see http://www.health.vic.gov.au/hdss/vaed/index.htm}\]
classification groups, so there is more than one system of labeling disease types. This study will use two of them:

- Major Diagnostic Categories (MDC), 23 in total.
- Diagnostic Related Groups (DRG), 661 in total.

This hierarchy of casemix classifications from the VAED is especially useful for creating the low and high variance scenarios for testing the analytical CIs. To create the two scenarios, the list of patients in the VAED is divided according to MDC or DRG definitions. For each aggregation, the incidence probabilities of those patients are estimated using a separate Logistic regression. Since there are fewer MDCs than DRGs, each MDC will contain more patients than each DRG. The result is that incidence probabilities estimated at the MDC level contains less variance than those estimated at the DRG method. This is preferable to artificially increasing the variance of existing data, since it is not certain how realistic
the chosen variance inflation will be.  

This empirical setup is motivated by the increased availability of information in datasets. The issue is that when a comprehensive dataset is available, a natural inclination might be to automatically use the most detailed patient variables available for regression. The more information there is, the more fully the regression will capture patient heterogeneity. However, the finer the patient classification system is, the fewer observations there are in each aggregation. This leads to higher variance in the risk-adjustment stage. Hence, what seems to be a strict improvement is in fact affected by a classic bias-variance trade-off, where fewer casemix groups may introduce bias as heterogeneity between patients in the same group is missed, while more but smaller groups may increase variance. This dynamic will be captured in the CI estimates for hospital SIR using the two different casemix levels.  

Admittedly, the fixed effects approach is the more common way to control for patient disease type. That is, each casemix group is identified using a separate variable and the entire patient sample is estimated all at once. The downside to this is that the marginal effects will be artificially restricted to be the same across all casemix types. This problem is eliminated by running separate regressions for each casemix group, allowing the same variables to take on different marginal effects for each casemix type. The result is that each disease type will have its own mean and variance for \( \lambda \), better capturing patient heterogeneity. Obviously, including fewer or more fixed effects parameters for each disease aggregation can also create a low and a high variance scenario.  

After risk-adjustment, the next step in preparing the data is to generate resamples of the patient makeup in each hospital. Resamples are needed for estimating coverage and location because there are only six years per hospital campus at most. This study generates two sets of simulated data, one for coverage and the other for location.  

For evaluating coverage, 100 resamples are generated for each hospital in the VAED dataset to identify common characteristics of hospitals that perform well under different CI estimation methods. A sample of 1123 hospitals and 100 resamples each gives a grand total of 112300 random resamples from a variety of hospitals with which to evaluate CI approximation methods. As for evaluating location, five hospitals of different sizes, at the [5th, 25th, 50th, 75th, 95th] percentiles, are chosen. These CIs are first estimated at 5000
resamples for more accuracy, and then compared to the percentile CI estimate as defined by the realised SIR values of the resamples.

For any individual hospital, 100 resamples to estimate coverage is inadequate, but in aggregate this is viable because the idea is to average the coverage estimates over many hospitals to make general observations. It is an arrangement that provides a balance between robustness with practical feasibility, as generating thousands of resamples for each of the hospitals in the entire VAED dataset would be impractical. There is no hard rule for how many repetitions are needed until the estimated coverage converges. As a precaution, the median hospital is singled out and a larger number of repetitions were made, with results presented in Appendix A.

2.5 Results

The coverage statistics of the analytical CIs are presented. Driving factors of coverage are examined, including hospital size. The locations of the CI bounds are first compared against each other using observed hospital data, and then compared against benchmark results using simulated data. The locations of simulated CIs are also aggregated into kernel density form and compared against benchmark results. A summary of findings is provided in the discussion section.

2.5.1 Coverage: comparing analytical CIs with the benchmark

The coverage results of the CI estimates in the low variance scenario are listed in Table 2.5, where fixed and variable denominator methods have similar levels of performance. This is most clearly observed by comparing the coverage statistics of the D-methods with the N-methods and S-methods. For the normal-based methods, N1 has very similar coverage statistics to N2 as expected. The slight differences between the two are mostly explained by the difference between convergence in distribution and convergence in probability.

Another observation is that the log-adjustment method S2 did not work for the SIR statistic at all. The CIs are too narrow to be useful, resulting in non-cover rates of over 90%. The likely reason is because dividing by \( N \) is only valid if the error structure of the expected incidence probabilities of the patients are approximately iid normal. The design
Table 2.5: Average coverage and 95% CI width of hospital resamples, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Case</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure %</td>
<td>2.40%</td>
<td>2.40%</td>
<td>11.53%</td>
<td>21.97%</td>
<td>22.57%</td>
<td>21.88%</td>
<td>2.51%</td>
<td>5.21%</td>
<td>30.27%</td>
</tr>
<tr>
<td>— zero CI bound(s)</td>
<td>2.37%</td>
<td>2.37%</td>
<td>11.45%</td>
<td>2.37%</td>
<td>2.37%</td>
<td>2.37%</td>
<td>2.37%</td>
<td>2.37%</td>
<td>2.37%</td>
</tr>
<tr>
<td>— fail, solution</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>16.36%</td>
<td>16.25%</td>
<td>0.14%</td>
<td>0.71%</td>
</tr>
<tr>
<td>—**fail, neg lbound</td>
<td>0.03%</td>
<td>0.03%</td>
<td>0.07%</td>
<td>3.24%</td>
<td>20.20%</td>
<td>3.26%</td>
<td>0.00%</td>
<td>2.14%</td>
<td>27.08%</td>
</tr>
<tr>
<td>cover</td>
<td>81.70%</td>
<td>83.83%</td>
<td>64.71%</td>
<td>73.29%</td>
<td>72.36%</td>
<td>73.37%</td>
<td>7.46%</td>
<td>89.63%</td>
<td>65.25%</td>
</tr>
<tr>
<td>— Avg. simulated CI width</td>
<td>1.1316</td>
<td>1.1901</td>
<td>0.7322</td>
<td>0.7405</td>
<td>0.7282</td>
<td>0.7425</td>
<td>0.2699</td>
<td>0.9746</td>
<td>0.6261</td>
</tr>
<tr>
<td>— Avg. benchmark CI width</td>
<td>1.591%</td>
<td>13.77%</td>
<td>23.76%</td>
<td>4.75%</td>
<td>5.07%</td>
<td>4.75%</td>
<td>90.02%</td>
<td>5.15%</td>
<td>4.48%</td>
</tr>
<tr>
<td>not cover</td>
<td>0.1395</td>
<td>0.0997</td>
<td>0.1684</td>
<td>0.0340</td>
<td>0.0321</td>
<td>0.0340</td>
<td>0.0742</td>
<td>0.0299</td>
<td>0.0264</td>
</tr>
<tr>
<td>— Avg. CI width</td>
<td>0.0232</td>
<td>0.1771</td>
<td>0.2532</td>
<td>0.0552</td>
<td>0.0594</td>
<td>0.0554</td>
<td>1.0247</td>
<td>0.0663</td>
<td>0.0567</td>
</tr>
</tbody>
</table>

The percentage of CIs computed from all resamples that belong to each coverage category, as defined in the methods section.

*The same resamples are used to compute all CI intervals. However, D3 operates by bootstrapping from the resamples, giving it an extra chance to draw a sample that produces a lower and/or upper bound of zero.

**Includes negative \(\mu_Y\) produced during the bias-adjustment step of producing the resamples. This explains the positive percentages for methods that otherwise cannot possibly produce negative lower bounds.

of the resampling mechanism in this study demonstrates the extreme dependence of S2 to that assumption, which in this example is very inaccurate. This explanation is also consistent with the location of the CI bounds that will be presented later on.

At higher variances, the risk of solution failure becomes a more important problem, as listed in Table 2.6. Analytical CIs in this instance are much narrower than their benchmark counterparts. This problem appears across all the methods in this study, more strongly for fixed denominator methods. Normal approximations do not work well either because their CI lower bound become negative too often. Methods algebraically related to Fieller’s Interval, such as the Chi-squared ratio (S1), also suffer a similar problem, with around 69% of simulation failures. S4 is better at avoiding negative lower bounds, but at the cost of reducing coverage to below that of simpler fixed expected incidence methods. Cube-root skewness adjustment of S3 failed in some cases and provided better coverage than fixed denominator methods in others.

Another significant result is the lower coverage of D3 as compared to D1 and D2, running against the intuition that bootstrapping is usually better. This occurred in both the low and high variance scenarios. The likely reason is that D3 completely ignores all the dispersion in expected incidence probabilities, while the algebraic structure of D1 and D2 happens to capture some of this dispersion. The importance of understanding the problem before directly applying a bootstrap method is clearly illustrated here.
Table 2.6: Average coverage and 95% CI width of hospital resamples, high-variance (MDC) case

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Failure %</td>
<td>5.70%</td>
<td>5.71%</td>
<td>16.44%</td>
<td>75.32%</td>
<td>75.30%</td>
<td>6.18%</td>
<td>31.64%</td>
<td>75.51%</td>
<td></td>
</tr>
<tr>
<td>—zero CI bound(s)</td>
<td>5.64%</td>
<td>5.64%</td>
<td>12.08%</td>
<td>5.64%</td>
<td>5.64%</td>
<td>5.64%</td>
<td>5.64%</td>
<td>5.64%</td>
<td>5.64%</td>
</tr>
<tr>
<td>—fail, solution</td>
<td></td>
<td></td>
<td></td>
<td>69.14%</td>
<td></td>
<td>69.11%</td>
<td></td>
<td>5.53%</td>
<td>4.63%</td>
</tr>
<tr>
<td>—fail, neg lbound</td>
<td>0.06%</td>
<td>0.07%</td>
<td>4.36%</td>
<td>0.54%</td>
<td>69.65%</td>
<td>0.54%</td>
<td>0.00%</td>
<td>21.37%</td>
<td>61.49%</td>
</tr>
<tr>
<td>cover</td>
<td>19.09%</td>
<td>20.22%</td>
<td>20.01%</td>
<td>12.10%</td>
<td>1.27%</td>
<td>12.11%</td>
<td>7.37%</td>
<td>38.56%</td>
<td>3.12%</td>
</tr>
<tr>
<td>—Avg. simulated CI width</td>
<td>0.2204</td>
<td>0.2765</td>
<td>0.3977</td>
<td>0.5644</td>
<td>0.0164</td>
<td>0.5671</td>
<td>0.2516</td>
<td>1.0225</td>
<td>0.0524</td>
</tr>
<tr>
<td>—Avg. benchmark CI width</td>
<td>0.9201</td>
<td>0.9942</td>
<td>0.8112</td>
<td>0.1665</td>
<td>0.0201</td>
<td>0.1674</td>
<td>0.4591</td>
<td>0.8985</td>
<td>0.0821</td>
</tr>
<tr>
<td>not cover</td>
<td>75.21%</td>
<td>74.08%</td>
<td>63.56%</td>
<td>12.58%</td>
<td>23.45%</td>
<td>12.59%</td>
<td>86.45%</td>
<td>29.80%</td>
<td>21.37%</td>
</tr>
<tr>
<td>—Avg. CI width</td>
<td>0.4023</td>
<td>0.3957</td>
<td>0.2618</td>
<td>0.0605</td>
<td>0.0913</td>
<td>0.2618</td>
<td>0.3886</td>
<td>0.1384</td>
<td>0.1109</td>
</tr>
<tr>
<td>—Avg. benchmark CI width</td>
<td>2.1181</td>
<td>2.0421</td>
<td>1.4789</td>
<td>0.2225</td>
<td>0.3666</td>
<td>0.2225</td>
<td>0.6634</td>
<td>0.3655</td>
<td></td>
</tr>
</tbody>
</table>

Coverage statistics similar to the ones in Table 2.5, but for the high-variance (DRG) case.

One way of explaining the differences in coverage between the low and high variance scenarios is to examine the factors that drive their differing levels of variance. This lever is the risk-adjustment process; specifically, the low-variance MDC and the high-variance DRG disease aggregation levels. These differences are captured with the parameters $\mu_Y$ and $\sigma_Y^2$. Table 2.7 is a summary of the values of the five parameters for the hospital sample used to define the analytical methods.

Table 2.7: The effect of variance on analytical CIs: a breakdown

<table>
<thead>
<tr>
<th></th>
<th>MDC (low-variance)</th>
<th>DRG (high-variance)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hospital mean</td>
<td>Hospital Std. Dev.</td>
</tr>
<tr>
<td>$\mu_X$</td>
<td>96.3232</td>
<td>168.8581</td>
</tr>
<tr>
<td>$\mu_Y$</td>
<td>90.1977</td>
<td>158.5079</td>
</tr>
<tr>
<td>$\sigma_X^2$</td>
<td>94.7889</td>
<td>156.9914</td>
</tr>
<tr>
<td>$\sigma_Y^2$</td>
<td>11.8848</td>
<td>20.7858</td>
</tr>
<tr>
<td>$\sigma_{XY}$</td>
<td>0.0024</td>
<td>0.0038</td>
</tr>
</tbody>
</table>

Analytical CIs are defined by the five variables $\mu_X$, $\mu_Y$, $\sigma_X^2$, $\sigma_Y^2$, and $\sigma_{XY}$. The derivation for these values is explained in the methods section. This table contain the mean and standard deviation of these values in the hospital sample. Both low and high variance scenarios are included.

$\mu_X$ is a hospital-level variable and is not affected by risk-adjustment. The high standard deviation for $\mu_X$ shows that there is a wide range of hospital sizes. The high standard deviation for $\sigma_X^2$ shows that there is a range of hospital types, from those that only treat a few types of diseases to those equipped for a diversity of problems. The variables $\mu_Y$ and $\sigma_{XY}$ could have also turned out to be different, but in fact they were very similar. These observations confirm that the set of hospitals in this study are diverse in size and types, increasing the likelihood that the results in this study are broadly applicable beyond hospitals examined in this study.
Figure 2.2: Kernel density plot comparison of $\sigma_Y^2$ for low (MDC) and high (DRG) variance aggregations, by hospitals

Each hospital has an associated $\sigma_Y^2$, ranked from lowest on the left to highest on the right, mapped by kernel density plot. The dashed line represents the low-variance (MDC) case and the solid line the high-variance (DRG) case. The horizontal axis is log-scaled.

The remaining variable to examine is $\sigma_Y^2$, which shows a large difference between the MDC (broad) and DRG (narrow) casemix classification systems. This is expected, but the reason for this phenomenon needs to be identified. Therefore, the next step is to compare hospital-level differences or outlier disease types that produced extra-high standard errors for their incidence probabilities.

Figure 2.2 contains estimates of $\sigma_Y^2$ for each hospital in kernel density form. The horizontal axis contains log-scale values of $\sigma_Y^2$ and the vertical axis contains the kernel density. Here, the difference between the MDC and DRG distribution is a shift on the horizontal axis. This indicates that the difference in the distribution of variance estimates between risk-adjustment methods is one of scale rather than of type. Consequently, there is no evidence that there are particular types of hospitals that are disproportionately affected by a change
Figure 2.3: Kernel density plot comparison of $\sigma^2_Y$ for low (MDC) and high (DRG) variance aggregations, by disease types

Each hospital has an associated $\sigma^2_Y$, ranked from lowest on the left to highest on the right, mapped by kernel density plot. The dashed line represents the low-variance (MDC) case and the solid line the high-variance (DRG) case. The horizontal axis is log-scaled.

in risk-adjustment method.

Figure 2.3 contains a similar-looking kernel density graph, but this time by disease type. There is a clear difference in the shapes of the two curves that represent their respective risk-adjustment method, with most of the probability density in the DRG case shifted towards the right. This represents a group of diseases in the upper end of the variance scale with higher $\sigma^2_Y$ estimates. There are no obvious outliers on the graph either. Therefore, the difference in CI coverage between the methods is caused by the DRG risk-adjustment method itself, with more patient categories but fewer patients in each group.

The important message here is the confirmation of the bias-variance trade-off as explained
earlier. A further demonstration of this phenomenon and its implications on empirical research are addressed in more detail in Chapter 4.

### 2.5.2 Coverage: hospital size and performance of methods

The coverage statistics presented in Tables 2.5 and 2.6 are averaged across all hospitals. The next question to ask is whether certain methods are preferred to others for different types of hospitals. However, there are many hospital characteristics available and not all of them can be examined. This study will focus on hospital size because it is clearly defined and one that is certain to affect CI estimates.

Each hospital has a set of analytical CIs, each with its associated coverage rate. The coverage of these CIs are compared pairwise to identify which percentage is closer to 95%. The comparison is two-sided: e.g. a coverage rate of 97% and 93% is considered equally accurate for this exercise. A one-sided comparison is possible but its relevance is more dependent on specific applications. The goal is to identify methods that perform better for hospitals of different sizes.

To summarise results for each pairwise combination, the average number of patients within the set of hospitals found to perform better using one method relative to the other is listed in Tables 2.8 and 2.9, for the low and high variance scenarios respectively.

Table 2.8: Patient count of hospitals whose coverage is closer to 95% in a pairwise comparison exercise, low-variance (MDC) method

<table>
<thead>
<tr>
<th>Method to evaluate</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>-</td>
<td>6381</td>
<td>13015</td>
<td>10806</td>
<td>10625</td>
<td>10787</td>
<td>nil</td>
<td>9290</td>
<td>10965</td>
</tr>
<tr>
<td>D2</td>
<td>1022</td>
<td>-</td>
<td>24224</td>
<td>11680</td>
<td>11528</td>
<td>11680</td>
<td>nil</td>
<td>10636</td>
<td>11528</td>
</tr>
<tr>
<td>D3</td>
<td>6079</td>
<td>6091</td>
<td>-</td>
<td>7657</td>
<td>7762</td>
<td>7649</td>
<td>752</td>
<td>6628</td>
<td>8870</td>
</tr>
<tr>
<td>N1</td>
<td>2480</td>
<td>2519</td>
<td>2860</td>
<td>-</td>
<td>5083</td>
<td>385</td>
<td>2869</td>
<td>5488</td>
<td>9348</td>
</tr>
<tr>
<td>N2</td>
<td>2493</td>
<td>2506</td>
<td>2746</td>
<td>2513</td>
<td>-</td>
<td>2502</td>
<td>2877</td>
<td>5317</td>
<td>9920</td>
</tr>
<tr>
<td>S1</td>
<td>2480</td>
<td>2519</td>
<td>2860</td>
<td>409</td>
<td>5174</td>
<td>-</td>
<td>2869</td>
<td>5480</td>
<td>9311</td>
</tr>
<tr>
<td>S2</td>
<td>6447</td>
<td>6447</td>
<td>6725</td>
<td>6977</td>
<td>7055</td>
<td>6963</td>
<td>-</td>
<td>6593</td>
<td>7469</td>
</tr>
<tr>
<td>S3</td>
<td>2309</td>
<td>2494</td>
<td>2032</td>
<td>8610</td>
<td>7858</td>
<td>8838</td>
<td>1300</td>
<td>-</td>
<td>9762</td>
</tr>
<tr>
<td>S4</td>
<td>2536</td>
<td>2531</td>
<td>2417</td>
<td>3088</td>
<td>3099</td>
<td>3075</td>
<td>2521</td>
<td>3046</td>
<td>-</td>
</tr>
</tbody>
</table>

Read across, then down. For example, 6381 is the average number of patients of the hospitals where the estimated CI using method D2 has coverage values closer to 95% than the estimated CI using method D1.

For the low variance scenario, fixed denominator methods represented by the first three
columns show that they are more accurate for smaller hospitals. D1 and D2 are also better for hospitals averaging around 6000 patients, which is basically the sample average, or the majority of the sample hospitals. Both of these observations are consistent with previous results about the naive bootstrap. Comparing between fixed and variable denominator methods, the cases where the latter out-perform the former are ones with larger hospitals. The interpretation is that although variable expected incidence methods are more accurate, the algebraic approximations involved do not always work with small sample sizes.

Table 2.9: Patient count of hospitals whose coverage is closer to 95% in a pairwise comparison exercise, high-variance (DRG) method

<table>
<thead>
<tr>
<th>Method to evaluate</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>-</td>
<td>2087</td>
<td>5721</td>
<td>13461</td>
<td>16660</td>
<td>13461</td>
<td>127</td>
<td>8091</td>
<td>19405</td>
</tr>
<tr>
<td>D2</td>
<td>(none)</td>
<td>-</td>
<td>5858</td>
<td>13474</td>
<td>17026</td>
<td>13474</td>
<td>220</td>
<td>8219</td>
<td>19851</td>
</tr>
<tr>
<td>D3</td>
<td>2725</td>
<td>2696</td>
<td>-</td>
<td>14475</td>
<td>17389</td>
<td>14475</td>
<td>598</td>
<td>7873</td>
<td>19744</td>
</tr>
<tr>
<td>N1</td>
<td>2586</td>
<td>2600</td>
<td>2825</td>
<td>-</td>
<td>(none)</td>
<td>1517</td>
<td>1591</td>
<td>4286</td>
<td>866</td>
</tr>
<tr>
<td>N2</td>
<td>4732</td>
<td>4727</td>
<td>5047</td>
<td>8436</td>
<td>-</td>
<td>8414</td>
<td>3480</td>
<td>6541</td>
<td>6642</td>
</tr>
<tr>
<td>S1</td>
<td>2586</td>
<td>2600</td>
<td>2825</td>
<td>3796</td>
<td>(none)</td>
<td>-</td>
<td>1591</td>
<td>4286</td>
<td>869</td>
</tr>
<tr>
<td>S2</td>
<td>5059</td>
<td>5017</td>
<td>5859</td>
<td>11295</td>
<td>14690</td>
<td>11275</td>
<td>-</td>
<td>7201</td>
<td>16064</td>
</tr>
<tr>
<td>S3</td>
<td>2092</td>
<td>2028</td>
<td>2171</td>
<td>20739</td>
<td>444</td>
<td>20739</td>
<td>1565</td>
<td>-</td>
<td>1399</td>
</tr>
<tr>
<td>S4</td>
<td>4286</td>
<td>4282</td>
<td>4509</td>
<td>9568</td>
<td>2026</td>
<td>9579</td>
<td>3238</td>
<td>6319</td>
<td>-</td>
</tr>
</tbody>
</table>

Read across, then down. For example, 2087 is the average number of patients of the hospitals where the estimated CI using method D2 has coverage values closer to 95% than the estimated CI using method D1.

Fixed denominator methods do not behave very differently at higher variance but variable denominator methods do. Normal-based methods N1 and N2 are only better than fixed denominator methods when hospital sizes are large enough to avoid the negative lower bound problem. S1 is similar to N1 in derivation and exhibits similar results. The difference between S3 and S4 is also more clearly flashed out. S4 is preferred for larger hospitals where kurtosis matters more, while S3 is better in most other situations.

As a sidenote, S2 did not perform as badly for the smaller hospitals. This lends further support to the hypothesis that it performed badly because of the iid normal assumption, as the distortionary effect of dividing by N as explained earlier is weaker if N is smaller. In higher variances, this problem becomes more clearly visible.
2.5.3 Location: comparison of analytic methods

CIs with similar levels of coverage can have different lower and upper bounds because they are not always centered at the median. Some of them may be shifted more to the left or right, while others may have wider or narrower CI widths. These traits define the characteristic of the analytical methods and may be useful depending on the hypothesis associated with the research question. For example, if a new policy dictates that hospitals with an SIR beyond a certain threshold will face disciplinary action, then a right-skewed CI with a lower bound of a greater value will reduce the probably of mistake in applying punitive measures, whereas for cross-comparison between hospitals, a two-sided CI centered around the median value of the CDF would be more sensible.

To identify these characteristics, the location of CIs computed using the analytic methods are compared against each other. The possible cases are as defined in Figure 2.3 of the methods section, four of which have relevant empirical interpretations:

1. Category 2 ensures the lower limit is covered.
2. Category 3 ensures the upper limit is covered.
3. Category 1 ensures the CI is more narrow and conservative.
4. Category 6 ensures the CI is wide enough to cover all the possibilities.

For this exercise only, observed data is used rather than simulated ones. The reason for doing this is that results computed using simulated data are more easily distorted by a small number of hospitals with more unusual characteristics. For example, there may be one hospital where method A is preferred over method B in 99% of the resamples, while for the other three hospitals, method A is only preferred 40% of the time. In this case, a straight average will produce biased results. But weighted averages are also impractical because results will then be largely driven by how the weights are assigned, and there is also no clear and logical basis to do so. There are 1123 hospitals and 9 analytical CIs for pair-wise comparisons in total. Both low-variance (MDC) and high-variance (DRG) scenarios are tested, with detailed tables of results included in Appendix A.

Results indicate that stochastic methods have lower bounds that stretch further towards zero than deterministic ones. This effect is most clearly seen in S3 and S4, with N2 included for higher variances. Amongst the D-methods, D3 has lower bounds close to zero than D1 and D2. Similarly, for upper bounds that stretch further, methods D1 and S1 are
more appropriate in the low variance case, while N1 is also usable for high-variance cases. In terms of CI widths, D1 produces narrow CI estimates while N1 and N2 produces wider ones. However, N1 and N2 do not work well because they fail to generate solutions too often.

2.5.4 Location: analytic methods versus benchmark

The locations of analytical CIs are also compared against the benchmark. The purpose is to measure how closely the distribution implied by the analytical method matches that of the simulated one. As explained in the methods section, a sample of five hospitals is selected for simulated patient samples that span the range of hospital sizes in the dataset, due to computational constraints. To compensate, a larger number of simulations are computed. Information about the chosen hospitals are limited to approximate ranges of number of episodes a year and only approximate mortality rates are provided in Table 2.10.

Table 2.10: Approximate profile of five selected hospitals for simulating benchmark CI locations

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance</th>
<th>High variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp 1</td>
<td>1-10</td>
<td>1-10</td>
</tr>
<tr>
<td>Hosp 2</td>
<td>10-100</td>
<td>10-100</td>
</tr>
<tr>
<td>Hosp 3</td>
<td>100-200</td>
<td>100-200</td>
</tr>
<tr>
<td>Hosp 4</td>
<td>1000-2000</td>
<td>1000-2000</td>
</tr>
<tr>
<td>Hosp 5</td>
<td>10000-20000</td>
<td>10000-20000</td>
</tr>
</tbody>
</table>

Descriptive statistics of 5 hospitals chosen from the dataset for detailed analysis, with number of patients at the 5th, 25th, 50th, 75th and 95th percentiles. Incidence count is obscured, variances are given in full and mortality rates are rounded to the nearest whole percentage.

Tables of summary statistics and kernel density graphs comparing the locations of simulated CI approximations to benchmark results are both included in Appendix A.

A CI is considered accurate if the majority of the simulated CIs are concentrated in categories 2 and 3, with similar proportions for both sides indicating little bias. For low variances, variable denominator methods often do not give more accurate CI bounds as
compared to fixed denominator ones and are probably not worth the effort. For hospitals 1 and 2 with fewer patients, variable denominator methods are unstable and suffer from too many solution failures. Hospital 4 is included to see the effects of low death rates, with the result that many variable denominator methods become too wide. In contrast, fixed denominator methods perform much better, except with a slight upward bias relative to the benchmarks. As for the larger hospitals 4 and 5, CI approximations are too narrow for all cases, with a clear majority of simulated CIs having both bounds within the benchmarks. However, the superiority of variable denominator methods over fixed denominator methods is also clearly observed in the large hospital cases.

At higher variances, most CIs estimated using fixed denominator methods are contained within the benchmark bounds. With a higher variance in expected incidence, the inaccuracy that results in assuming a fixed expected incidence becomes more pronounced. Similar to the low variance case, approximations with variable expected incidence in this study demonstrate a marked improvement over fixed denominator methods for larger hospitals. This is most clearly seen for hospital 5 for methods N1, N2, S1 and S3, where most of bounds are located in categories 1, 2 and 3. As for smaller hospitals with high variance, they proved to be more difficult to approximate well in general. Variable denominator methods failed to produce a CI too often; fixed denominator methods suffered less from this problem because they had narrow CI bounds. Approximation methods surveyed in this study proved inadequate in such cases and better methods are yet to be discovered.

2.6 Discussion

Hospital quality is not something physical that can be directly observed. Instead, it is indirectly measured by counting observable outcomes that are correlated to underlying quality. For a fair assessment, these measurements also need to be adjusted for differences in patient composition between hospitals. This process entails statistical dispersion that needs to be captured. Non-parametric methods are good at doing so, but they are also time-consuming and difficult to use. The alternative is to approximate the statistical dispersion using analytical methods.

The SIR is a typical example of a risk-adjusted statistic in the form of a ratio. Our study examined a number of common analytical methods for estimating its CI as a means to
Two main messages are expressed in these results. First, accuracy of CIs for ratio statistics may be compromised if the method used does not account for the variance of the estimated probabilities. This has direct consequences for conclusions based on these CI estimates. For example, a recent study that compared the additional risk of leukemia from residing close to nuclear power plants found a significant relationship based on CIs of odds ratios computed using Byar’s Method (Sermage-Faure et al 2012). However, Byar’s method does not account for the variance of the denominator, which may be significant considering that the lower bound of the CI for the ratio was just above 1.0, and that the sample size was only 14. By extension, similar risks may be embedded in other studies.

Second, methods that account for the skewness and kurtosis of the SIR distribution seemed to better capture the SIR distribution. The marginal improvement is especially noticeable for larger hospitals. As for hospitals with fewer patients, we recommend conventional methods, ignoring the stochastic property of the denominator. In a high-variance environment with small sample sizes, analytical approximations may fail to generate a solution, while the logistical barriers to bootstrapping is usually much lower for small numbers of patients regardless.

The SIR is more commonly used today as a variable within some larger study. Examples include Stochastic Frontier Analysis and the Malmquist productivity index. If the variability of the SIR is not sufficiently captured, the accuracy of estimated scores may be reduced. This can be shown by introducing a risk-adjusted measure of output quality into the production function, and is the focus of chapters 3 and 4. Datasets are also becoming more detailed and will continue to do so in the foreseeable future. Having more patient information to work with is obviously desirable as it opens up new research opportunities. In that regard, results from this study serve as a reminder that the corresponding statistical challenges must also be handled with care.
Table 2.11: Evaluation of analytical methods: a summary

<table>
<thead>
<tr>
<th>Method</th>
<th>LowVar, SmallHosp</th>
<th>LowVar, LargeHosp</th>
<th>HighVar, SmallHosp</th>
<th>HighVar, LargeHosp</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1, Wilson’s Score</td>
<td>Good</td>
<td>Poor</td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>D2, Byar’s Method</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3, Nave Bootstrap</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>N1, Fieller’s Interval</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>N2, Gaussian Ratio</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S1, Chi-squared Ratio</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td></td>
</tr>
<tr>
<td>S2, Log-adjusted</td>
<td>Poor</td>
<td>Poor</td>
<td>Poor</td>
<td>Good</td>
</tr>
<tr>
<td>S3, Cube-root adjusted</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S4, Edgeworth adjusted</td>
<td>Poor</td>
<td>Good</td>
<td>Poor</td>
<td>Poor</td>
</tr>
</tbody>
</table>

Method Comments
- D1, Wilson’s Score: Good for low variance, and also very high variance cases if solution failure is a possibility.
- D2, Byar’s Method: Dominated by D1 and D2, do not use.
- D3, Nave Bootstrap: Solution fail often, do not use.
- N1, Fieller’s Interval: Dominated by S3, do not use.
- S2, Log-adjusted: Unusable.
- S3, Cube-root adjusted: Good for high variance when small risk of solution failure or only upper bound matters.
- S4, Edgeworth adjusted: Good for low variance, large hospital cases.
Notation used for chapter 2

The following is a list of notation used in this chapter. Common symbols used in economics or statistics may be excluded.

Literature review

- SIR: Standardised Incidence Ratio.
- F, K: Special symbols, uniform distribution motivating example.

Methods, introduction example

- h: Index of hospitals from 1 to H (implied).
- p: Index of patients in a given hospital, from 1 to P ($P_h$).
- n: Index of samples drawn from a given hospital, from 1 to N.

Computing the SIR statistic

- X: Numerator of SIR, defined as the sum of mortalities of patients.
- Y: Denominator of SIR, defined as the sum of probabilities of mortality of patients.
- $\mu_X$: Mean of X.
- $\sigma_X^2$: Variance of X.
- $\hat{\mu}_X$: Estimate of the mean of X.
- $\hat{\sigma}_X^2$: Estimate of the variance of X.
- $\delta$: Vector of patient demographic variables.
- $\lambda$: Parameter estimates corresponding to elements in $\delta$.
- u: Index of patient demographic variables from 1 to U.
- $\mu_Y$: Mean of Y.
- $\hat{\mu}_Y$: Estimate of the mean of Y.
- $\text{Var}(\lambda)$: Variance-covariance matrix of associated with $\lambda$.
- $\sigma_Y^2$: Variance of Y.
• $\hat{\sigma}_Y^2$: Estimate of the variance of Y.
• $\sigma_{XY}$: Covariance of X and Y.
• $\hat{\sigma}_{XY}$: Estimate of the covariance of X and Y.
• $\bar{X}$: Mean of X.
• $\bar{Y}$: Mean of Y.

Methods, simulation

• $\psi$: Coverage value.
• $\mathbb{1}$: Indicator function.
• $\text{lb/ub}$: Lower/upper bound.
• $m$: Index of resamples for method D3 and benchmark, from 1 to M.
• $\epsilon$: Vector of standard errors of elements in $\lambda$.
• $b(\mu_Y)$: Bias-adjustment for $\mu_Y$.
• $s$: Index of bias-adjustment resamples, from 1 to S.

Analytic Methods for estimating the CI of SIRs

• $z$: Normal distribution (various uses).
• $\alpha$: Significance levels.
• $\chi^2$: Chi-squared distribution.
• $t$: Student’s t distribution, for method N1.
• G, $\Gamma$, g1, g2, g3, h: Special symbols for method N2.
• $\eta$, $\theta$, $\bar{r}$, $\bar{p}$, S(), K(), $\zeta$: Special symbols for S4.
Chapter 3

Estimating the cost of improving hospital performance using quality measures with statistical error

3.1 Introduction

Rising costs of healthcare and improving hospital performance are both policy concerns of broad interest across many countries. An important research question is whether improving hospital performance has a significant impact on cost. While this problem has led to a branch of research with a very large and growing literature, no dominant point of view yet emerges. A major reason for the lack of a consensus is caused by differences in methodology and datasets between studies.

To show this, we introduce two methodological variations that are commonly observed in the literature. The first is to account for the stochastic nature of the hospital performance measure by explicitly capturing its error structure using bootstrap regression. The second is to compare the differences that the choice of performance measure can make on the relationship between cost and performance. Two measures are used for the latter: mortality and unplanned readmissions.

For both exercises, we are interested in the results across different specifications, rather than the estimates of the parameters themselves. In the first instance, the bootstrap-
adjusted estimation resulted in a negative and more statistically significant correlation between hospital quality and cost. As for the second exercise, when mortality was replaced by unplanned readmissions as the quality measure, the correlation between the quality and cost largely disappeared.

Cost is the most important factor driving policy decisions for improving hospital quality. Therefore, if the dollar value and the degree of certainty reported to policymakers were to change considerably with a different quality measure during analysis, then the credibility of its conclusion may be compromised. The choice of quality measure is also important because there are truly many measures available, both from economics and from the medical literature that economists can borrow. Furthermore, the correlation between these measures are also weak (Thomas, Ashish 2008), making the choice of what to measure a significant factor influencing hospital performance ratings.

The rest of the paper is organised as follows. Section 3.2 provides a brief overview of the existing literature. Then, the method used for this study is explained in section 3.3. A description of the data source is given in section 3.4, followed by the results from analysis in section 3.5. A concluding discussion is provided in section 3.6.

3.2 Literature Review

The relationship between hospital cost and quality is a research question that is discussed in both health economics and medicine. Economists are well-equipped for capturing the cost structure of hospitals, which is also the primary concern for policymakers. However, this is only a part of the research question. The other part, defining and measuring hospital quality, is a topic that is more comprehensively researched in the medical literature, because patient outcome is the more important metric from the medical professional's point of view. The latter can be inferred by the prevalent use of patient-oriented comprehensive measures such as Patient Safety Indicators (PSI) (Holt et al 2010), both for specific diseases and for hospitals in general. This study draws ideas from the existing literature from both disciplines.

Economists traditionally use cost functions to model the supply of healthcare services from hospitals. A common feature of these studies is the minimalist manner in which the quality differences between hospitals is captured in their specification. An illustrative example is
a study conducted in New York State that examined the cost function of nursing homes (Gertler, Waldman 1992). The main focus of the authors was the underlying cost structure of the nursing homes; quality was a set of external factors to be controlled for. As such, the effects of quality were summarily removed by the use of latent variable methods.

Data limitations are the most commonly cited reason for the lack of a comprehensive specification that accounts for hospital quality. The problem is that most variables that are available and usable for measure hospital quality also tend to have unobserved endogeneity with other variables. If left alone, estimated parameters may be biased, which then leads to inaccurate hypothesis testing results.

An instrumental variable method can be effective at reducing endogeneity, but finding a good instrument is often challenging. This is because a good instrument needs to be closely related to the variable in question, and at the same time uncorrelated with the other variables (i.e. cost and quality). A classic example of an effective instrument is the patient’s geographical distance from the nearest hospital (Gowrisankaran, Town 1999). The rationale is that patients who do not choose the closest hospital had a reason not to.

Occasionally, it is possible to construct an instrumental variable by making modifications to existing ones. An example of this is a study measuring the effect of a reduction in reimbursements on quality of care (Shen 2003), where the change in funding is separated into exogenous and hospital-adjusted components. Time differences can also be used as instruments in hospital cost-quality studies (Carey, Burgess 1999).

Hedonic cost functions have been used in more recent work on hospital cost and quality. It combines a list of variables, or characteristics, and models their interaction with the subject variable using a function that can be freely defined based on the research question. The challenge is to find estimation methods that capture the complex error structure or, sometimes for non-linear hedonic function, finding an algorithm that converges for the data in question.

As an example, an earlier paper on the cost structure of transport networks aggregated a continuum of route distances into discrete groups and then analysed them as different types of outputs (Chiang, Friedlaender 1983). Another way to use hedonic functions is to
aggregate a few values into a single variable using a hedonic aggregator function. A study of the utility of public transport (Wardman 2004) estimated the dis-utility from waiting for the next ride using a statistic called average headway disutility.

Non-parametric methods allow more flexibility in the way hospital quality is captured. As an example, a US public hospital on several states (Clement et al 2008) included five Inpatient Quality Indicators (IQI) into their Data Envelopment Analysis (DEA) as separate parameters. DEA is a deterministic frontier method that measures hospital efficiency, a different but related concept to cost. It also forms the basis for Chapter 4, where the benefits and weaknesses of DEA will be discussed in more detail.

With improved access to government data, there was a renewed interest in empirical questions, including the impact of changes in quality to hospital cost. In the US, a study that investigated the revealed preference of pneumonia patients in Los Angeles (Romley, Goldman 2008) claims that a change in quality at the margins is negative, provided that the correlation between quality and productivity is controlled for. The authors estimated a hedonic translog function, using hospital dummy variables as a measure of quality, with direct access to hospital data from the Office of Statewide Health Planning and Development (OSHPD) in the state of California. Similar studies were soon conducted in other developed countries such as the UK (Dusheiko et al 2011) and Japan (Hashimoto et al 2011).

The current literature lacks a clear consensus as to whether the relationship between cost and quality is positive, negative or not related. Without a representative conclusion, there is no benchmark to use for empirically validating new theories claiming to explain the phenomenon. This may be due to a related problem that there is a lack of discussion about the nature of quality measures.

Many innovations in estimation strategies are related to dealing with error terms, but this is less effective if there is a lack of discussion about the statistical properties of the hospital quality measures themselves. That is, while econometrics is well-equipped with handling error structures, and this study will employ some existing tools to do so, health economics also has an extra issue of measuring hospital quality, a detail that is at least as important as the econometrics itself. To address the latter, we turn to the medical literature for insights.
Early studies on measuring the financial impact of adverse incidents came from medical journals. An example is one by Anderson and Steinberg (1984), studying Medicare expenditures caused by patient readmissions. Expenditures were measured by examining Medicare payments, using a longitudinal database of patients into different groups. Certain demographic and geographical characteristics of patients and hospitals that drove up readmissions were identified, and more stringent controls for patients with these traits were advocated.

Another common topic in the medical literature is the type of incidence to use. There is a wide variety of metrics that can be used to measure patient outcomes (Donabedian 2005). Some examples that are directly related to hospital studies are different types of ambulatory care (Finegan et al 2010) or various hospital-acquired infections. There is little consensus on which type of incidence to count, except that more than one measure should be used if the data and research question allow (Iezzoni 1997). Much of this literature may be instructive to current and future economics research.

Besides selecting an appropriate metric, it is also important to carefully choose a risk-adjustment strategy, a non-trivial decision because it is a current research topic with no consensus. Moreover, observed differences in methods used amongst existing literature are often driven by differences in hospital operation and data collection rather than any theoretical disagreements.

The latter can be demonstrated by the result from a project that investigated 34 different studies as part of a meta-analysis on how estimation strategy affects readmission predictions (van Walraven et al 2011). It turns out that risk-adjusted quality is just as heavily affected by the definition of what constitutes “avoidable” readmissions, most of which were determined by subjective criteria by a small number of individuals, with results varying from 5% to 79%.

Recently, the medical literature made use of techniques in health economics to explore questions in relation to hospital cost and quality. An example is a study written by a group of doctors that measured the marginal costs of reducing patient readmission, while controlling for patient casemix by applying the “30-day odds-ratio” to a selection of acute illnesses that are representative of the dataset (Wong et al 2011). There is also a study that investigated the accuracy of adverse incidence rates for measuring hospital performance,
with a strong emphasis on patient risk-adjustment (Hauck et al 2012).

Applying methods from the medical literature to studies in economies is harder because the former are generally aimed at specific set of patients and their associated problems (Silverstein et al 2008). For example, in the case of patient readmissions, which is a single type of adverse incidence, there are different operational measures such as patient complication events (Encinosa, Bernerd 2005) or potentially avoidable readmission (Medicare Payment Advisory Commission 2007, Chap 5).1

Furthermore, this measure may be different based on the type of disease. A recent paper that reviews the readmission definition of one particular type of rheumatic disease employed an automated search system to read through a large number of relevant papers, counting the features of existing definitions that appear most frequently (Dejaco et al 2011). A study that focuses on the economic impact of a change in quality for hospitals would require definitions that generalise across many disease types.

3.3 Method of Analysis

This section describes how hospital quality is estimated and outlines the main concepts of production functions and the cost function. It also describes how the cost function is estimated and how to incorporate hospital quality measures into the estimation process.

3.3.1 Quality measure and its statistical properties

The statistic used for measuring hospital quality in this study is the Standardised Incidence Ratio (SIR), discussed in chapter 2. The SIR is a generic measure that can be applied to most hospital incidents. This study will use two incidence types, mortality and unplanned readmissions (henceforth referred to as ‘readmissions’). Mortality is clearly defined, while one readmission count is defined in the data as an unplanned revisit within 28 days for the same disease. This is similar to the 30-day time-span by Wong et al (2011).

1The Medicare report cited a commercial service to identify treatments that are linked to initial admissions versus new cases. See http://solutions.3m.com/wps/portal/3M/en_US/3M_Health_Information_Systems/HIS/Products/PPR/
There are \( H \) hospitals in the dataset, indexed \( h = 1, \ldots, H \). Each hospital \( h \) has a specific number of mortality incidents per year \( D_h \) and readmissions \( R_h \). \( D \) and \( R \) are random variables, with corresponding expected values of \( E(D) \) and \( E(R) \). The measure of quality in this chapter is the SIR statistic of each hospital, \( z_D \) and \( z_R \), defined as follows:

\[
\begin{align*}
    z_{Dh} &= \frac{D_h}{E(D_h)} \\
    z_{Rh} &= \frac{R_h}{E(R_h)}
\end{align*}
\]

The SIR spans the range of non-negative values, with smaller values indicating higher hospital quality.

Computing point estimates of \( z_{Dh} \) and \( z_{Rh} \) is straight-forward. However, when the SIR is included in a cost function for estimation, its error structure needs to be accounted for in order to reduce unobserved bias. Standard computed-regressor methods exist for this purpose, but they only work for variables with Gaussian distributions or some other simple functions. This is not the case for \( z_{Dh} \) and \( z_{Rh} \), where each is formed as the ratio of two random variables. This algebraic structure does not lend itself to clean, analytical solutions.\(^2\)

Approximating the SIR distribution using parametric simulation can get around this problem. Assume that the probabilities attached to each patient are mutually independent. The simulation process used here is very similar to the one described in chapter 2; readers can refer to the respective sections there. For this chapter, parametric simulation is repeated for a total of \( N=1000 \) values for each \( z_{Dh} \) and \( z_{Rh} \). Results are denoted as \( z_{Dh1} \) to \( z_{DhN} \) and \( z_{Rh1} \) to \( z_{Rhn} \) respectively. The use of these simulated quality values are explained later.

The process that is specific to this chapter is capturing the distributions of \( D_h, R_h, E(D_h) \) and \( E(R_h) \) that make up the SIR measures as defined in this study. Their distributions are estimated at the individual patient level in a way that is similar to Huack et al (2012), but with a more complicated algebraic structure that accounts for the SIR as a distribu-

\(^2\)This is not to say an analytical solution does not exist, only that the author is not aware of one. Chapter 2 also explains reasons why a solution is not likely to appear.
To begin, patients in a given hospital $h$ are indexed $p = 1, \ldots, P_h$. The mortality status of each patient is a binary variable $0$ or $1$, denoted as $D_{hp}$. Similarly, the readmissions count of each patient is modelled as a non-negative integer $R_{hp}$. The estimate of $D_h$ and $R_h$ of a hospital are denoted as $\hat{D}_h$ and $\hat{R}_h$ respectively, defined as follows:

$$\hat{D}_h = \sum_{p=1}^{P_h} D_{hp}$$

$$\hat{R}_h = \sum_{p=1}^{P_h} R_{hp}$$

Since $D_h$ is the sum of binary variables $D_{hp}$, it follows a binomial distribution with probability of incidence equal to $\hat{D}_h / P_h$. Similarly, $R_h$ is the sum of count variables $R_{hp}$, which follows the Poisson distribution with incidence count rate $\hat{R}_h$.

The other two random variables $E(D_h)$ and $E(R_h)$ in the denominator are defined as the sum of the expected incidence counts of the patients:

$$E(D_h) = \sum_{p=1}^{P_h} E(D_{hp})$$

$$E(R_h) = \sum_{p=1}^{P_h} E(R_{hp})$$

The variables $E(D_{hp})$ and $E(R_{hp})$ are estimated using regression because probabilities are not directly observed. Consider each patient as an entity that carries with them the following properties:

1. Mortality: alive (0) or dead (1).

---

3The Poisson is the simplest count model in common use. Obviously there are many other. Also, the process here as described is actually Poisson-Binomial. This is because the each patient has a different rate of occurrence for unplanned readmissions. But for collectively small p-values it is close enough and saves computation time. See Winkelmann (2008) for a detailed treatment of other count models.
2. Readmissions: a non-negative integer.

3. A vector of demographic characteristics.

4. A disease type, defined when a patient is admitted.

The mortality and readmissions count for each patient are directly observed. Demographic characteristics used for risk-adjustment include age, gender (Leng et al 1999), and the Charlson comorbidity index (Charlson et al 1987). Age and gender are directly taken from data, while the Charlson index is computed using risk-factors as defined in the ICD-10 diagnostic coding system. The marginal effects of these variables on mortality and readmission rates are estimated using regression. Disease type is not included as a variable. Instead, the parameters for patients of each disease type are separately estimated; refer to chapter 2 for a more detailed explanation.

Let \( \delta_{hp} \) be the vector of variables associated with each patient \( p \) in hospital \( h \). Each element in that vector has a corresponding parameter estimate that captures their marginal effect on incidence probability. Denote the vector of parameters for mortality using \( \lambda_D \) and, similarly for readmissions, \( \lambda_R \). Co-variances between parameters are assumed to be zero. Mortality is a binary dependent variable, estimated using Logistic regression:

\[
P(D_{hp} = 1) = \frac{exp(\lambda_D' \delta_{hp})}{1 + exp(\lambda_D' \delta_{hp})}
\]

Unplanned readmissions is a count variable, estimated using the standard Poisson model (Halfon et al 2002):

\[
P(R_{hp} = r \mid E(R_{hp})) = \frac{exp(E(R_{hp}))E(R_{hp})^r}{r!}
\]

\[
E(R_{hp}) = exp(\lambda_R' \delta_{hp})
\]

Both Logistic and Poisson regressions are estimated using standard maximum-likelihood estimation (MLE). This study used STATA v12; it is also available in most major off-the-shelf statistical software.
3.3.2 The hospital production function

Firms are entities that employ inputs and then process them into outputs. Hospitals are an example of such entities because they also employ inputs, such as medical staff and equipment, to produce outputs that treat patients. The process that turns the inputs into outputs is called the production technology. The algebraic way of representing a hospital’s production technology is a production function.

A production function is a relationship between a vector of inputs and a vector of outputs, not necessarily of the same length. It is usually assumed to be continuous, increasing and quasi-concave. Continuity is an approximation to ease computation and is accurate for input and output vectors with large numbers. Increasing means having more inputs will not reduce any outputs, and rests on the free disposibility assumption. The concavity requirement has the practical interpretation that inputs in complement with each other can produce outputs at lower cost than extremes.

Hospitals differ from firms in most other industries because the former are not always profit-maximising entities. The goal of setting up a hospital is often not profit maximisation, even though some hospitals end up with profits. Most physicians are not purely profit-motivated and have some regard for the welfare of their patients. Hospitals in some countries may be obligated to treat emergency cases even if there is a chance the patient cannot pay afterwards. The prices of many medical services are regulated, so hospitals do not have complete freedom to increase revenue in this way. The effect of these factors on hospitals is the increased emphasis on cost-minimisation and providing better quality of care.

The discussion about cost minimisation starts with the cost structure of a hospital, algebraically expressed using cost functions. Let total cost be C, input quantities be x, output quantities as y and input prices as w. The length of the vectors x and w are equal to the number of input types, indexed i = 1,...,I. The length of y is the number of outputs. Let T be the set of possible combinations of x and y. The cost function is defined as follows:

\[ C(w, y) = \min_x \{wx \text{ s.t. } (x, y \in T(x, y))\} \]
where \( \min_x \) is the combination of inputs that produce \( y \) at the lowest cost \( C \).

As an important note, the specification above does not account for hospitals performing below optimal efficiency. This is a strong simplifying assumption that risks conflating discrepancies in efficiency between hospitals with differences in quality. However, addressing this problem would require knowledge of the way quality and efficiency interact, which is an ongoing research topic in itself. Chapter 4 addresses this problem in more detail.

This study will focus on cost functions because many of its relevant properties are similar to production functions. In particular, for the production function, minimising cost and maximising profits are equivalent when hospitals are price-takers. There are various possible reasons for prices to be fixed for hospitals, such as in a regulated environment or when hospitals are competing for inputs in the same market as other producers in the economy. Note that the term fixed prices does not have to mean the same prices for all hospitals, but rather the same price schedule, such that the production choices of a hospital cannot change the prices it faces. For example, larger hospitals might enjoy bulk buying discounts for its inputs, but cannot bargain for further discounts because it is one of many buyers from the supplier.

There are a number of common properties related to cost functions:

1. Non-negative and non-decreasing in \( y \). Algebraically, \( \frac{\partial C}{\partial y_i} \geq 0 \) for all \( i \).
2. Twice (continuously) differentiable on input prices and output units.
3. Non-decreasing in input prices \( w \).
4. Homogeneity of degree one in prices: \( \text{const} \cdot C(w, y) = C(\text{const} \cdot w, y), \forall \text{const} \in \mathbb{R}_+ \)
5. Slutsky symmetry: \( \frac{\partial w_1}{\partial x_2} = \frac{\partial w_2}{\partial x_1} \)
6. Concave in prices: \( \nabla_{ww} C \) is a negative semi-definite matrix.

Non-negative and non-decreasing means producing more \( y \) always costs more. Twice continuously differentiable means input and output units can be infinitely sub-divided. Continuity is an approximation made to ease computation; large pieces of expensive medical equipment obviously cannot be infinitely sub-divided. Non-decreasing in prices means higher prices will never lead to lower total cost to produce the same vector of outputs. Homogeneity of degree 1 means if all input prices are changed by some percentage, the
total cost will also change by the same proportion. Slutsky symmetry excludes state dependence in the interaction terms and depends on an implicit assumption that changing proportions of both inputs is costless.

Additionally, if \( C(w, y) \) is strictly quasi-concave, then Shephard's Lemma holds:

\[
\frac{\partial C}{\partial w_i} = x_i
\]

Shephard’s Lemma is a result of cost-minimisation, which occurs when the quantity demanded for each input is equal to its marginal price. Otherwise, the hospital could have altered the relative proportions of inputs for cost savings as long as the transaction cost is not significant. This is an acceptable assumption as long as the overall efficiency levels of hospitals are not systematically correlated to marginal effects of hospital quality, which will be derived shortly.

### 3.3.3 Functional form specification

The hospital cost function used in this study contains the total cost as the subject variable and all the explanatory factors contributing to the total cost,

\[
C = f(w, y)
\]

where \( w \) is the vector of input prices and \( y \) is the vector of outputs.

The input price vector \( w \) contains three variables: price of capital \( (w_k) \), price of labour \( (w_l) \) and price of materials \( (w_m) \), all with standard economic interpretations. These values are a result of aggregation from a more detailed list of hospital inputs. Aggregation is necessary because cost functions cannot be estimated with the available data if there are too many variables. Similarly, the output measure \( y \) is defined as the sum total of the number of patient-days of all treatments in a hospital. An approximate distribution of patient-days for each episode is shown in Table 3.1.
Table 3.1: Overview of distribution of length-of-stay for hospitals

<table>
<thead>
<tr>
<th>Length of stay</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (sameday)</td>
<td>66.15%</td>
<td>66.75%</td>
<td>67.42%</td>
</tr>
<tr>
<td>2-7 days</td>
<td>24.10%</td>
<td>23.66%</td>
<td>23.17%</td>
</tr>
<tr>
<td>8-28 days</td>
<td>7.95%</td>
<td>7.82%</td>
<td>7.75%</td>
</tr>
<tr>
<td>29+ days</td>
<td>1.80%</td>
<td>1.77%</td>
<td>1.66%</td>
</tr>
</tbody>
</table>

Aggregate length-of-stay percentages for the patient separation data used in this study. More details are available in the data description section.

There are three reasons for using the number of patient-days as a singular output unit. First, preliminary testing was done using simple linear regression, estimating the effect of total cost against number of patient days. The result is that number of patient-days already explains a large portion of variations in total hospital expenses. Second, there was no identifiable pattern in the way parameter estimates changed as the number of output units increased. This makes it difficult to identify the additional heterogeneity that would have been missed when using only one output unit. The standard error of estimates also increased noticeably. For reference, estimates from the two-output and three-output case are presented later on in Section 5.

Specification of a functional form requires prior knowledge of the underlying cost structure of the hospital but this is not always available. A less demanding approach is to approximate the underlying function using series expansion, then estimate parameters that provides the best fit empirically. This study will use the translog function, derived from the second-order Taylor series expansion. Below is a brief exposition, adapted from the original work of Christensen et al (1973).

For ease of exposition, introduce an arbitrary function $f$, where $f = \log(C)$, and an arbitrary vector $\psi$, where $\psi$ is the combined vector of $\log(w)$ and $\log(y)$, indexed from $v = 1,...,V$. Denote $a$ as an arbitrary point in the function $f$, in this case a vector of constants of the same length as $\psi$. Take the second-order Taylor series expansion of $f$ centered around the point $a$, meaning the first two terms of the (infinite) Taylor series expansion,

---

4Other types of series expansions, such as Fourier (Gallant 1981) or Laurent (Caves, Christensen 1980) expansions, have also been advocated. The issue is they generate more terms, consequently requiring more parameters to be estimated, a significant disadvantage for many empirical studies with limited number of data points.
\[ f(\psi + a) = f(a) + \psi \nabla g(a) + \frac{1}{2!} \psi^T \nabla H(a) \psi + O^3 \]

where \( \nabla g(\cdot) \) is the gradient vector, \( \nabla H(\cdot) \) is the Hessian matrix and \( O^3 \) is the sum of the remaining higher-order terms.

The first term is a constant, which we represent using \( \alpha_0 \). The second term contains a vector of first-order derivatives evaluated at point \( a \), each one denoted using \( \alpha_v \). The third term contains a matrix of second-order derivatives also evaluated at point \( a \), each one denoted using \( \beta_{v1v2} \). The final term \( O^3 \) are the sum of the remaining terms and any measurement errors, written as \( \epsilon \).

Recall that the approximation applies for any arbitrary point \( a \), so we can set \( a = 0 \). Using Slutsky’s symmetry, which implies that \( \beta_{v1v2} = \beta_{v2v1} \), we obtain the following expression:

\[ f(\psi) \approx \alpha_0 + \sum_{v=1}^{V} \alpha_v \log(\psi_v) + \sum_{v1=1}^{V} \sum_{v2=1}^{V} \frac{1}{2} \beta_{v1v2} \psi_v \psi_{v2} + \epsilon \]

To obtain the translog cost function, substitute \( \log(C) \) back into \( f \) and \( (w,y) \) back into \( \psi \) respectively:

\[ \log(C(w,y)) = \alpha_0 + \sum_{i=1}^{I} \alpha_{wi} \log(w_i) + \alpha_y \log(y) \]
\[ + \sum_{i=1}^{I} \frac{1}{2} \beta_{ii} \log(w_i) \log(w_i) + \sum_{i=1}^{I} \beta_{iy} \log(y) \log(w_i) \]
\[ + \sum_{i=1}^{I} \sum_{j>i}^{I} \beta_{ij} \log(w_i) \log(w_j) + \epsilon \]

where it is assumed that there are \( I=3 \) input prices and the production unit.
The final step is to add the two proxy measures of hospital quality $z_D$ and $z_R$. This study will assume a basic multiplicative relationship, expressed by an addition sign in log-log form. This leads to the final cost function to be estimated:

$$
\log(C(w, y), z) = \alpha_0 + \sum_{i=1}^{I} \alpha_{wi} \log(w_i) + \alpha_y \log(y) \\
+ \sum_{i=1}^{I} \frac{1}{2} \beta_{ii} \log(w_i) \log(w_i) + \sum_{i=1}^{I} \beta_{iy} \log(y) \log(w_i) \\
+ \sum_{i=1}^{I} \sum_{j>i}^{I} \beta_{ij} \log(w_i) \log(w_j) + \alpha_d \log(z_D) + \alpha_r \log(z_R) + \epsilon
$$

This function belongs to the broader group of multi-output cost functions (Mishra 2007). This method is the usual basis of most existing empirical literature in health economics.

Finally, for the translog cost function to satisfy regularity conditions, some restrictions to the parameter values needs to be added. Some are enforced by default, such as making sure that parameters appearing in different equations have the same values. One of them, linear homogeneity in prices, is enforced by adding four extra conditions:

$$
\alpha_k + \alpha_l + \alpha_m = 1 \\
\beta_{kk} + \beta_{kl} + \beta_{km} = 0 \\
\beta_{kl} + \beta_{ll} + \beta_{lm} = 0 \\
\beta_{km} + \beta_{lm} + \beta_{mm} = 0
$$

We now turn to one estimation strategy that enforces these restrictions.

---

6For an overview of methods used in current literature on cost and efficiency, see Rosko and Mutter (2008).
3.3.4 Estimating the cost function

The translog cost function estimation is usually done using seemingly-unrelated regressions (SUR) (Zellner 1962). To do this, the derivatives of the cost function are estimated. Let $S_i$ represent the percentage share of the total cost contributed by input $i$. Shephard’s Lemma (Shephard 1953) as stated leads to the following:

$$S_i = \frac{\partial \log(C)}{\partial \log(w_i)}$$

This generates three additional equations to be estimated, one for each category of input prices:

$$S_{wi} = \alpha_{wi} + \beta_{ii} \log(w_i) + \sum_{j=1}^{J} \beta_{ij} \log(w_j) + \beta_{iy} \log(y)$$

Combined with the cost function, this leaves a total of four equations with to be estimated by regression. A common estimation technique is to use a two-step feasible generalised least-squares (FGLS) estimator (Zellner 1962) for the translog cost function. Maximum Likelihood is another popular alternative.

SUR as described enforces Slutsky’s symmetry and linear homogeneity in prices. The benefit of SUR is it captures the correlation between the error terms, since these equations are related by construction, whereas estimating them using four separate regressions do not.

Besides the regularity conditions mentioned above, there are others as well. But the translog function does not naturally lend itself to curvature restrictions. It also does not require non-negativity, meaning that it needs to be checked after estimation for each hospital by computing the predicted total cost. Non-decreasing in input prices and output units are also to be checked, by computing their respective predicted first derivatives. Note that these checks only confirm local curvature properties rather than the theoretical global curvature requirements. However, local curvature is usually sufficient because one can argue that the data spans the range of values that are relevant for practical use.
One idea to note is that while the translog function does not enforce global concavity, it is acknowledged here that there are flexible functions that do have this property, such as the Generalized McFadden (Diewert, Wales 1987) and the third-order translog function (Goh, Yong 2006). There are also examples where some of these functions are currently in use specifically for hospital evaluation purposes (Gunning et al. 2011). However, the trade-off is having many more parameters to estimate and is only feasible if the dataset is sufficiently large. There is ample empirical evidence that concavity conditions can be violated (Singh 2006) (Ogawa 2011).

The statistical uncertainty in the quality measures \( z_D \) and \( z_R \) also needs to be accounted for. We do this using a bootstrap regression (Freedman 1981). The advantage over parametric approaches to quality-adjusted hospital studies such as Bayesian hierarchical models (Burgess et al. 2000a) is the ability to maintain a parametric quality specification without being limited to expressions with closed-form solutions. The tradeoff is a reduced ability to make other adjustments such as time-series effects (Burgess et al. 2000b).

Recall that each hospital has two observed quality measures \( z_{Dh} \) and \( z_{Rh} \), each associated with a set of \( B=1000 \) simulated quality values. If each of the observed hospital quality measures is replaced with a simulated one, then the resulting parameter estimates will be different as well.

This study will conduct a total of \( B = 1000 \) bootstrap regressions. For each bootstrap \( b \), where \( b = 1,\ldots,B \), denote the resulting parameters as \( \psi_b \) and the corresponding covariance matrix \( \Omega_b \). These bootstrap-estimates form the distribution of the parameter estimates. The next step is to compute the weighted-average of the parameters, written as \( \hat{\beta} \). The method used here is similar in concept to Lindley and Smith (1972), where each successive bootstrap is treated as additional information added to the prior. This gives us the weighted-average estimate the covariance matrix, estimated as given:

\[
\left( \sum_{b=1}^{B} \Omega_b^{-1} \right)^{-1}
\]

\footnote{For the translog specifically, Diewert and Wales (1987) explains how concavity can be checked at every point, if necessary. Another established method is to enforce ‘local concavity’ as described by Ryan and Wales (2000), where concavity is enforced only at each data point. We consider such restrictions overly artificial.}

58
The weighted-average estimates of $\tilde{\beta}$ follows:

$$
\tilde{\beta} = \left( \sum_{b=1}^{B} \Omega_b^{-1} \psi_b \right) \left( \sum_{b=1}^{B} \Omega_b^{-1} \right)^{-1}
$$

Point estimates from both the unadjusted and adjusted versions will be presented for comparison.\(^8\)

### 3.3.5 Interpreting estimated results

Estimated parameters serve to define the function that is used to approximate the cost structure of the hospitals. For ease of interpretation, they are used to compute statistics of economic significance. This study will measure marginal effects and Allen-Uzawa elasticities, commonly used measures in empirical cost function analysis.\(^9\)

Marginal effects measure the influence that a variable has on total cost. It is an estimate of the increase in total cost when the variable increases by one unit. Four variables are measured for their marginal effects in this study, three input prices and the output unit. The marginal effects of a change in average prices are computed as follows:

$$
\begin{align*}
\frac{\partial C}{\partial w_k} &= S_{wk} \cdot \frac{C}{w_k} \\
\frac{\partial C}{\partial w_l} &= S_{wl} \cdot \frac{C}{w_l} \\
\frac{\partial C}{\partial w_m} &= S_{wm} \cdot \frac{C}{w_m} \\
\frac{\partial C}{\partial y} &= \frac{\partial \log(C)}{\partial \log(y)} \cdot \frac{C}{y}
\end{align*}
$$

\(^8\)Their standard errors, defined by the square root of the diagonal elements in the covariance matrix, are presented next to the point estimates. The off-diagonal elements are omitted for brevity, but they are embedded into the estimation method by construction.

\(^9\)Another common statistic is Global Economies of Scope (GSO). There are two difficulties with computing this statistic in this setting. First, some diseases always require services from different output types within the same hospital, a reality that is not reflected in our cost function. Second, the theoretically correct way to estimate economies of scope is to set the other outputs to 0. However, an output quantity of 0 will show up as $\log(0)$ in the translog function.
Allen-Uzawa (input) elasticities (Uzawa 1962) measure the degree of complementarity or substitution between inputs. When the price of one input increases, the quantity of another input used may increase or decrease. If the statistic is positive, the two inputs are substitutes, otherwise they are complements. There are six statistics in total, three for cross-elasticities and three for self-elasticities. Let $\eta$ represent the elasticity estimate, defined as follows:

$$
\eta_{kk} = \frac{\beta_{kk}}{S_{wk}S_{wk}} + 1 - \frac{1}{S_{wk}}
$$

$$
\eta_{ll} = \frac{\beta_{ll}}{S_{wl}S_{wl}} + 1 - \frac{1}{S_{wl}}
$$

$$
\eta_{mm} = \frac{\beta_{mm}}{S_{wm}S_{wm}} + 1 - \frac{1}{S_{wm}}
$$

$$
\eta_{kl} = \frac{\beta_{kl}}{S_{wk}S_{wl}} + 1
$$

$$
\eta_{lm} = \frac{\beta_{lm}}{S_{wl}S_{wm}} + 1
$$

$$
\eta_{km} = \frac{\beta_{km}}{S_{wk}S_{wm}} + 1
$$

The statistics listed above also have standard errors, reflecting the statistical uncertainty of the parameters used to compute them. The challenge is that each parameter used to compute these statistics are normally distributed, making the distribution of the various combined statistics difficult to solve. Simulation is a viable alternative to the lack of analytical solutions.

1. For each regression parameter, simulate a new set of values, generated using the multivariate normal distribution with mean and covariance matrix as given by the regression.

2. Using the new set of simulated parameters, compute the values of the statistics using the expressions as listed above.

3. Repeat the process 1000 times. The vector of estimates represents the approximation of the statistic’s distribution.

4. The standard error of the vector of 1000 values is presented as the standard error of the statistic.
This study will present statistics for the hypothetical ‘representative hospital’ in the dataset. This is derived by taking the average of each variable of all the hospitals. The purpose of the simplification is to distill the results to essential differences between estimation methods. Obviously each hospital will have different values for computed statistics because their input prices and output quantities differ. However, interpreting statistics for an entire list of hospitals is difficult and distracts from the main research questions.

### 3.4 Data Description

Two data sources are used in this study, one for hospital inputs and the other for patient episodes. Hospital input data are available for a selection of public hospitals in Victoria, Australia from the years 2002 to 2004, while patient episodes are available for six years from 1999-2004. Throughout this study, individual hospitals cannot be identified and summary statistics are presented only in aggregate.

Hospital input data are extracted from financial accounting statements that are available for most public hospitals. To be included in this study, a hospital must have enough entries in their financial statement to compute total expenses, capital(K), labour(L), and materials(M). It must also have had four or more mortality and readmissions incidents. The latter condition serves to filter out small generalist clinics classified as hospitals only in an administrative sense, and to avoid some ambiguous methodological judgements about quality measures.\(^{10}\)

An overview of the vital statistics of hospitals in the dataset are presented in Table 3.2. The first row indicates the average number of episodes in each data point. There is a wide range of sizes, as indicated by the standard deviation of admission episodes. This means when the cost function is estimated, the range of patient numbers is also wider. This is beneficial because the standard error of a prediction made using the estimated equation will be smaller when the variables are within the range of values used during regression. Acute diseases form the majority of cases, a consequence of including hospitals with mortality incidences.\(^{10}\)

\(^{10}\)For example, if the mean of the Poisson distribution in the numerator is less than 4, the lower bound of the 95% CI must be 0, even though hospitals cannot be completely risk-free. A few others are similarly caused by data.
Table 3.2: Selected hospitals in the VAED, 2002-2004: an overview

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of admissions</td>
<td>13690</td>
<td>19033</td>
</tr>
<tr>
<td>— (acute)</td>
<td>11846</td>
<td>17267</td>
</tr>
<tr>
<td>— (maternity)</td>
<td>1421</td>
<td>2777</td>
</tr>
<tr>
<td>— (rehabilitation, alcohol, drugs)</td>
<td>423</td>
<td>783</td>
</tr>
<tr>
<td>Average Length of Stay (LOS), days</td>
<td>4.12</td>
<td>2.25</td>
</tr>
<tr>
<td>— (acute)</td>
<td>4.07</td>
<td>2.40</td>
</tr>
<tr>
<td>— (maternity)</td>
<td>3.11</td>
<td>1.25</td>
</tr>
<tr>
<td>— (rehabilitation, alcohol, drugs)</td>
<td>6.19</td>
<td>3.69</td>
</tr>
<tr>
<td>Mortality rate, by episodes</td>
<td>1.72%</td>
<td>1.22%</td>
</tr>
<tr>
<td>Unplanned readmission, by episodes</td>
<td>1.45%</td>
<td>0.96%</td>
</tr>
<tr>
<td>Avg. expenses ($'000)</td>
<td>69,658</td>
<td>108,687</td>
</tr>
<tr>
<td>— (capital, $'000)</td>
<td>3,665</td>
<td>5,735</td>
</tr>
<tr>
<td>— (labour, $'000)</td>
<td>51,587</td>
<td>79,213</td>
</tr>
<tr>
<td>— (materials, $'000)</td>
<td>14,406</td>
<td>24,635</td>
</tr>
<tr>
<td>Number of hospitals in 2002/03</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>Number of hospitals in 2003/04</td>
<td>83</td>
<td></td>
</tr>
<tr>
<td>Number of hospitals in 2004/05</td>
<td>46</td>
<td></td>
</tr>
</tbody>
</table>

Mean and standard deviation of vital statistics of the hospitals in the VAED sample, 2002-2004. A group of hospitals were merged at the administrative level in 2004. Hospital identities are encoded and cannot be matched from year to year.
Another important observation is the decrease in the number of hospitals during 2004, when administrative changes occurred that resulted in the merging of many smaller hospitals into larger administrative units. The actual ground-level operation did not undergo similar changes, and there is little reason to suspect that a merger on the balance sheet level has any effect on hospital operations. However, since individual hospitals cannot be identified, they cannot be linked across the years either. This means these data points have to be treated as a cross-section and panel-data methods cannot be applied to them. The biggest impact of such limitations is the inability to capture auto-regressive effects.

For patient separations, each hospital in Victoria is recorded in the Victorian Admitted Episodes Database (VAED). These records contain basic information such as demographics, disease diagnostics and how the patient was discharged. The treatments are also categorised into different disease types under various classification systems that can be used for risk-adjustment purposes. For risk-adjustment, this study will separate the patient population into 23 groups at the Major Diagnostic Category (MDC) level. There is also a more detailed level of about 600 Diagnostic-Related Categories (DRG) but many of them are rare or terminal illnesses with very few entries. Refer to Chapter 2 for more details about the VAED.

The patient population is taken from both public and private hospital campuses over six years, summarised in Table 3.3. Both the number and distribution of disease types have been stable throughout the years. The decision to include private hospital patients is to allow a larger patient population to be used for risk-adjustment. A final note is that since patient data for this study begins at the year 2002/03 and ends at the year 2004/05, the readmissions rate for episodes that occurred near the first and last 28 days of the observation period are under-represented.

### 3.5 Results

#### 3.5.1 Cost function estimates

Results in Table 3.5 show a significant negative relationship between quality and total expenses. In contrast, when the bootstrap-adjustment is removed as in Table 3.4, the parameter estimates of the quality variables become positive and is no longer significant. This suggests that the variability of the quality indicators are considerable and that the
### Table 3.3: Summary of patient mortality rates in VAED by year and Major Diagnostic Categories

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Total separations</td>
<td>1,198,434</td>
<td>1,260,530</td>
<td>1,285,402</td>
<td>1,369,548</td>
<td>1,401,942</td>
<td>1,441,641</td>
</tr>
<tr>
<td>-respiratory system</td>
<td>5.299</td>
<td>5.269</td>
<td>5.122</td>
<td>5.135</td>
<td>5.200</td>
<td>5.080</td>
</tr>
<tr>
<td>-hepatobiliary system and pancreas</td>
<td>1.959</td>
<td>1.946</td>
<td>1.928</td>
<td>1.892</td>
<td>1.866</td>
<td>1.883</td>
</tr>
<tr>
<td>-musculoskeletal system and connective tissue</td>
<td>10.278</td>
<td>10.354</td>
<td>10.494</td>
<td>10.240</td>
<td>10.196</td>
<td>10.221</td>
</tr>
<tr>
<td>-skin, subcutaneous tissue and breast</td>
<td>4.851</td>
<td>4.887</td>
<td>4.773</td>
<td>4.954</td>
<td>4.940</td>
<td>4.955</td>
</tr>
<tr>
<td>-Endocrine, nutritional and metabolic diseases and disorders</td>
<td>1.215</td>
<td>1.298</td>
<td>1.345</td>
<td>1.403</td>
<td>1.447</td>
<td>1.464</td>
</tr>
<tr>
<td>-male reproductive system</td>
<td>1.875</td>
<td>1.803</td>
<td>1.809</td>
<td>1.761</td>
<td>1.777</td>
<td>1.832</td>
</tr>
<tr>
<td>-female reproductive system</td>
<td>5.358</td>
<td>5.207</td>
<td>4.779</td>
<td>4.800</td>
<td>4.634</td>
<td>4.451</td>
</tr>
<tr>
<td>-Pregnancy, childbirth and the puerperium</td>
<td>8.446</td>
<td>8.171</td>
<td>7.967</td>
<td>8.335</td>
<td>8.347</td>
<td>8.225</td>
</tr>
<tr>
<td>-Newborns and other neonates</td>
<td>4.434</td>
<td>4.115</td>
<td>3.855</td>
<td>3.736</td>
<td>3.813</td>
<td>3.726</td>
</tr>
<tr>
<td>-blood and blood forming organs and immunological disorders</td>
<td>1.000</td>
<td>1.030</td>
<td>1.115</td>
<td>1.160</td>
<td>1.216</td>
<td>1.281</td>
</tr>
<tr>
<td>-Neoplastic disorders (haematological and solid neoplasms)</td>
<td>1.599</td>
<td>1.578</td>
<td>1.666</td>
<td>1.642</td>
<td>1.712</td>
<td>1.783</td>
</tr>
<tr>
<td>-Infectious and parasitic diseases</td>
<td>1.014</td>
<td>1.067</td>
<td>1.070</td>
<td>1.081</td>
<td>1.085</td>
<td>1.084</td>
</tr>
<tr>
<td>-Mental diseases and disorders</td>
<td>2.462</td>
<td>2.475</td>
<td>2.372</td>
<td>2.322</td>
<td>2.294</td>
<td>2.248</td>
</tr>
<tr>
<td>-Alcohol or drug use and alcohol or drug induced organic mental disorders</td>
<td>0.450</td>
<td>0.423</td>
<td>0.413</td>
<td>0.400</td>
<td>0.395</td>
<td>0.400</td>
</tr>
<tr>
<td>-Injuries, poisoning and toxic effects of drugs</td>
<td>2.332</td>
<td>2.465</td>
<td>2.643</td>
<td>2.467</td>
<td>2.442</td>
<td>2.524</td>
</tr>
<tr>
<td>-Burns</td>
<td>0.163</td>
<td>0.089</td>
<td>0.094</td>
<td>0.100</td>
<td>0.086</td>
<td>0.091</td>
</tr>
<tr>
<td>-Factors influencing health status and other contacts with health services</td>
<td>5.955</td>
<td>4.875</td>
<td>5.298</td>
<td>5.693</td>
<td>5.789</td>
<td>5.907</td>
</tr>
</tbody>
</table>

Summary information of all patients on the VAED list from 1999/00 - 2005/05, including hospitals that are not included in the cost function. Patient types are sorted according to 2004/05 definitions of MDC.
resulting biases from ignoring it during regression are large.

A similar correlation is not observed for readmissions, which is consistent with the lack of literature that includes this indicator. There are two ways to see this. First, for the non-bootstrap version, the parameter estimate is very close to zero. Second, in the bootstrap version, the parameter estimate for readmissions is negative and significant when both quality parameters are included, but moved much closer to zero again when mortality is removed. Even if there is an effect, it may not be significant enough to warrant inclusion in this data.

The cost function also provides information on the marginal effects of input prices and output quality. The first three rows contain marginal effects for the three input categories and they are similar to each other. This suggests that their input share is close to the optimal arrangement. The next three rows contain self-elasticity estimates. Normally they should be negative, and it is indeed the case for $\eta_{kk}$, but not for $\eta_{ll}$ and $\eta_{mm}$. This may be due to qualitative differences in labour and materials, where better medical staff and medication are more highly sought-after. The last three rows contain cross-elasticities of substitution. Labour and materials are complements, while capital is a substitute to the two. It may be that good medical staff can substitute for more expensive capital equipment. These are all interesting possibilities to be considered in future studies.

### 3.5.2 Static comparison exercise

A comparison between estimates generated with and without bootstrap adjustment serves to highlight the importance of controlling for the statistical properties of hospital quality. However, the practical implications of these differences are not easily understood just by inspecting these numbers in isolation. To put these differences in policy context, a static comparison exercise can be used to evaluate the possible financial effects of an adjustment in cost function estimation methods.

The setup is a hypothetical policy intervention whereby hospitals below a certain industry standard for patient outcome are given resources to bring their operations up to a minimum target level:
Table 3.4: Elasticities and marginal effects estimated using SUR, with simulated errors

<table>
<thead>
<tr>
<th>Variable</th>
<th>Both (A4)</th>
<th>Mortality only (A2)</th>
<th>Readmissions only (A3)</th>
<th>(no quality) (A1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point estimate</td>
<td>Standard Error</td>
<td>Point estimate</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$\alpha_D$</td>
<td>0.003367</td>
<td>0.003334</td>
<td>0.003346</td>
<td>0.003331</td>
</tr>
<tr>
<td>$\alpha_R$</td>
<td>-0.000328</td>
<td>0.001681</td>
<td>-0.000265</td>
<td>0.001683</td>
</tr>
<tr>
<td>$\dfrac{dC}{dw}$</td>
<td>19,961 [18849, 21119]</td>
<td>19,950 [18757, 21140]</td>
<td>19,965 [18875, 21087]</td>
<td>19,957 [18774, 21120]</td>
</tr>
<tr>
<td>$\dfrac{dC}{dl}$</td>
<td>13,670 [13583, 13754]</td>
<td>13,671 [13587, 13753]</td>
<td>13,669 [13583, 13755]</td>
<td>13,670 [13577, 13761]</td>
</tr>
<tr>
<td>$\dfrac{dc}{dy}$</td>
<td>0.5473 [0.5423, 0.5523]</td>
<td>0.5472 [0.5426, 0.5518]</td>
<td>0.5472 [0.5423, 0.5521]</td>
<td>0.5472 [0.5474, 0.5498]</td>
</tr>
<tr>
<td>$\eta_{kk}$</td>
<td>-0.7632 [-0.8960, -0.5785]</td>
<td>-0.7632 [-0.9012, -0.5798]</td>
<td>-0.7632 [-0.8972, -0.5754]</td>
<td>-0.7632 [-0.9021, -0.5780]</td>
</tr>
<tr>
<td>$\eta_{ll}$</td>
<td>0.0102 [0.0058, 0.0145]</td>
<td>0.0102 [0.0059, 0.0142]</td>
<td>0.0102 [0.0054, 0.0148]</td>
<td>0.0102 [0.0056, 0.0149]</td>
</tr>
<tr>
<td>$\eta_{mm}$</td>
<td>0.3336 [0.2947, 0.3752]</td>
<td>0.3333 [0.2926, 0.3755]</td>
<td>0.3333 [0.2929, 0.3730]</td>
<td>0.3330 [0.2929, 0.3782]</td>
</tr>
<tr>
<td>$\eta_{kl}$</td>
<td>0.0897 [0.0649, 0.1138]</td>
<td>0.0897 [0.0642, 0.1141]</td>
<td>0.0898 [0.0640, 0.1146]</td>
<td>0.0897 [0.0649, 0.1125]</td>
</tr>
<tr>
<td>$\eta_{km}$</td>
<td>-0.1155 [-0.1360, -0.0959]</td>
<td>-0.1153 [-0.1360, -0.0963]</td>
<td>-0.1157 [-0.1364, -0.0928]</td>
<td>-0.1155 [-0.1365, -0.0945]</td>
</tr>
<tr>
<td>$\eta_{lm}$</td>
<td>0.2138 [0.1614, 0.2663]</td>
<td>0.2137 [0.1584, 0.2707]</td>
<td>0.2139 [0.1532, 0.2683]</td>
<td>0.2138 [0.1568, 0.2697]</td>
</tr>
</tbody>
</table>

Point estimates and estimation errors of quality variables, marginal effects and price elasticities. Four specifications are included: both quality variables, one of each quality variable, or none of the quality variables.
Table 3.5: Elasticities and marginal effects estimated using bootstrap regression, with bootstrapped errors

<table>
<thead>
<tr>
<th>Variable</th>
<th>Both (B3)</th>
<th>Mortality (B1)</th>
<th>Readmissions (B2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point Estimate</td>
<td>Standard Error</td>
<td>Point Estimate</td>
</tr>
<tr>
<td>$\alpha_D$</td>
<td>-0.001602</td>
<td>0.000011</td>
<td>-0.003057</td>
</tr>
<tr>
<td>$\alpha_R$</td>
<td>-0.001920</td>
<td>0.000013</td>
<td>0.000455</td>
</tr>
</tbody>
</table>

| $\frac{\partial C}{\partial d_C}$ | 19,931 [18712, 21115] | 21,115 [18811, 21165] | 18,811 [18514, 21068] |
| $\frac{\partial C}{\partial d_L}$ | 13,674 [13586, 13764] | 13,764 [13589, 13764] | 13,589 [13595, 13774] |
| $\frac{\partial C}{\partial d_M}$ | 14,177 [13995, 14354] | 14,354 [14002, 14354] | 14,002 [13996, 14373] |
| $\frac{\partial C}{\partial \eta_{kk}}$ | 0.5471 [0.0874, 0.0896] | -0.7801 [-0.9246, -0.5753] | -0.7828 [-0.9086, -0.5980] |
| $\frac{\partial C}{\partial \eta_{ll}}$ | 0.0111 [0.0063, 0.0157] | 0.0107 [0.0061, 0.0156] | 0.0110 [0.0064, 0.0158] |
| $\frac{\partial C}{\partial \eta_{mm}}$ | 0.3590 [0.3232, 0.4010] | 0.3492 [0.3139, 0.3845] | 0.3616 [0.3255, 0.4000] |
| $\frac{\partial C}{\partial \eta_{kl}}$ | 0.0921 [0.0622, 0.1195] | 0.0933 [0.0676, 0.1172] | 0.0930 [0.0618, 0.1231] |
| $\frac{\partial C}{\partial \eta_{lm}}$ | -0.1216 [-0.1417, -0.1005] | -0.1190 [-0.1386, -0.0975] | -0.1222 [-0.1431, -0.1006] |
| $\frac{\partial C}{\partial \eta_{km}}$ | 0.2160 [0.1561, 0.2686] | 0.2159 [0.1593, 0.2657] | 0.2156 [0.1603, 0.2729] |

Point estimates similar to Table 3.4, this time with bootstrap regression using simulated quality estimates.
\[ \log(\tilde{C}(w, y), z) = \alpha_0 + \sum_i \alpha_w \log(w_i) + \alpha_y \log(y) \]
\[ + \sum_i \frac{1}{2} \beta_{ii} \log(w_i) \log(w_i) + \sum_i \beta_{iy} \log(y) \log(w_i) \]
\[ + \sum_i \sum_j \beta_{ij} \log(w_i) \log(w_j) + \alpha_d \log(\tilde{z}_D) + \alpha_r \log(\tilde{z}_R) + \epsilon \]

Here the symbols \( \tilde{z}_D \) and \( \tilde{z}_R \) represent the minimum quality level that hospitals need to raise themselves up to.\(^{11}\) The two quality levels are SIR=1.0 and SIR=0.5. Hospitals already operating at or above those levels are left alone. The quantity of interest is the difference in total operating cost between hospitals operating at the observed versus hypothetical levels of quality. The number of hospitals with a starting SIR value above the pre-determined values are listed in Table 3.6:

**Table 3.6: Static comparison: number of hospitals requiring intervention**

<table>
<thead>
<tr>
<th></th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mort.</td>
<td>34</td>
<td>28</td>
<td>29</td>
</tr>
<tr>
<td>Readm.</td>
<td>80</td>
<td>72</td>
<td>77</td>
</tr>
<tr>
<td>Total hospitals</td>
<td>85</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>Total hospitals</td>
<td>83</td>
<td>83</td>
<td>46</td>
</tr>
<tr>
<td>Total hospitals</td>
<td>46</td>
<td>46</td>
<td>46</td>
</tr>
</tbody>
</table>

Number of hospitals with SIR values of above 1.0 and 0.5. The static comparison is a hypothetical policy intervention where hospitals with SIR values above a certain value are given resources to lower their incidence rates.

Two main observations can be gleaned from the static comparison results, displayed in Table 3.7. First, results from the bootstrap-adjusted regressions indicate that improving hospital quality will result in increased hospital cost. This is more consistent with intuition than the result implied by the non-adjusted regression estimates, which says that improving hospital quality can actually lower cost. This is important considering that the estimate from the bootstrap adjusted version is also statistically significant at 95% and not the plain version.

Second, mortality and unplanned readmissions exert a different effect on hospital expenses. Both are measures of hospital quality, suggesting that the financial impact of quality improvements does depend on which variables are used to measure quality.

\(^{11}\)Traditional SIR estimates re-adjust the mean to 1 but this does not affect ranking
Table 3.7: Static comparison: estimated cost of intervention

<table>
<thead>
<tr>
<th></th>
<th>Bootstrap</th>
<th>Non-bootstrap</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>At 100%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mort. Only (B1)</td>
<td>1,317,119</td>
<td>1,374,551</td>
<td>115,160</td>
<td>1,462,539</td>
</tr>
<tr>
<td>Readm. Only (B2)</td>
<td>71,312</td>
<td>134,650</td>
<td>124,109</td>
<td>-</td>
</tr>
<tr>
<td>Both (B3)</td>
<td>987,555</td>
<td>1,282,816</td>
<td>578,500</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>At 50%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mort. Only (B1)</td>
<td>8,892,433</td>
<td>8,888,892</td>
<td>5,061,540</td>
<td>9,826,389</td>
</tr>
<tr>
<td>Readm. Only (B2)</td>
<td>1,275,473</td>
<td>1,357,788</td>
<td>1,108,635</td>
<td>742,374</td>
</tr>
<tr>
<td>Both (B3)</td>
<td>9,985,842</td>
<td>10,329,197</td>
<td>7,283,414</td>
<td>8,968,999</td>
</tr>
</tbody>
</table>

Sum total of the difference in predicted total cost before and after hypothetical policy interventions for hospitals in the data, by year. The bootstrap regression specification is used.

3.5.3 Robustness checks

To complement our main findings, three realistic alternative model specifications are examined as a robustness check, in case there are additional insights that are missed by the main specification as described above.

The first is the number of output types included in the cost function. This is a clear choice because hospitals treat many different types of patient cases and over-aggregation may conceal their differences.

Table 3.8 is similar to the one-output version, except for the addition of the Returns to scale (RTS) statistic, which is measure of the proportional change in total cost in response to a change in output. It is defined as follows for the two-output and three-output cases:

\[
RTS_2 = 1 - \frac{\partial \log(C)}{\partial \log(u_1)} - \frac{\partial \log(C)}{\partial \log(u_2)}
\]

\[
RTS_3 = 1 - \frac{\partial \log(C)}{\partial \log(v_1)} - \frac{\partial \log(C)}{\partial \log(v_2)} - \frac{\partial \log(C)}{\partial \log(v_3)}
\]

A positive value indicates that hospitals should be expanded, while a negative value suggests that they should be scaled down.

The two-output and three-output specifications do not seem to be well-supported. The variables \(\frac{dC}{du_1}\), \(\frac{dC}{dv_2}\) and \(\frac{dC}{dv_3}\) are negative, which violates regularity conditions.

\[\text{12}\] The grouping of patients into output types, based on their MDC disease code, is based on a similar study conducted by the Productivity Commission of Australia. See chapter 4 for more details.
Table 3.8: Robustness check: more output types, SUR estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Point estimate</th>
<th>Standard Error</th>
<th>Variable</th>
<th>Point estimate</th>
<th>Standard Error</th>
<th>Variable</th>
<th>Point estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_D$</td>
<td>0.003367</td>
<td>0.003334</td>
<td>$\alpha_D$</td>
<td>-0.014954</td>
<td>0.001536</td>
<td>$\alpha_D$</td>
<td>-0.016148</td>
<td>0.002386</td>
</tr>
<tr>
<td>$\sigma_D$</td>
<td>-0.000328</td>
<td>0.001681</td>
<td>$\sigma_D$</td>
<td>0.009340</td>
<td>0.000540</td>
<td>$\sigma_D$</td>
<td>0.014840</td>
<td>0.002133</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Point estimate</th>
<th>Confidence Interval</th>
<th>Variable</th>
<th>Point estimate</th>
<th>Confidence Interval</th>
<th>Variable</th>
<th>Point estimate</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$dC_m$</td>
<td>0.5473</td>
<td>[0.5423, 0.0897]</td>
<td>$dC_m$</td>
<td>-0.0541</td>
<td>[-0.0735, -0.0358]</td>
<td>$dC_m$</td>
<td>0.0932</td>
<td>[-0.0186, 0.1982]</td>
</tr>
<tr>
<td>$dC_u$</td>
<td>0.0700</td>
<td>[0.0261, 0.1116]</td>
<td>$dC_u$</td>
<td>-0.0541</td>
<td>[-0.0735, -0.0358]</td>
<td>$dC_u$</td>
<td>-3.9698</td>
<td>[-4.0364, -3.9041]</td>
</tr>
<tr>
<td>$dC_v$</td>
<td>0.0700</td>
<td>[0.0261, 0.1116]</td>
<td>$dC_v$</td>
<td>-0.0541</td>
<td>[-0.0735, -0.0358]</td>
<td>$dC_v$</td>
<td>-3.9698</td>
<td>[-4.0364, -3.9041]</td>
</tr>
</tbody>
</table>

| $\eta_{kk}$ | -0.7632        | [-0.8960, -0.5848] | $\eta_{kk}$ | 0.2266        | [0.2202, 0.2328]    | $\eta_{kk}$ | 0.8537        | [0.8518, 0.8558]    |
| $\eta_{ll}$ | 0.0102         | [0.0058, 0.0146]    | $\eta_{ll}$ | 3.4549        | [3.4262, 3.4849]    | $\eta_{ll}$ | 0.1125        | [-0.1218, -0.1041]  |
| $\eta_{mm}$ | 0.3336         | [0.2947, 0.3743]    | $\eta_{mm}$ | -0.2870       | [-0.2923, -0.2805]  | $\eta_{mm}$ | -3.3471       | [-3.4746, -3.2219]  |
| $\eta_{kl}$ | 0.0897         | [0.0649, 0.1133]    | $\eta_{kl}$ | 0.7998        | [0.7897, 0.8101]    | $\eta_{kl}$ | 1.0353        | [1.0347, 1.0359]    |
| $\eta_{km}$ | -0.1155        | [-0.1360, -0.0957]  | $\eta_{km}$ | -2.0566       | [-2.1133, -2.0217]  | $\eta_{km}$ | 5.3026        | [5.2263, 5.3838]    |
| $\eta_{lm}$ | 0.2138         | [0.1614, 0.2688]    | $\eta_{lm}$ | 0.8769        | [0.8578, 0.8872]    | $\eta_{lm}$ | 0.9176        | [0.9142, 0.9211]    |

| RTS | 1.0817         | [1.0473, 1.1184]    | RTS | 2.5840         | [2.4176, 2.7641]    |

Quality variables, elasticity estimates and scale/scope of production when including more than one output type. Output types are based on a similar study by the Productivity Commission on hospital efficiency in the years 2010 and 2011. v1, v2 and v3 stands for acute, maternity and rehabilitation type diseases in the three-output model. u1 and u2 are used for the two-output model, with u2=v2+v3.
The changes in elasticity estimates from one to two output types do not follow the same trend as the changes from two to three outputs either. This means that the change in elasticity estimates is probably not capturing additional information that was concealed by aggregating all outputs into one type, but rather instability in the model specification itself. Therefore, without additional data, it is better to rely on the single-output specification.

The second is the possibility of structural differences within the set of hospitals. To explore this possibility, a finite mixture model is estimated for a Cobb-Douglas cost function to see if hospitals naturally fall under a number of groups. Results in Table 3.9 indicate that there may be two latent types of hospitals, differentiated by sensitivity to input prices. There is however no easily identifiable traits predicting which group each hospital is classified under. There are also too few observations in the smaller group for a separate cost function estimation.

The third is the potential significance of interaction effects between quality parameters.

---

13Convergence difficulties were encountered during the maximum likelihood estimation when estimating the FMM model for the translog specification. No trial-and-error attempt was made in this study for alternative starting values to see if the process can converge to an estimate. See Karlis, Xekalaki (2003) for more details.
and other variables in the cost function. If such interactions are significant, omitting these terms may reduce the accuracy of the parameter estimates and exacerbate any endogeneity problems. To check for this, interaction terms of $z_d$ and $z_r$ with $y$ and $w$ are included in the expanded cost function specification. This extension also introduces a modification to the share equations:

$$S_{wi} = \alpha_{wi} + \beta_{ii} \log(w_i) + \sum \beta_{ij} \log(w_j) + \beta_{Dw} \log(z_d) + \beta_{Rw} \log(z_r) + \beta_{iy} \log(y)$$

as well as elasticity estimates for quality:

$$\frac{dC}{dD} = (\beta_{Dw} \log(w_k) + \beta_{Dw} \log(w_l) + \beta_{Dw} \log(w_m) + \beta_{Dy} \log(y)) \cdot \frac{C}{D}$$

$$\frac{dC}{dR} = (\beta_{Rw} \log(w_k) + \beta_{Rw} \log(w_l) + \beta_{Rw} \log(w_m) + \beta_{Ry} \log(y)) \cdot \frac{C}{R}$$

According to the results in Table 3.10, the relationship between standardised mortality, readmissions and cost has now changed signs. However, the standard errors of all the marginal effects and elasticities are also noticeably higher, a direct result of the high standard errors of the interaction parameters. This, in conjunction with results from the two-output and three-output specifications, suggests that the data does not support additional parameters in the cost function. A log-likelihood ratio test of the significance of the interaction terms in the non-bootstrap model also failed to reject the null hypothesis ($p=0.2690$).

### 3.6 Discussion

Predicting the financial impact of improving public hospital care is an established research question, but without an agreed answer. Empirical studies differ in method and hospital sample used, making cross-comparison difficult. To fully control for these differences, the relative merits of most major empirical methods in the literature will have to be evaluated, which is beyond the scope of a single study.
Table 3.10: Robustness check: interaction terms between quality and other variables

<table>
<thead>
<tr>
<th>Point estimate</th>
<th>Confidence Interval</th>
<th>Point estimate</th>
<th>Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{\partial r}{\partial C}$</td>
<td>94</td>
<td>[-824 , 987]</td>
<td>-290</td>
</tr>
<tr>
<td>$\frac{\partial r}{\partial D}$</td>
<td>93</td>
<td>[-376 , 522]</td>
<td>422</td>
</tr>
<tr>
<td>$\frac{\partial r}{\partial p}$</td>
<td>19,990</td>
<td>[18766 , 21283]</td>
<td>20,131</td>
</tr>
<tr>
<td>$\frac{\partial r}{\partial w}$</td>
<td>13,604</td>
<td>[13462 , 13754]</td>
<td>13,657</td>
</tr>
<tr>
<td>$\frac{\partial r}{\partial y}$</td>
<td>14,547</td>
<td>[14034 , 14970]</td>
<td>14,143</td>
</tr>
<tr>
<td>$\eta_{kk}$</td>
<td>0.5469</td>
<td>[0.5406 , 0.5532]</td>
<td>0.5475</td>
</tr>
<tr>
<td>$\eta_{ll}$</td>
<td>-0.7690</td>
<td>[-0.9328 , -0.5518]</td>
<td>-0.7890</td>
</tr>
<tr>
<td>$\eta_{mm}$</td>
<td>0.0094</td>
<td>[0.003 , 0.0155]</td>
<td>0.0093</td>
</tr>
<tr>
<td>$\eta_{kl}$</td>
<td>0.2547</td>
<td>[0.1333 , 0.4084]</td>
<td>0.3348</td>
</tr>
<tr>
<td>$\eta_{lm}$</td>
<td>0.0873</td>
<td>[0.0532 , 0.1180]</td>
<td>0.0949</td>
</tr>
<tr>
<td>$\eta_{km}$</td>
<td>-0.1015</td>
<td>[-0.1364 , -0.072]</td>
<td>-0.1168</td>
</tr>
<tr>
<td>LR test: log(qD)log(pw) = logqRlog(y) = logqRlog(pw) = logqRlog(y) = 0; p=0.2690</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Marginal effects of quality and elasticities of the cost function with interaction terms for quality.

We focus on parametric cost functions with adjustments for hospital quality. Two issues about estimating hospital cost functions are addressed: the choice of quality measure, and the statistical variability of risk-adjusted measures.

### 3.6.1 Main findings

We have made the argument that future studies on hospital cost structure should explicitly allow for the stochastic property of quality and other computed variables. To show this, the same cost function was estimated with and without bootstrap-adjustment. In the bootstrap-adjusted version, there is a negative correlation between risk-adjusted incidence rate and hospital cost. In contrast, when the adjustment is removed, the correlation became positive. This comparison provides a reference to the literature supporting the notion that the distribution of quality measures needs to be captured. It also forms the basis for exploring the stochastic properties of other commonly-used quality measures.

As an example, consider again the study by Hashimoto et al (2011) mentioned earlier on hospital spending and quality in Japan. Their main research question is why Japan is able to have superior quality for lower costs as compared to other OECDs. The core analysis consists of a log-log constant elasticity function where hospital quality is the dependent variable and hospital characteristics are the explanatory variables. One of
their central measures of quality, standardised mortality ratio, is an estimated measure with standard errors that are not accounted for. This may distort the evaluation of the quality of individual hospitals in the analysis, possibly resulting in less accurate parameters for determinants of quality.

We have also reinforced the importance of the choice of variable used to measure hospital quality by comparing between two separate variables, mortality and unplanned readmissions. The two variables had different marginal effects on hospital cost, even under controlled experiment settings, where the same dataset, cost function and estimation strategy were used. This demonstrates that quality is a multi-faceted concept and should not be aggregated into a single index. As an extension, it is also possible to explore other, more effective methods for capturing the multiple dimensions of quality in an empirical setting.

The weakness of aggregating quality measures into a single index can be further illustrated using the study conducted by Clement et al (2008) mentioned previously. Their main message is that the negative impact of poor quality on hospital efficiency is underestimated in existing literature. The basis for their argument is that they found a stronger correlation using a multiple-indicator measure as compared to a more popular single-indicator measure.\textsuperscript{14} However, since their analysis did not include estimating the impact of individual components in the aggregated measure, there is a chance that their result was driven by a couple of outlier components rather than a general underlying relationship.

### 3.6.2 Limitations

Depending on data availability, we have identified three viable extensions to improve the robustness of the results.

The first and most important one is to account for any endogenous relationships between hospital quality and cost more explicitly. Endogeneity happens when the error terms between two or more variables are correlated, distorting the estimates for the cost functions. This can happen, for example, when sicker patients choose better hospitals or when hospitals select healthier patients to treat. Therefore, the extent of the endogeneity issue

\textsuperscript{14}Acute Myocardial Infarction Indicator (AMI).
depends on how heavily constrained public hospitals are in terms of patient selection as compared to private ones.

Obviously the best practice is to use an instrument, but finding a good one is always difficult, more so for quality because it is a computed value and not something directly observed. If there is a good understanding of which variable(s) is endogenous and their relationship, traditional methods like maximum-likelihood or 2SLS, or even further bootstrapping, may reduce endogeneity bias. The lack of a conceptual understanding increases the difficult of finding an econometric solution.

The second one depends on having better access to financial information to measure hospital costs. This is relevant in two ways. One, financial statements of public hospitals are recorded in administrative units rather than the number of physical locations. Some of these broad administrative units contain multiple hospital campuses. This study applied a straight proportional appropriation of dollar costs by the number of output units. This may explain the small RTS results in the multi-output alternative specifications. Two, financial statements only state input variables in dollars. Without quantity, prices are approximated by total dollar amounts divided by number of output units. This may cause differences in RTS to be mis-attributed to a difference in operational scale, when in fact these variations might have been due to differences in price-schedules enjoyed by larger hospitals instead.

The third one depends on the availability of private hospital financial statements, and is related with the attribution of fault for patient incidents. In this study, when patients suffer mistakes in hospitals, the risk-adjustment process assumes that the hospital in which a patient suffers a medical incident is solely to blame for it.

It is easy to imagine scenarios where this can be misleading. For example, most hospitals in Victoria are smaller, regional ones that deal with less serious problems. When a patient suffers complications, either because the illness is more serious or because the regional center provided inappropriate care, the patient is almost always transferred to a larger hospital as soon as possible.

Few existing studies that we know of mention the third issue or address it adequately.
Transferred patients may simply be excluded from the analysis, resulting in sample bias (Groeger et al 1999). Alternatively, a simple indicator variable may be added into the list of variables used in the patient risk-adjustment process (ACSQHC 2012). This adjustment captures some patient bias but the attribution issue remains; the same study later on indicates that the inclusion or exclusion of transferred patients has a significant impact on expected mortality rates.
Notation used for chapter 3

The following is a list of notation used in this chapter. The most basic symbols used in Economics or Statistics may be excluded.

Quality variables:

- **h**: Index of hospitals from 1 to H.
- **D_h**: Number of observed deaths.
- **R_h**: Number of observed readmissions.
- **E(D_h)**: Number of expected deaths.
- **E(R_h)**: Number of expected readmissions.
- **z_{Dh}**: Risk-adjusted incidence count, mortality.
- **z_{Rh}**: Risk-adjusted incidence count, unplanned readmissions.
- **n**: Index of quality measure simulations from 1 to N.

Risk-adjustment parameters:

- **p**: Index of patients from 1 to \( P_h \).
- **δ_{hp}**: Vector of patient characteristics of patient p in hospital h.
- **λ_D**: Vector of parameters for patient characteristics of patient p in hospital h, for mortality.
- **λ_R**: Vector of parameters for patient characteristics of patient p in hospital h, for unplanned readmissions.

Production function theory:

- **C**: Total cost.
- **x**: Vector of input quantities.
- **y**: Vector of output quantities.
- **w**: Vector of prices for corresponding input types.
• T: Possible pairs of input and output quantities.
• \( \min_x \): Combination of \( x \) producing \( y \) at lowest \( C \).

Cost function variables:
• \( i, j \): Index for inputs and their input prices from 1 to \( I \).
• \( k \): Capital.
• \( l \): Labour.
• \( m \): Materials.

Translog cost function:
• \( \psi \): Special symbol, Taylor expansion.
• \( v \): Index of cost function variables, from 1 to \( V \).
• \( a \): Special symbol, Taylor expansion.
• \( \nabla g(a), \nabla g(h) \): Gradient / Hessian matrix, Taylor Expansion.
• \( \alpha \): Cost function parameters (first-order).
• \( \beta \): Cost function parameters (second-order).
• \( \epsilon \): Error term, cost function.
• \( S \): Share equations.

Estimation strategy:
• \( b \): Index of bootstraps, from 1 to \( B \).
• \( \Omega \): Covariance matrix.

Others:
• \( \eta \): Allen-Uzawa elasticities.
• RTS: Returns to Scale.
• \( t, u, v \): Robustness check, multiple output types.
Chapter 4

Hospital Quality, Disease Aggregation and Risk Adjustment

4.1 Introduction

Risk-adjustment is an essential part of any study that seeks to improve hospital quality. Hospitals treat patients with vastly different diseases and underlying traits, so the yardstick used to measure quality also has to account for such differences. Consequently, the investigation of statistical methods for doing so is an established branch of research in health economics.

Disease aggregation is the process of sorting individual patients into groups, necessary for computing risk-adjusted quality. All patients are arguably unique, but some aggregation is necessary due to data and computational constraints, even though it leads to some loss of information. There are also many different ways to aggregate patients in the existing literature. However, there is little in the literature that investigates how the choice of disease aggregation method may affect quality estimates.

This study aims to demonstrate the importance of disease aggregation, or the classification of patients, in the context of risk-adjustment for empirical analysis. First, risk-adjusted hospital quality measures are estimated using three different levels of disease aggregation,
ranging from broader to narrower groups. Then, these values are used to estimate the impact that a change in quality may have for a typical empirical study. The object of demonstration is the correlation between technical efficiency and quality, because it is a topic of widespread and immediate policy interest. A standard two-stage model is used, first to estimate efficiency with data-envelopment analysis, then to determine how quality drives efficiency scores using Tobit regression.

When patients are sorted into smaller and more detailed disease aggregations, the differences in quality between hospitals were reduced. This result is consistent with the literature. However, the variance of the risk-adjusted measures also becomes higher, suggesting that there may be a bias-variance trade-off involved. In the subsequent empirical analysis, the relationship between efficiency and quality were different across disease aggregation levels at the 95% significant level. Both of these results confirm that the choice of disease aggregation can have a significant impact on empirical estimates.

Traditionally, the biggest barrier against accurate risk-adjustment has been the lack of data. Researchers struggled to obtain enough data to conduct their study in the first place, so any discussions about disease aggregation strategy were largely irrelevant. But as data quality and accessibility improve, more decisions will have to be made during each step of the study, and disease aggregation is one of them. This study also advocates that future studies should either include a robustness check whereby another level of disease aggregation is used, or provide justification for the chosen aggregation.

The rest the paper is organised as follows: section 4.2 provides a literature review, section 4.3 explains the methodology, section 4.4 provides a data description, section 4.5 presents the results, section 4.6 summarises the findings and potential for further investigation.

4.2 Literature Review

The earliest research on patient risk-adjustment may have originated from the medical literature, where researchers have been investigating risk adjustment methods for a very long time. One of the earliest well-known examples was a discussion about how the unit of time measurement can produce very different incidence measures (Farr 1838). What makes it interesting is its perceived relevance even after all the years, along with its supporters
(Vandenbroucke, Vandenbroucke 1996) and detractors (Iezzoni 1996). More recently, the level of interest in statistical methods for computing them seemed to have decreased over the years (Iezzoni et al. 1996). Instead, current medical studies prefer to adapt tools that already exist in other disciplines.

In contrast, the economics literature traditionally focused more on the consequences of varying levels of hospital performance than the process of performance evaluation itself. As such, more simplistic measures such as a simple count of the number of secondary diagnosis (Dranove, Ludwick 1998) are commonly used. Details such as disease aggregation are seldom discussed, except when there is a known empirical issue in a specific context.

An example of the latter is a study concerning the way health service utilisation is captured when more than one database system is involved (Burgess et al. 2011). The two example databases are the Department of Veterans Affairs and the Medicare system, during which the authors had to adjust both disease aggregation systems to make them comparable. It is also one of the few studies where the construction of an aggregation system is carefully considered. Unfortunately, the context of the study did not lend itself to consider the implications it has on risk-adjustment.

Another motivating example of the potential significance of disease aggregation comes from a paper investigating a policy change that demanded increased transparency in hospital quality statistics (Dranove et al. 2003). Their aim was to measure the effect of increased transparency on physician and patient choice, with a concern that difficult patients would be turned away by hospitals because accepting them may have a negative impact on their hospital report card. The performance evaluation method is a secondary issue.

However, readers from outside economics may wonder whether the risk-adjustment mechanism behind the quality measure sufficiently controlled for the differences in patient makeup. This question is important because, if the quality measure was fair to hospitals treating sicker patients, then they may be less averse to accepting high-risk patients even with increased transparency. Therefore, some form of robustness checking to verify the validity of the risk-adjustment process would have strengthened the results of that study. And if such robustness checks were in place, the disease aggregation method would have definitely been examined, as it is one of the most important factors for patient risk-adjustment.
Hence, we believe that this is a suitable empirical setting to demonstrate the importance of disease aggregation. The rest of this section provides an overview of common methods for estimating quality-adjusted hospital efficiency.

A cost function is the most common way of measuring hospital efficiency. For example, a study looking at a sample of hospitals operating in the US in 1983 estimated a translog cost function for efficiency scores (Morey et al 1992), followed by estimating an efficiency-quality frontier function. Their results show that reducing hospital incidents will lead to lower costs. An open question is if disease aggregation matters in this context.

Another common measure of efficiency is via distance functions. The simplest form is the radial distance function (Farrell 1957, Shephard 1970), from which most efficiency measures in use today are derived from. One example is a decomposed Malmquist Index with quality (Färe et al 1995) that separates efficiency gains from technological changes. The Malmquist index is useful when the data span many years. The distance function definition of efficiency also leads to the use of frontier methods. An advantage of such methods is that they do not require prices. This is important for situations where accurate prices are not available, such as when analysing public hospitals.

Stochastic frontier analysis (SFA) is an example of a commonly used frontier method. SFA estimates a cost function with two error terms, one representing random errors spanning the real line while the other representing hospital efficiency spanning only the positive half of the real line. An example of SFA is a study that measures the impact of hospital inefficiency on risk-adjusted mortality rates (Deily, McKay 2006). By comparing cost functions with and without the inefficiency error term, it was shown that the relationship between efficiency and mortality rates are weaker than once believed.

Another example closely related to our study is a series of reports on hospital efficiency that was produced by the Productivity Commission (PC) of Australia (Productivity Commission 2010).\footnote{The PC reports have data on private hospitals and additional variables, although some of their parameter estimates are confidential and only accessible to insiders.} There, an entire chapter is dedicated to examining the quality-efficiency relationship, using a multiplicative adjustment on quality (Zuckerman et al 1994). The PC study is noteworthy not just because of similar data sources, but also because they make two general claims related to this study. First, public hospitals that are larger tend
to have higher efficiency scores. Second, hospitals with higher standardised incidents have lower efficiency. This study would complement this line of research by comparing what happens when different disease aggregation methods are used.

Non-parametric methods such as data-envelopment analysis (Banker et al 1984) (DEA) can also be used to estimate hospital efficiency. The benefit is it does not require excessive prior assumptions about the hospital’s production process in order to establish an algebraic expression to estimate. However, it is especially susceptible to outliers because the estimation process is completely deterministic, which is an important weakness when detecting outliers is difficult. The use of DEA is discussed in more detail in the next section.

4.3 Method of Analysis

This analysis consists of two parts: evaluating hospital quality, and estimating the marginal effect of hospital quality on efficiency scores. For the first part, hospital quality is evaluated using risk-adjusted measures. The goal is to compare hospital quality values estimated from different levels of disease aggregation and, if possible, to provide reasons for their differences. For the second part, the empirical setting is to estimate quality-adjusted efficiency values and to examine a number of factors that may explain these values. Once again, the emphasis is on comparing and contrasting results derived from different aggregation levels.

4.3.1 Risk-adjusted quality and disease aggregation

Hospitals with sicker patients are expected to have higher incidence rates. This can be predicted using the characteristics of the patients being treated. Consistent with previous chapters, the Standardised Incidence Ratio (SIR) is used as the measure of hospital quality. The numerator is the total number of mortality incidents in a hospital. The denominator is the sum of the expected mortalities of all its patients. The expected probabilities are estimated using Logistic regression. Variables used include age, gender (Chambers, Clarke 1990) and the ICD-10-based Charlson index. Each disease aggregation is separately estimated.

Let h and p index respectively hospitals and patients, h = 1,...,H and p = 1,...,P. Each
patient has a vector of demographic characteristics $\delta_{hp}$, with the marginal effect of each characteristic represented by the corresponding element in the vector $\lambda$. Patients are admitted into hospitals and then discharged either alive ($D_{hp}=0$) or dead ($D_{hp}=1$). The probability of death for patient $p$ in hospital $h$ is represented by $ED_{hp}$. The functional form for the Logistic regression is given as follows:

$$ED_{hp} = \frac{e^{\lambda'\delta_{hp}}}{1 + e^{\lambda'\delta_{hp}}}$$

Assume that the expected number of deaths in a hospital is the sum of the probabilities of all its patients. It is denoted as $ED_h$, defined as follows:

$$E(D_h) = \sum_{p=1}^{P_h} ED_{hp}$$

The definition of SIR for each hospital follows:

$$SIR_h = \frac{D_h}{ED_h} \in [0, \infty)$$

Each hospital will have an SIR estimate for each level of disease aggregation. The higher the SIR, the worse the hospital quality is. A total of three increasingly refined disease aggregation levels are used. They are explained in more detail in the data description. This change in aggregation level will be reflected in different estimates of $ED_h$ and consequently SIR values.

The importance of disease aggregation on quality statistics is easy enough to see by direct comparisons. However, if they turn out to be different, simple comparisons would not indicate whether broader or narrower aggregations are preferred. The key is whether patients in broader aggregations share similar risk characteristics. An informal way of doing this is to group together the narrower aggregations into broader ones based on risk characteristics, and then compare the result with the existing administrative definitions of the broader aggregations as defined in the patient dataset.
Risk characteristics are represented by estimates of $\lambda$. More specifically, the probability of death for each patient in each aggregation is estimated using separate regressions, so we can write the parameter estimates as $\lambda_a$, where $a$ is the index of aggregations from 1 to $A$ from the broadest to the narrowest aggregation levels. Thus narrower smaller aggregations can then be categorised into broader ones based on how similar their parameter estimates are.

The following is a brief restatement of an aggregation method commonly used for summarising data with many variables (see Hirschberg et al 1994). Let $\epsilon_a$ be the variance-covariance matrix of the parameter estimates, so that parameters with a lower standard error receive more weight. Each element in $\lambda$ is indexed using $u$, where $u = 1, \ldots, U$.\footnote{For this study $U=4$, similar to chapters 2 and 3. They include age, gender, Charlson Index and the constant term.}

Starting with the smallest disease aggregation method, we write the distance between two arbitrary disease aggregations 1 and 2 as $c_{12}$, where ‘distance’ is defined as follows:

$$c_{12} = \frac{1}{U} (\lambda_1 - \lambda_2)'(\epsilon_1 + \epsilon_2)^{-1}(\lambda_1 - \lambda_2)$$

This statistic is computed for each pairwise combination of disease aggregations. Then, the two disease groups with the smallest distance are combined, by the reasoning that they have the highest probability of actually belonging to the same disease group. This is done by combining the parameter estimates of these two disease groups. Again, without loss of generality, the vector of combined parameters is written as $\lambda_3$ and its variance-covariance matrix $\epsilon_3$, both of which are computed as follows:

$$\lambda_3 = (\epsilon_1^{-1} + \epsilon_2^{-1})^{-1} (\epsilon_1^{-1} \lambda_1 + \epsilon_2^{-1} \lambda_2)$$

$$\epsilon_3 = (\epsilon_1^{-1} + \epsilon_2^{-1})^{-1}$$

Each iteration reduces the total number of aggregations by one, and can be iterated until a target number of aggregations is reached. For this study, the iteration ends when the number of clusters is reduced to the same number of aggregations as the next broader level, as defined by the patient dataset. The interpretation is that, if the components of the broader aggregations in the patient dataset are similar to the ones formed by this
process, then it is a sign that patients in the broader aggregations share similar risk characteristics.

4.3.2 Empirical setting for risk-adjusted quality measure

Besides direct comparisons, hospital quality measures are also used as variables for answering other research questions. Therefore, if the estimated values of quality variables can be affected by disease aggregation method, then they might also have an effect on the results of any subsequent estimation. This study will investigate this possibility using a hospital efficiency study, a common research question in health economics. The goal is to compare results across broader versus narrower disease aggregation methods.

We use a two-stage model with a first-stage DEA followed by a second-stage Tobit regression bounded by full efficiency. This gives us two sets of results to work with, the technical efficiency of each hospital and factors driving those efficiency scores. The first stage is a DEA frontier model solved using linear programming. Details of the estimation process can be found in operations research textbooks (e.g. Charnes et al (1994), Ramanathran and Ramanathan (2003)); hence, only the intuition is provided here. The second stage is a regression where the subject variable is bounded by a certain value. This is useful because technical efficiency is a quantity bounded between 0 and 1.

In DEA, each hospital contains a vector of inputs and a vector of outputs. The combination of inputs and outputs of each hospital is compared against other hospitals. If there is no convex combination of hospitals that can produce more outputs for the same inputs, or vice versa, then that hospital is considered fully efficient. Otherwise, it is not performing at the efficient frontier.

A visual representation is provided in Figure 4.1. Inputs are represented on the horizontal axis and outputs on the vertical axis. There are four hospitals, \( H_1, H_2, H_3 \) and \( H_4 \). The frontier line represents the set of input-output combinations of a hospital that is technically efficient.

The shape of the estimated frontier line depends on the assumptions used. The two main ones are constant returns to scale (CRS), represented by the straight line, and variable returns to scale (VRS), represented by the segmented line. In this example, hospital \( H_2 \)
Figure 4.1: Frontier concept of hospital efficiency in DEA
sets the efficient frontier under CRS, while all of \( H_1 \), \( H_2 \) and \( H_3 \) define that frontier under VRS. \( H_4 \) is inefficient under both assumptions because in each case there exists a combination of other hospitals that can (weakly) dominate it.

Let \( \hat{\theta}_h \) be the computed efficiency score of hospital \( h \). DEA has more than one definition of efficiency, so the efficiency score for \( H_4 \) depends on which method is used. The two simplest methods are input-oriented and output-oriented efficiency, illustrated using Figure 4.1 as follows:

\[
\begin{align*}
\text{Input-oriented-CRS} & = \frac{X_{4\text{CRS}}}{X_4} \\
\text{Input-oriented-VRS} & = \frac{X_{4\text{VRS}}}{X_4} \\
\text{Output-oriented-CRS} & = \frac{Y_4}{Y_{4\text{CRS}}} \\
\text{Output-oriented-VRS} & = \frac{Y_4}{Y_{4\text{VRS}}} 
\end{align*}
\]

We use input-oriented, VRS efficiency scores with BCC specification. Input-oriented means efficiency is measured by how much input is used to produce the observed set of outputs, relative to the minimum quantities of inputs implied by the estimated frontier. The BCC specification is one of a number of computation algorithms used to implement DEA. In practice, differences in DEA specification used seldom produce significant variations in the results (Golany, Yu 1997).

Additionally, the extra effort required for hospitals to improve service is captured by including the quality measure as an input variable. Specifically, the negative of the SIR rating is included because the less negative the SIR rating is, the better the hospital quality. Details of the analytical properties associated with similar approaches in quality-adjusted frontier analysis are available in Eckermann and Coelli (2013).

A weakness of DEA is that the generated efficient frontier is defined by the hospitals themselves. This means that the hospitals at the frontier are considered to be operating at 100% efficiency, which is unlikely in reality. This is important because hospital efficiency scores can be used for subsequent analysis. If a large number of data points are portrayed
to operate at perfect efficiency, then subsequent results may be biased. Some adjustment for this bias is necessary. But unlike OLS, DEA does not easily account for measurement error because it is a deterministic method.

This study will simulate the DEA distribution using naive bootstrap with subsamples (Simar, Wilson 2006). Recall that there are H hospitals in the entire sample. For each hospital, 2000 bootstrap resamples are generated, indexed n = 1,..., N. Each resample contains R hospitals drawn with replacement, where R is less than H; we set R at one-third of the sample size.\(^3\) We then compute the DEA efficiency score for that resample, written as \(\hat{\theta}_{hn}\), where n is the index of resamples. This results in a total of 2000 bootstrapped efficiency scores per hospital.

The aim of computing the bootstrapped efficiency scores is to estimate the bias of the original estimate for hospital h, defined as follows:

\[
\text{bias}(h) = \left( \frac{R}{H} \right) \frac{1}{2000} \sum_{n=1}^{2000} (\hat{\theta}_{nh} - \hat{\theta}_h)
\]

The first term is a sample size adjustment, the second term is the arithmetic mean of all the bootstrapped efficiency scores for each subsample and the third term is the original efficiency estimate. The bias-adjusted efficiency estimate \(\hat{\theta}_h\) is defined as follows:

\[
\hat{\theta}_h = (\hat{\theta}_h - \text{bias}(h))
\]

The second stage is concerned with the factors that are related to the technical efficiency of hospitals. This is estimated using a bootstrap-adjusted Tobit regression as specified by Simar and Wilson (2007).

The dependent variable is the vector of bias-adjusted efficiency scores \(\tilde{\theta}\). Let \(\eta = \frac{1}{\tilde{\theta}}\), a measure of (in)efficiency. The purpose is to reduce the truncation to one side and simplify.

\(^3\)Simar and Wilson (2006) stated that R should be less than H, but no explanation or guidelines for choosing R was given.
later analysis by mapping from (0,1] to [1, ∞). The vector of hospital-level independent variables affecting efficiency scores is represented by $z_h$, with corresponding parameters $\phi$ and error terms $\nu$. The regression equation is specified as follows,

$$\eta = \text{constant} + z_h^t \phi + \nu$$

where $\nu$ is an iid error term with variance $\sigma_\nu$.

To estimate confidence intervals that captures the distribution of $\phi$. Denote the predicted efficiency score as $\hat{\eta}$ and the residual as $\hat{\nu}$, where $\hat{\nu} = \hat{\eta} - \eta$. For notational simplicity, use $n$ again to index bootstraps of $\phi$, in a similar manner as the bias-adjustment for the efficiency scores $\theta$ earlier.

For each bootstrap iteration:

- Draw an error term $\tilde{\nu}_h$ for each hospital from a truncated normal distribution. The distribution has mean $= 0$, variance $= \sigma_\nu$ and left-limit $= 1 - z_h^t \phi$.
- Use this error term to compute a new predicted efficiency score $\tilde{\eta}_h$ for each hospital, defined as follows:
  $$\tilde{\eta}_h = z_h^t \phi + \tilde{\nu}_h$$
- This produces a set of bootstrapped efficiency scores $\tilde{\eta}_{hn}$, one for each hospital. With these new scores, estimate a truncated regression:
  $$\hat{\eta}_{hn} = \text{constant} + z_h^t \phi_n + \nu$$

The end result of these iterations is $N$ sets of parameter estimates, $\phi_1, ..., \phi_N$.

Percentile bootstrap confidence intervals are usually more accurate than Gaussian or other analytical approximations. However, the bootstrap distribution does not asymptotically converge to the sample distribution except in very specific cases. The most popular way to improve the accuracy of these intervals is the BCa percentile interval method pioneered by Efron and Tibshirani (1987). In simple terms, BCa makes two adjustments to the
simple percentile interval, “bias” and “acceleration”. These terms roughly translate to unaccounted differences in skewness and kurtosis between the bootstrap and the sample distribution. This is implemented in four steps: (i) compute the percentile CI; (ii) compute the bias constant; (iii) compute the acceleration constant; and (iv) apply the adjustment to the simple CI to form a BCa-adjusted CI. Each of these steps is outlined below.

(i) For computing the simple percentile interval, sort the estimates of each $\phi$ in ascending order from smallest to largest. The lower and upper limits of the simple CI are defined by the bootstrapped parameter value placed at its corresponding percentile rank. For example, if there are 9 values in a list, then the 25th percentile will be the 3rd smallest value in that list.

(ii) The bias constant of each parameter is computed as follows,

$$\text{Bias} = \Phi\left(\frac{\sum_{n=1}^{N} 1(\hat{\phi}_n < \phi)}{N}\right)$$

where $1(x)$ is the indicator function with $1(x) = 1$ if $x$ is true, otherwise $1(x) = 0$; and $\Phi$ is the CDF of the standard normal.

(iii) For estimating the acceleration constant, Efron and Tibshirani suggested using an old bootstrap method called the jackknife. Denote the jackknife estimate of each parameter as $\phi_{-h}$, defined as the parameter estimate using Tobit regression when the $h$-th observation is removed. This produces a total of $H$ jackknife estimates of each $\phi$. Additionally, denote $\overline{\phi}_{-h}$ as the mean of the $H$ jackknife estimates of $\phi$. The acceleration constant is defined as follows:

$$\text{Accel} = \frac{\sum_{h=1}^{H} (\overline{\phi}_{-h} - \phi_{-h})^3}{6(\sum_{h=1}^{H} (\overline{\phi}_{-h} - \phi_{-h})^2)^{\frac{3}{2}}}$$

(iv) For applying the adjustment based on the computed constants, denote the lower and upper percentiles of the straight percentile intervals as $l$ and $u$ respectively. Denote the

---

4Efron and Tibshirani’s paper explained one way to do this, but also mentioned that there are other ways of obtaining the acceleration constant.
BCa-adjusted locations as $\tilde{l}$ and $\tilde{u}$.\(^5\) The adjustment function is given as follows:

$$\tilde{l} = \phi(Bias + \frac{Bias + \phi^{-1}(l)}{1 - \text{Accel}(Bias + \phi^{-1}(l))})$$

$$\tilde{u} = \phi(Bias + \frac{Bias + \phi^{-1}(u)}{1 - \text{Accel}(Bias + \phi^{-1}(u))})$$

The values $\tilde{l}$ and $\tilde{u}$ are given in terms of the range of the standard normal distribution rather than percentile values. For example, if $\tilde{l} = -2.1$, then the lower bound is the bootstrapped value that is closest to the 1.79 percentile in size.

4.4 Data Description

Two sources of data are used in this study, a patient admissions database and a list of assembled hospital financial statements. Patient data is extracted from the Victorian Admitted Episodes Database (VAED) for the years 1999-2004. Patient names and identities of private hospitals are encoded due to confidentiality reasons. See chapter 2 for further details.

A useful feature of this data is the three-tiered hierarchy of disease definitions. The first two are directly listed from the VAED. The third is assembled from the Productivity Commission Reports. They are summarised in Table 4.1.

<table>
<thead>
<tr>
<th>Acute MDC 1-13, 16-18, 21-22</th>
<th>Maternity / Prenatal MDC 14-15</th>
<th>Alcohol, Mental, Drugs MDC 19-20</th>
<th>Rehabilitation / others MDC 23</th>
</tr>
</thead>
<tbody>
<tr>
<td>570 DRG’s(^6)</td>
<td>46 DRG’s</td>
<td>20 DRG’s</td>
<td>13 DRG’s</td>
</tr>
</tbody>
</table>

Administrative definitions for the DRG, MDC and Service-type aggregations. DRG and MDC definitions are directly taken from the VAED and the service types from the Productivity Commission (2010).

The smallest aggregation level is the Diagnostic-Related Group (DRG). There are around 660 DRGs in total, reduced to 570 after removing administrative entries.\(^7\) These DRGs are then aggregated into the next level called Major Diagnostic Categories (MDC). There are

\(^5\)Another common notation is $\alpha$ and $1-\alpha$, which implicitly assumes symmetry around the median.

\(^7\)A handful of rare illnesses at the DRG level did not have enough data points for sufficient variation to support Logistic regression. A simple average is taken for those cases.
23 MDCs in total during the studied period. Additionally, the Productivity Commission Reports (2010) aggregates the MDCs into four service types. They roughly translate to: 1) acute diseases, 2) pregnancy and births, 3) mental, drug and alcohol, 4) rehabilitation and long-term care. This was presumably done to reduce variance in risk-adjustment. This study will investigate if this way of defining aggregation makes a significant difference.

The other data source required for estimating hospital efficiency is operating expenses. Similar to chapter 3, these are extracted from the financial statements for public hospitals in Victoria during the years 2002-2004. The data are separated into metropolitan and rural hospitals in Table 4.2:

Table 4.2: Summary statistics of hospital finances and types of patients

<table>
<thead>
<tr>
<th></th>
<th>Pooled mean</th>
<th>Pooled sd</th>
<th>Metro mean</th>
<th>Metro sd</th>
<th>Rural mean</th>
<th>Rural sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capital ('000)</td>
<td>2,868</td>
<td>4,428</td>
<td>5,362</td>
<td>6,273</td>
<td>543</td>
<td>372</td>
</tr>
<tr>
<td>Labour ('000)</td>
<td>40,582</td>
<td>61,890</td>
<td>87,366</td>
<td>80,688</td>
<td>5,800</td>
<td>3,870</td>
</tr>
<tr>
<td>Materials ('000)</td>
<td>11,296</td>
<td>19,149</td>
<td>26,090</td>
<td>25,685</td>
<td>1,116</td>
<td>648</td>
</tr>
<tr>
<td>Acute</td>
<td>26,369</td>
<td>42,755</td>
<td>57,017</td>
<td>56,983</td>
<td>2,256</td>
<td>1,940</td>
</tr>
<tr>
<td>Maternity</td>
<td>3,753</td>
<td>9,341</td>
<td>8,075</td>
<td>14,618</td>
<td>245</td>
<td>271</td>
</tr>
<tr>
<td>Alcohol / drugs</td>
<td>2,820</td>
<td>5,407</td>
<td>6,484</td>
<td>7,199</td>
<td>129</td>
<td>111</td>
</tr>
<tr>
<td>Rehabilitation</td>
<td>5,700</td>
<td>9,836</td>
<td>12,728</td>
<td>13,679</td>
<td>654</td>
<td>677</td>
</tr>
<tr>
<td>Mort. Rate</td>
<td>3.52%</td>
<td>7.57%</td>
<td>4.60%</td>
<td>12.10%</td>
<td>2.22%</td>
<td>1.58%</td>
</tr>
<tr>
<td>Number of hospitals</td>
<td>298</td>
<td>100</td>
<td>198</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Profile of basic statistics for hospitals in the VAED sample for the years 2002-2004. The sample is treated as a cross-section. Since hospitals cannot be identified, they are treated as separate even though many of them are repeat appearances. Dollar values are not adjusted for inflation.

In the first stage of the analysis for estimating hospital efficiency, financial statement entries are aggregated into the three input prices of capital(K), labour(L) and materials(M). This leaves a total of five variables in the frontier estimate: three input types in dollars, output in number of patient-days and quality in estimated SIR values. Hospitals are then further separated into two hospital types, metropolitan and rural hospitals, and the estimate is repeated. This separation is relevant because Victoria has a clear contrast.

---

8The exact breakdown from MDC to these categories is available in P.182 of Productivity Commission Reports, Dec 2010
9Slight discrepancies exist due to investment income and donations.
between urban and rural regions, leading to very different operational needs and funding arrangements. Results generated from both the entire hospital set and from the separated metropolitan/rural sets will be presented.

Quality is one of the five variables used to estimate the frontier, measured by the SIR statistic of each hospital. The risk-adjustment portion of the SIR can be estimated using one of three tiers of disease aggregations. That is, for each aggregation tier, patients are divided into their corresponding aggregations. The expected mortality of patients from each aggregation are estimated using Logistic Regression. This produces three versions of expected mortality for each patient, each culminating into different SIR estimates for the hospitals. Other variables remain unchanged each time the frontier analysis is conducted.

In theory, there is enough information in financial statements and the VAED sample to use more variables and hospital types. However, we choose to limit the model to 3 input units and 1 output unit, separated into metropolitan and rural types, in order to strike a balance between bias and robustness. For DEA this is especially important because, unlike regression models, it has a deterministic structure that places even greater strain on degrees of freedom. The number of variables that can be supported is primarily a function of sample size. There is no definitive rule for this, so we use a rule of thumb by Cooper et al (2000) as the next best reference.¹¹

In the second stage, explanatory variables used include the quality measure, a computed Herschmann-Herfinhad index (HHI) and a dummy variable for teaching hospital status.¹² Quality is the primary variable and is included by default. HHI is a common measure of hospital competition and is usually considered to be an important factor for efficiency. This study computes an adjusted HHI according to Zwanziger and Melnick (1988).¹³ The basic idea is that a hospital’s catchment is only defined by regions where a significant percentage of patients living there attend that hospital. Teaching status is, according to Hureta et al (2008), a basic measure of ‘organizational and market characteristics’ that is widely considered to affect efficiency.

---

¹¹Cooper (2000) suggests the rule max(inputs*outputs, 3*(inputs+outputs))
¹²The HHI is the most common measure of market concentration, but it assumes that the sample sizes between markets are approximately equal. More refined versions of the HHI that account for this has been proposed (Hirschberg et al 2003).
¹³The exact adjustments made to the original HHI and its reasons are detailed in their paper.
4.5 Results

4.5.1 Estimates of hospital quality by level of disease aggregation

Table 4.3 contains summary statistics of hospital SIR values, separated into pooled data, metropolitan hospitals and rural hospitals. For each hospital grouping, the first column is the sample mean and the second column is the standard deviation. In general, the more refined the disease aggregation method, the less varied the estimated quality scores become. A consistent explanation is that a part of the perceived differences in quality can be attributed to unobserved patient differences. This effect is stronger for metropolitan hospitals, perhaps because they deal with more severe diseases.

<table>
<thead>
<tr>
<th>Aggregation type</th>
<th>Pooled mean</th>
<th>Pooled sd</th>
<th>Metro mean</th>
<th>Metro sd</th>
<th>Rural mean</th>
<th>Rural sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service (4)</td>
<td>1.5551</td>
<td>1.2026</td>
<td>1.5878</td>
<td>1.8421</td>
<td>1.5386</td>
<td>0.6886</td>
</tr>
<tr>
<td>MDC (23)</td>
<td>1.3879</td>
<td>0.8820</td>
<td>1.3769</td>
<td>1.2776</td>
<td>1.3934</td>
<td>0.5937</td>
</tr>
<tr>
<td>DRG (570~)</td>
<td>1.2675</td>
<td>0.6108</td>
<td>1.0876</td>
<td>0.7041</td>
<td>1.3584</td>
<td>0.5372</td>
</tr>
</tbody>
</table>

SIR estimates for hospitals in the dataset, presented in aggregate with mean and standard deviations. Results are separated into all (pooled), metropolitan hospitals and rural hospitals. SIR values are separately computed using three levels of disease aggregation during the risk-adjustment stage. See the data description section for details.

Tables 4.4 and 4.5 compare the sorting of smaller aggregations into broader ones using parameter clustering versus using administrative definitions. The first column lists the names of the broader aggregations defined administratively. The first row on top lists the same number of aggregations derived using clustering, labeled with generic numbers. The value in each cell represents the number of smaller aggregations that are present in a given clustered aggregation and administrative aggregation. The number of smaller aggregations present in a given broad aggregation is equal to the sum of its row or column.

For most administrative aggregations, their constituent DRGs are scattered evenly across the hypothetical clusters, with very little discernible pattern between the two aggregation methods. If they were similar, the majority of cells should be empty or contain small values, with much higher values in a small number of cells, but that is not the case here. This demonstrates that similarity in medical terms may not always be a good predictor of patient risk patterns.
Table 4.4: Cluster analysis comparing the MDC and DRG aggregation levels

<table>
<thead>
<tr>
<th>By VAED MDC</th>
<th>By clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Nervous system</td>
<td>0</td>
</tr>
<tr>
<td>Eye</td>
<td></td>
</tr>
<tr>
<td>Ear, nose, mouth and throat</td>
<td>1</td>
</tr>
<tr>
<td>Respiratory system</td>
<td>1</td>
</tr>
<tr>
<td>Circulatory system</td>
<td>2</td>
</tr>
<tr>
<td>Digestive system</td>
<td>4</td>
</tr>
<tr>
<td>Hepatobiliary system and pancreas</td>
<td>1</td>
</tr>
<tr>
<td>Musculoskeletal system and connective tissue</td>
<td>1</td>
</tr>
<tr>
<td>Skin, subcutaneous tissue and breast</td>
<td>3</td>
</tr>
<tr>
<td>Endocrine, nutritional and metabolic diseases and disorders</td>
<td>1</td>
</tr>
<tr>
<td>Kidney and urinary tract</td>
<td>1</td>
</tr>
<tr>
<td>Male reproductive system</td>
<td></td>
</tr>
<tr>
<td>Female reproductive system</td>
<td></td>
</tr>
<tr>
<td>Pregnancy, childbirth and the puerperium</td>
<td></td>
</tr>
<tr>
<td>Newborns and other neonates</td>
<td></td>
</tr>
<tr>
<td>Blood and blood forming organs and immunological disorders</td>
<td>1</td>
</tr>
<tr>
<td>Neoplastic disorders (haematological and solid neoplasms)</td>
<td>1</td>
</tr>
<tr>
<td>Infectious and parasitic diseases</td>
<td>1</td>
</tr>
<tr>
<td>Mental diseases and disorders</td>
<td>1</td>
</tr>
<tr>
<td>Alcohol or drug use and alcohol or drug induced organic mental disorders</td>
<td>1</td>
</tr>
<tr>
<td>Injuries, poisoning and toxic effects of drugs</td>
<td>1</td>
</tr>
<tr>
<td>Burns</td>
<td>1</td>
</tr>
<tr>
<td>Factors influencing health status and other contacts with health services</td>
<td>1</td>
</tr>
</tbody>
</table>

Comparison of the 23 administrative MDCs, versus 23 clusters sorted by similarity in risk-characteristics. The first column to the left lists the names of each MDC. The first row on top are the risk-characteristic clusters arbitrarily labeled from 1 to 23. Each row contains the number of DRGs in each MDC sorted into one of 23 hypothetical clusters. For example, in the ‘nervous system’ row, 6 of the DRGs in that MDC are assigned to cluster 6, out of a total of 32.
Table 4.5: Cluster analysis comparing the service category and DRG aggregation levels

<table>
<thead>
<tr>
<th>Service Type</th>
<th>By clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Acute</td>
<td>25</td>
</tr>
<tr>
<td>Pregnancy and neonate</td>
<td>17</td>
</tr>
<tr>
<td>Mental and alcohol</td>
<td>6</td>
</tr>
<tr>
<td>Others</td>
<td></td>
</tr>
</tbody>
</table>

Similar to Table 4.4, but between DRG and service categories. As an example, the number 25 in the upper-left-hand corner means 25 DRGs belonging to the ‘Acute’ service type was assigned to cluster 1.

4.5.2 Effect of disease aggregation on empirical results

The empirical setting is a two-stage DEA/Tobit analysis of efficiency. Summary statistics of first-stage results are included in Table 4.6. Four specifications are estimated: service-level risk-adjustment, MDC-level, DRG-level and a control set without quality. There are two main observations. The first one is a confirmation that quality is important. This can be seen by the large increase in efficiency score after quality is included. This is consistent with the notion that hospitals with sicker patients may appear to be less efficient. The second is that a change in disease aggregation did not make a large difference to hospital efficiency levels. The mean efficiency scores were similar across the aggregation methods and the correlation coefficients between the different sets of efficiency scores are close to 1. A larger dataset of hospitals could take this analysis further.

Table 4.6: DEA estimates of hospital efficiency, by aggregation levels

<table>
<thead>
<tr>
<th>Hospital Type</th>
<th>Pooled mean</th>
<th>Pooled sd</th>
<th>Metro mean</th>
<th>Metro sd</th>
<th>Rural mean</th>
<th>Rural sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>(no mortality rates)</td>
<td>0.3936</td>
<td>0.2109</td>
<td>0.5609</td>
<td>0.2478</td>
<td>0.5568</td>
<td>0.1811</td>
</tr>
<tr>
<td>Service - SIR</td>
<td>0.5782</td>
<td>0.1839</td>
<td>0.7723</td>
<td>0.1726</td>
<td>0.6317</td>
<td>0.1575</td>
</tr>
<tr>
<td>MDC - SIR</td>
<td>0.5634</td>
<td>0.1878</td>
<td>0.7725</td>
<td>0.1674</td>
<td>0.6237</td>
<td>0.1582</td>
</tr>
<tr>
<td>DRG - SIR</td>
<td>0.5647</td>
<td>0.1856</td>
<td>0.7685</td>
<td>0.1565</td>
<td>0.6335</td>
<td>0.1529</td>
</tr>
<tr>
<td>Corr(DRG, MDC)</td>
<td>0.9615</td>
<td>0.9602</td>
<td>0.9688</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr(DRG, Service)</td>
<td>0.9404</td>
<td>0.9445</td>
<td>0.9494</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr(MDC, Service)</td>
<td>0.9876</td>
<td>0.9879</td>
<td>0.9797</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimates of hospital efficiency by aggregation level using pooled, metropolitan and rural hospitals. Mean and standard deviations of the scores are presented, as well as the correlation coefficients between the aggregation levels.
The second stage examines the relationship between quality and efficiency, with hospital-level factors added as control variables. The result of interest from the Tobit regression is whether the 95% CIs of the quality parameters overlap or not. It is a measure of whether disease aggregation has a statistically significant impact on the quality-efficiency relationship. Comparison results are listed in Table 4.7, with complete tables of parameter estimates and their corresponding confidence intervals listed in Appendix C.

Table 4.7: Comparing the 95% CI of quality parameters in the second-stage Tobit regression

<table>
<thead>
<tr>
<th>Different using 95% CI</th>
<th>Services vs DRG</th>
<th>Services vs MDC</th>
<th>MDC vs DRG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>Disjoint</td>
<td>Overlap</td>
<td>Disjoint</td>
</tr>
<tr>
<td>Metro</td>
<td>Overlap</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
<tr>
<td>Rural</td>
<td>Overlap</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
</tbody>
</table>

A summary of whether the 95% CIs of the quality parameters, estimated during the second stage of the DEA/Tobit estimation, overlap or not. ‘Disjoint’ means the confidence intervals do not overlap. It is taken as a supporting sign that disease aggregation matters. Similarly, ‘Overlap’ means there is a range of values where the CIs overlap. Overlapping CI means there is insufficient evidence proving that disease aggregation makes a statistically significant impact. Tables containing all the parameter estimates and their confidence intervals are listed in Appendix C.

For the pooled data, estimates obtained at the DRG level are significantly different from those under the MDC and service-type levels. This is consistent with our claim that the choice of disease aggregation mechanism can have a significant effect on subsequent empirical analysis. As such, risk-adjustment must be repeated using different possible aggregations to ensure that the results are not dependent on this empirical detail. Interestingly, a similar effect is not observed when hospitals are divided.

4.5.3 Robustness check: risk-adjustment by fixed effects

We conduct two simple robustness checks. The first concerns the risk-adjustment process. Computed quality statistics can be sensitive to the method of risk-adjustment and logistic regression is not the only method available. There is also a more classical method of risk-adjustment that uses hospital dummy variables as a measure of quality. It should be noted that dummy variables capture all types of heterogeneity, not just differences in patient makeup. It is still prevalent in the literature, particularly when there are concerns about observed heterogeneity that correlates with error terms. See Chua et al (2010) for a recent example of a more elaborate two-stage fixed-effects model in a similar context.
For each element in the vector of risk-factors $\delta$ used for risk-adjustment, introduce a new set of variables $\delta_{au}$, where $a = 1$ to $A$; $u = 1, .. U$. The value of each element in $\delta_{au}$ is positive if the patient belongs to aggregation $a$ and zero otherwise. Additionally, introduce a new set of hospital-level dummies $\omega_1$ to $\omega_H$. The first stage risk-adjustment estimates the following OLS specification:

$$\text{Death (0 or 1)}_{hp} = \text{constant} + \sum_{a=1}^{A} \sum_{u=1}^{U} \alpha_{au} \delta_{au} + \sum_{h=1}^{H} \alpha_h \omega_h + \epsilon_{hp}$$

The aim of estimating this regression is to obtain estimates for the hospital-level effects $\alpha_h$. They represent the marginal change in probability of death by virtue of the hospital’s identity and acts as an alternative quality measure for the Tobit regression. The role of disease aggregation is that the finer the aggregation method is, the more patient risk controls $\alpha_{au}$ there are. The intention is to better capture patient heterogeneity between hospitals, ensuring that hospital dummies actually capture hospital-level differences.

Table 4.8: Robustness check: using hospital fixed-effects as a quality measure

<table>
<thead>
<tr>
<th>Hospital type</th>
<th>Pooled</th>
<th>Metro</th>
<th>Rural</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>sd</td>
<td>mean</td>
</tr>
<tr>
<td>Group - dummy</td>
<td>0.0577</td>
<td>0.0720</td>
<td>0.0694</td>
</tr>
<tr>
<td>MDC - dummy</td>
<td>0.0107</td>
<td>0.0669</td>
<td>0.0231</td>
</tr>
<tr>
<td>DRG - dummy</td>
<td>0.0150</td>
<td>0.0594</td>
<td>0.0248</td>
</tr>
</tbody>
</table>

Estimates of hospital efficiency using hospital fixed-effect parameter. Results are presented in the same way as Table 4.3.

Mean and standard errors of hospital quality estimates as defined by fixed effects parameters $\alpha_h$ is listed in Table 4.8. Similar to risk-adjustment using Logistic regression, the standard deviation of the estimates of $\alpha_h$ between hospitals become lower as disease aggregation becomes more refined. The mean of hospital quality estimates also become closer to zero with smaller disease aggregation. These two observations together strengthen the argument that, for most existing studies, hospital-level differences may have been exaggerated due to unobserved patient heterogeneity.
4.5.4 Robustness check: using differences as quality statistic

The second robustness check concerns the risk-adjustment statistic. Besides the SIR, there are other risk-adjusted quality measures used to evaluate hospital performance. To argue that disease aggregation is important in general, there needs to be indications of its influence when using some of these other measures. One of these measures is to take the difference between expected and observed number of incidents:

\[ \text{Diff}_h = \frac{D_h - E(D_h)}{P_h} \in [-E(D_h), \infty) \]

Table 4.9: Robustness check: estimates of hospital efficiency using alternative quality measure

<table>
<thead>
<tr>
<th>Hospital type</th>
<th>Pooled mean</th>
<th>Metro mean</th>
<th>Rural mean</th>
<th>Pooled sd</th>
<th>Metro sd</th>
<th>Rural sd</th>
</tr>
</thead>
<tbody>
<tr>
<td>(no mortality rates)</td>
<td>0.3936</td>
<td>0.5609</td>
<td>0.5568</td>
<td>0.2109</td>
<td>0.2478</td>
<td>0.1811</td>
</tr>
<tr>
<td>Service - Diff</td>
<td>0.4231</td>
<td>0.6355</td>
<td>0.5651</td>
<td>0.2191</td>
<td>0.2053</td>
<td>0.1775</td>
</tr>
<tr>
<td>MDC - Diff</td>
<td>0.4107</td>
<td>0.6151</td>
<td>0.5627</td>
<td>0.2197</td>
<td>0.2161</td>
<td>0.1775</td>
</tr>
<tr>
<td>DRG - Diff</td>
<td>0.3978</td>
<td>0.5804</td>
<td>0.5628</td>
<td>0.2175</td>
<td>0.2432</td>
<td>0.1770</td>
</tr>
<tr>
<td>Corr(DRG, MDC)</td>
<td>0.9536</td>
<td>0.9354</td>
<td>0.9977</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr(DRG, Service)</td>
<td>0.9364</td>
<td>0.9132</td>
<td>0.9974</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corr(MDC, Service)</td>
<td>0.9919</td>
<td>0.9879</td>
<td>0.9993</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

First-stage results estimated using alternative quality measure. Similar to Table 4.6, the mean, standard errors and correlation coefficients are listed.

First-stage results using the alternative measure are listed in Table 4.9 and second-stage results in Table 4.10, with parameter estimates in Appendix C. For the first stage results estimated using the alternative quality measure, the level of disease aggregation is still not a significant influence for efficiency scores. This is consistent with results estimated using SIR.

As for the second stage results, the relationship between HHI and rural hospital inefficiency remains positive and significant when the alternative quality measure is used. Quality no longer has a statistically significant effect on efficiency, but their point estimates are still significant different from each other at the service and MDC levels. This does not contradict the message that disease aggregation matters. However, there are not enough
Table 4.10: Robustness check: comparing CIs of the quality parameters estimated using alternative quality measure

<table>
<thead>
<tr>
<th>Different using 95% CI, alternative</th>
<th>Services vs DRG</th>
<th>Services vs MDC</th>
<th>MDC vs DRG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pooled</td>
<td>Overlap</td>
<td>Disjoint</td>
<td>Overlap</td>
</tr>
<tr>
<td>Metro</td>
<td>Overlap</td>
<td>Disjoint</td>
<td>Overlap</td>
</tr>
<tr>
<td>Rural</td>
<td>Overlap</td>
<td>Overlap</td>
<td>Overlap</td>
</tr>
</tbody>
</table>

The second stage analysis using the alternative quality measure. Similar to Table 4.7, ‘Disjoint’ means their 95% CIs do not overlap and ‘Overlap’ means they do.

results of statistical significance to further investigate the exact details of their functional forms driving these differences.

4.6 Discussion

Most existing literature aggregates patients into disease groups for the purposes of risk-adjustment. Aggregation is necessary because there are too many patients to individually account for. However, there is no agreed method for doing this, which led to a diversity of methods used without much empirical justification. Existing literature does not discuss whether the choice of aggregation method has any impact on their subsequent results.

We first estimated hospital risk-adjusted quality measures using three different levels of disease aggregation. The levels were: 4 service categories, 23 Major Diagnostic Categories, and around 600 Diagnostic-Related Groups. The marginal factors for probability of death are estimated separately in each aggregation, using logistic regression. Then, these quality measures were used in a two-stage DEA/Tobit analysis, first to estimate quality-adjusted efficiency, then to investigate factors that explain these differences in efficiency. Every stage of the exercise is repeated for all levels of disease aggregation so that their results can be compared.

4.6.1 Summary of findings

The variance in reported quality between hospitals decreased as disease aggregations became smaller and more numerous. This is consistent with the common belief that unobserved patient heterogeneity can exaggerate differences in hospital performance. How-
ever, the standard error of the estimates also increased, so there is a bias-variance trade-off.

During the second-stage of the empirical analysis, the estimated correlation between hospital quality and efficiency were also significantly different at the 95% level. Larger distances also increased inefficiency for rural hospitals. These observations are robust to other risk-adjustment methods and as well as an alternative quality statistic to the SIR.

These results demonstrate that risk-adjustment should be done using more than one disease aggregation level whenever possible. It is a necessary safety check because there is a chance that the reported estimates of previous studies may have been significantly different if an alternative aggregation was used instead.

Additionally, a comparison of disease aggregations derived by medical standards versus risk-characteristics also resulted in different higher-level aggregations. Medical or administrative definitions are often used to divide patients for risk-adjustment purposes, but in fact they may not share similar risk characteristics.

To explain this in an empirical setting, we return to the public and private hospital study conducted by the Productivity Commission (2010; pp 205). It is a comprehensive study that addresses multiple policy issues, one of which was to compare efficiency scores between private and public hospitals after controlling for compounding factors. Their method of analysis included two types of outputs, weighted-separations and hospital quality as measured using the hospital-standardised mortality ratio (HSMR).

The problem is that HSMR scores are only risk-adjusted at the Major Diagnostic Categories (MDC) level, without checking for other possible disease-aggregation methods. While we do not have access to the private hospital data needed to repeat the estimation, the comparison conducted in our study suggests that some of their conclusions may also be sensitive to changes in disease-aggregation in the risk-adjustment process.

4.6.2 Limitations and extensions

We note that the design of our empirical demonstration is constrained by three limitations. The first one is the hospital costing data. Hospital prices are derived from financial
statements, which only provide total dollar amounts for their inputs. Unit quantities are unavailable, so prices had to be imputed. This limits the dimensionality of the DEA analysis, which reduces the accuracy of the DEA efficiency scores. Also, the accounting definition of depreciation can also be quite different from the economic concept of capital expenses, which may have reduced the accuracy of estimated capital prices.

The second one is that DEA, unlike linear regressions, is not invariant to affine transformations. That is, if the each value of a variable is transformed by a function of the form $Ax + b$, the resulting efficiency estimates from DEA will also be different. This, rather than disease aggregation, could have also contributed towards variations in efficiency scores.

The third one is the characterisation of a general-purpose public hospital as a single decision-making unit (DMU). This is likely inaccurate because the technology function of each service type is different. A possible variation to this study is to focus on a specific ward. The challenge is obtaining data on inputs, particularly the appropriation of overheads. On the flip side, a DMU smaller in scope may allow physical input units rather than dollar amounts, or a combination thereof.

Finally, having demonstrated its potential effect on empirical results, we re-iterate the need for further investigation into the nuances of disease aggregation in future studies. The eventual goal is to derive a list of criteria that researchers can use to choose between possible aggregation methods. As a starting point, one could depend on statistical concepts such as minimising the mean-squared error of the risk-adjustment process, or estimate a loss function based on the distribution of each aggregation method and its implied impact on the expected value of a policy.
Notation used for chapter 4

The following is a list of notation used in this chapter. The most basic symbols used in Economics or Statistics may be excluded.

Hospitals, patients, SIR:
- $h$: Index of hospitals from 1 to $H$.
- $p$: Index of patients from 1 to $P$, in a given hospital.
- $\delta$: Vector of patient characteristics.
- $\lambda$: Vector of parameters for patient characteristics.
- $D_{hp}$: Death status of patient $p$ in hospital $h$, 0 or 1.
- $D_h$: Number of deaths in hospital $h$.
- $ED_{hp}$: Expected probability of death of patient $p$ in hospital $h$.
- $ED_h$: Expected number of deaths in hospital $h$.
- SIR: Standardised Incidence Ratio.

Disease aggregation comparison:
- $a$: Index of disease aggregations from 1 to $A$.
- $\epsilon$: Variance-covariance matrix of $\lambda$.
- $u$: Index of risk-adjustment variables from 1 to $U$.
- $c_{12}$: “Distance” between disease aggregations 1 and 2.

First-stage DEA and bootstrap:
- $H_1, H_2, H_3, H_4$: Special symbol, DEA diagram.
- $\hat{\theta}_h$: DEA efficiency score of hospital $h$.
- $n$: Index of bootstraps from 1 to $N$.
- $R$: Subsample of hospitals from the total list of $H$ hospitals.
- bias($h$): Estimated bias of computed efficiency score of hospital $h$. 

104
• $\tilde{\theta}_h$: bias-adjusted DEA efficiency score of hospital $h$.

Second-stage efficiency estimates, bootstrap DEA:

• $\eta$: Reciprocal of efficiency score.
• $z_h$: Vector of independent variables.
• $\phi$: Vector of parameters of independent variables.
• $\nu$: Residual of efficiency score.
• $\upsilon$: Error term of the distribution of efficiency score $\eta$.

Second-stage efficiency estimates, BCa confidence interval:

• $\mathbb{1}$: Indicator function.
• $\Phi$: CDF of standard normal.
• $\phi_{-h}$: jackknife estimate of $\phi$ with $h$ removed.
• $\tilde{l}, \tilde{u}$: BCa-adjusted lower and upper values of inverse standard normal CDF.

Hospital dummies risk-adjustment robustness:

• $\omega$: Hospital dummies.
• $\alpha$: parameter estimates of OLS risk-adjustment.
Chapter 5

Conclusion

This thesis is divided into three parts, each discussing a topic related to the measurement and evaluation of quality in public hospitals. The first part addresses the variability of risk-adjustment in quality measures, commonly ignored or approximated in simple ways in the literature. A number of analytical approaches to capture this dispersion is suggested. The second part investigates the lack of consensus about the relationship between hospital cost and quality. An important reason for this is the discrepancy in method and data used. The third part examines disease aggregation, an empirical detail during risk-adjustment that is not well-discussed in the literature. The concern is that the choice of aggregation may impact on estimated quality measures and subsequent empirical results.

Part one is about the statistical uncertainty of commonly-used quality measures, usually captured using confidence intervals (CI). Estimates of these CIs are often misrepresented because most existing methods do not adequately account for the variability embedded in the risk-adjustment stage. Practical and accurate methods to do so are important because these statistics are often used in subsequent analysis, such as computed regressors, in instrumenting for an endogenous variable, or in hypothesis testing. Bootstrapping or other non-parametric methods are accurate but can be difficult to use depending on the problem, despite continuing developments towards more efficient computational methods and more intuitive strategies for interpreting results. Alternative approaches need to be made available at the policymaker’s disposal.

We address this problem by borrowing a number of simple-to-use analytical approxima-
tions from the applied econometrics literature and adapting them for use in a hospital setting. Point estimates of each hospital’s quality is computed, along with CIs estimated using each of these methods. Results are then compared against each other as well as against simulated benchmarks. Two standards are used to judge the performance of these methods: coverage statistics and the similarity of their lower/upper bound values as compared to the simulated ones.

Results showed that true dispersion is under-estimated when using methods that ignore the variance during the risk-adjustment processes. Analytical methods also worked better for larger hospitals because there is a lower risk of solution failure. Additionally, accurately capturing the higher moments of the distribution seems to be the key, even if the approximation did not follow all theoretical conditions such as asymptotic convergence. This is indicated by the marginally better performance of the two analytical methods that adjusted for the skewness and kurtosis of the SIR distribution.

This study contributes to the literature in two ways. First, we provide a systematic comparison of frequently used analytical methods for capturing the variability of risk-adjusted quality measures. Methods that performed better in this comparison exercise also provide ideas for future research into more accurate and practical ways to use quality measures. Second, our comparison exercise is based on actual patient data, with parameters naturally resulting from differences in variable definition. The disadvantage of using simulated data is the need to justify the realism of the chosen parameters.

Part two is about the relationship between hospital efficiency and quality, an established research question without a consensus answer. Some of the variation in results comes from differences in data used, an inevitable consequence of an empirically-driven research topic. However, we also believe that differences in estimation strategy and measure of hospital quality are two more important reasons for this phenomenon. Unfortunately, most studies use their own proprietary data and method of analysis, making it hard to isolate the two effects. This makes it difficult to directly test our hypothesis.

The second-best approach is to provide evidence that a change in either of these two factors alone can significantly change the relationship between quality and cost. To show this, two methodological variations are introduced into the estimated hospital cost function. One is to vary the estimation strategy by introducing a bootstrap adjustment for the stochastic
properties of the quality measure. The other is to compare mortality versus unplanned readmissions, two common variables used to measure quality. The variable of interest is the marginal effect of quality on cost.

The first comparison was between the estimated effect of quality on cost with and without bootstrap adjustment for the distribution of the quality measure. The correlation estimated from the bootstrapped-version was negative and more statistically significant. The second comparison was between the use of relative mortality versus relative readmissions as the quality measure. When the default quality measure of mortality was replaced by unplanned readmissions, the relationship between hospital quality and cost are weaker. Additionally, a hypothetical static comparison exercise also generated a sizable implied difference in estimated cost for improving hospital quality for both methodological variations.

This study contributes to the literature by demonstrating the necessity of performing robustness checks on two methodological variations that may affect the answer to this research question. In the first instance, ignoring the error structure of quality measures may introduce bias in subsequent empirical results, which can then lead to misleading policy recommendations. In the second instance, if a different measure of underlying hospital quality could have produced a different result, then it is necessary to check by repeating the analysis with alternative quality measures. They can range from traditional measures based on observed patient outcome by a medical professional, or patient reported outcomes that captures the intangible side of health services as well, such as the pain threshold for alternative treatment options. In both cases, the aim is to further progress in answering this research question by controlling for differences in method of analysis.

Part three is about the process of disease aggregation, a necessary step towards patient risk-adjustment and estimating quality measures. Since there are too many disease types in a hospital to address individually, they are grouped into categories before analysis. This results in small losses of information but is also necessary for statistical risk-adjustment to happen. The problem is that while most studies are aware of the importance of risk-adjust method, there seems to be lack of concern for the way disease aggregation is handled. Our goal is to show that the choice of disease-aggregation affects both estimates of hospital quality and subsequent empirical results.
To achieve these goals, this study is organised into two parts. First, the risk-adjusted quality measure of each hospital is separately estimated using three graduated levels of disease aggregations. The resulting estimates are compared against each other for any similarities and differences. Second, these quality values are incorporated into an illustrative study measuring the relationship between hospital efficiency and quality, using a two-stage DEA/Tobit specification. The emphasis is on the comparison of parameter estimates between disease aggregation levels rather than the estimates themselves.

In the first part, we find that the observed heterogeneity in quality amongst hospitals is reduced when disease aggregation became more refined. This is consistent with the commonly cited concern that some of the perceived differences in hospital quality can be attributed to unobserved patient heterogeneity. In the second part, the estimated correlation between quality and efficiency in the two-stage estimate turned out to be different from each other at the 95% significance level. Both of these results are robust to variations in the function used for measuring quality and alternative risk-adjustment methods.

To our best knowledge, this is the first study that explicitly isolates the effect of disease aggregation on risk-adjusted quality measures. As our results demonstrate, the choice of disease aggregation may influence estimates of hospital quality, as well as other analysis results that depend on those measures. Therefore, risk-adjusted quality measures should, whenever possible, be computed using multiple disease aggregation methods to ensure that their findings are resilient to slight variations in estimation method. The next logical step is to derive rules-of-thumb that future researchers can depend on to determine the optimal disease aggregation method to use.

As a finishing note, it is acknowledged here that besides observable medical variables such as patient incidents, stakeholder-reported measures are also an important component of hospital performance. The key is to identify ones that are both feasible to collect and hold the most promise for new discoveries. This is a challenge for public patient records such as the VAED because its official nature places constraints on the variables that they are able to collect. In particular, measures that depend on reported ratings have to be widely supported as fair and impartial since they are subjective by construction.
Bibliography


Appendix A

Appendix to Chapter 2

A.1 Detailed derivation of analytical CI estimation methods

D1: Wilson’s score interval (Wilson 1927).

Let $z_{\frac{\alpha}{2}}$ and $z_{1-\frac{\alpha}{2}}$ be the standard deviations of the 2.5th and 97.5th percentiles of the normal distribution respectively. Wilson’s Score Interval for the binomial distribution is as follows:

\[
\text{CI}(\text{lb}, \text{ub}) = \mu_X + \frac{z_{1-\frac{\alpha}{2}}^2}{2P} \pm \frac{z_{1-\frac{\alpha}{2}}^2}{4P^2} \sqrt{\frac{\mu_X(1-\mu_X)}{P} + \frac{z_{1-\frac{\alpha}{2}}^2}{4P^2}} \cdot \frac{P}{\mu_Y} \]

D2: Byar’s Method (Rothman, Boice 1979).

The original Byar’s method was derived when there was not sufficient computing power for large values of Poisson. The following is a version developed by Sahai and Khurshid (1993) that dropped the simplification as it was no longer necessary. Let $\chi^2$ represent the Chi-squared distribution, with degrees of freedom given in the lower-right subscript. The equation is given below,
\[ \text{CI}(\text{lb,ub}) = \left[ \frac{\chi^2(2\mu_X, \frac{q}{2})}{2\mu_Y}, \frac{\chi^2(2\mu_X + 1, 1 - \frac{q}{2})}{2\mu_Y} \right] \]

D3: Bootstrap with replacement assuming fixed expected incidence.

Each patient in a given hospital contains two attributes, the death status \( \mu_{Xp} \) (0 or 1) and the probability of death \( \mu_{Yp} \). One way to estimate the CI of the SIR is to bootstrap a large number of resampled hospitals and compute the SIR for each one. Collectively, the SIR values form an Empirical Distribution Function (EDF) that approximates the SIR distribution.

2100 random samples indexed \( n = 1, ..., N \) are drawn to create bootstrapped resamples with replacement of equal size to the actual hospital. For each resample, compute the SIR and denote it \( SIR_N \). The \( (1 - \alpha) \)-th CI estimate is then defined as the \( \alpha \times N \)-th and \( (1 - \alpha) \times N \)-th values of the list of SIRs. The 95% CI is defined by the 53rd and 2047th values of the random samples.

N1: Fieller’s Interval (Fieller 1932).

Let \( t_{P-1} \) be the value of the inverse Student’s t distribution with \( P-1 \) degrees of freedom. Denote a new variable, \( q \), indicating the value for half the length of the inverse Student’s t distribution, defined as follows:

\[ q = \frac{1}{2} \cdot t_{(1 - \frac{\alpha}{2}, P-1)} \]

This study will use a geometric interpretation of Fieller’s interval:

\[ \text{CI}(\text{lb,ub}) = \left( \frac{\mu_X \mu_Y - q^2 \sigma_{XY}}{\mu_Y^2 - q^2 \sigma_Y^2} \right) \pm \sqrt{\left( \frac{\mu_X \mu_Y - q^2 \sigma_{XY}}{\mu_Y^2 - q^2 \sigma_Y^2} \right) - \left( \mu_X^2 - q^2 \sigma_X^2 \right) \left( \mu_Y^2 - q^2 \sigma_Y^2 \right)} \]
To check for regularity conditions, the following constants are also computed:

\[ q_{\text{exclusive}}^2 = \frac{\mu_Y^2}{\sigma_Y^2} \]
\[ q_{\text{complete}}^2 = \frac{\mu_X^2 \sigma_Y - 2 \mu_Y \mu_X \sigma_{XY} \mu_Y^2 \sigma_X}{\sigma_Y^2 \sigma_X - \sigma_{XY}^2} \]

The constants are then used to check if a continuous CI exists:

\[
CI = \begin{cases} 
[0, \infty], & \text{if } q_{\text{complete}}^2 \leq q^2 \\
[0, l_1] \cup [l_2, \infty], & \text{if } q_{\text{exclusive}}^2 \leq q^2 \leq q_{\text{complete}}^2 \\
(CI(\text{lb}), CI(\text{ub})) & \text{otherwise}
\end{cases}
\]

N2: Gaussian Ratio (Hinkley 1969).

Let \( G(\text{SIR}) \) be the CDF of the Gaussian Ratio distribution, written as follows:

\[
G(\text{SIR}) = \Gamma(g_1, g_2; g_3) + \Gamma(-g_1, -g_2; g_3)
\]

where

\[
\Gamma(g_1, g_2; g_3) = \frac{1}{2\pi \sqrt{(1 - g_3^2)}} \int_{g_1}^{\infty} \int_{g_2}^{\infty} \exp \left(-\frac{\mu_X^2 - 2g_3 \mu_X \mu_Y + \mu_Y^2}{2(1 - g_3^2)}\right) dX dY
\]

\[
g_1 = \frac{\mu_X - \mu_Y \text{SIR}}{\sigma_X \sigma_Y h(\text{SIR})}
\]
\[
g_2 = -\frac{\mu_Y}{\sigma_Y}
\]
\[
g_3 = \frac{\sigma_Y \text{SIR} - \rho \sigma_X}{\sigma_X \sigma_Y h(\text{SIR})}
\]
\[
h(\text{SIR}) = \left(\frac{\text{SIR}^2}{\sigma_X^2} - \frac{2 \rho \text{SIR}}{\sigma_X \sigma_Y} + \frac{1}{\sigma_Y^2}\right)^{\frac{1}{2}}
\]
For numerical purposes, the single-variable form of the integral is used:

\[
\Gamma(g_1, g_2; g_3) = \frac{1}{2\pi} \int_0^{g_3} \frac{1}{\sqrt{1 - (g_3)^2}} \exp\left(-\frac{g_1^2 - 2g_3g_1g_2 + g_2^2}{2(1 - (g_3)^2)}\right) dg_3
\]

The lower bound of the CI is the SIR value such that \(G(\text{SIR}) = \frac{\alpha}{2}\). Similarly, the upper bound is the SIR value such that \(G(\text{SIR}) = 1 - \frac{\alpha}{2}\).

S1: Chi-square ratio approximation (Silcock 1994)

Let \(\mu_X\) and \(\mu_Y\) be normally distributed. This leads to the following approximation:

\[
\frac{(\mu_X - \mu_Y.SIR)^2}{\text{Var}(\mu_X - \mu_Y.SIR)} \approx \chi^2_{\alpha,1}
\]

Take the variance term from the above expression and expend it:

\[
\text{Var}(\mu_X - \mu_Y.SIR) = \text{Var}(\mu_X) - 2SIR\sigma_{XY} + SIR^2\text{Var}(\mu_Y)
\]

Silcock (1994) then used features of the Poisson approximation for simplifying out further. This can be adapted to a binomial approximation in this study where the individual quantities are estimated separately:

\[
SIR^2\mu_Y^2 - \sigma_Y^2\chi^{-1}_{\alpha,1} + SIR(2\sigma_{XY}\chi^{-1}_{\alpha,1} - 2\mu_X\mu_Y) + \mu_X^2 - \sigma_X^2\chi^{-1}_{\alpha,1} = 0
\]

CI(lb,ub) are the two solutions to the quadratic equation above. The solutions might not exist.

S2: Log-adjusted confidence intervals against skewness (Sherman et al 2011).

The key innovation is using an expansion to avoid using a simplistic Delta Method ap-
proximation, but only for logs:

$$\text{CI}(\text{lb}, \text{ub}) = SIR \cdot \exp \left( \pm t_{\frac{\alpha}{2}, P-1} \sqrt{\frac{\sigma_Y^2}{\mu_Y^2 P} + \frac{\sigma_X^2}{\mu_X^2 P} - 2 \frac{\sigma_{XY}}{P \mu_Y}} \right)$$

The algebra within the square-root signs is similar to that of Fieller’s Interval.

S3: Cube-root adjusted confidence intervals (Hall 1992).

The natural log used in S2 can be viewed as root of the constant e. With some Delta-Method intermediary steps, other roots such as the cube-root can be used. First, apply the delta method approximation to estimate \( \text{Var}(SIR) \):

$$\text{Var}(SIR) \approx \frac{\mu_X^2}{\mu_Y^2} \sigma_Y^2 + \frac{1}{\mu_Y^2} \sigma_X^2 - 2 \frac{\mu_X}{\mu_Y} \sigma_{XY}$$

This is then used to estimate the variance of the cube-root of the above:

$$\text{Var} \left( SIR^{\frac{1}{3}} \right) \approx \frac{V(SIR)}{9SIR^{3/4}}$$

Finally the confidence interval is estimated as follows (\(z\) is the corresponding value of the inverse CDF of the standard normal of the given percentile):

$$\text{CI}(\text{lb}, \text{ub}) = \left( SIR \pm z_{1-\frac{\alpha}{2}} \sqrt{\text{Var}(SIR)^{\frac{1}{3}}} \right)^3$$

S4: Edgeworth adjustment for skewness and kurtosis (Ng et al 2008).

The Edgeworth Series is supposed to adjust for the skewness and kurtosis of the entire SIR. This study will adjust for the denominator. Higher moments for the entire SIR are not solved here. To begin, Poisson-Binomial distributions with large N can be roughly
approximated by a Poisson distribution with mean equal the sum of expectations of all the patients in the hospital. This allows the use of the Poisson and Negative Binomial compound distribution relationship. Re-parameterise the first two moments of $\mu_Y$

$$\begin{align*}
\mu_Y & = \eta \theta \\
\sigma_Y^2 & = \eta \theta^2
\end{align*}$$

The compound distribution is negative binomial:

$$\text{NegBin} \left( \eta, \frac{1}{\theta + 1} \right) \sim \text{NegBin}(\hat{r}, \hat{p})$$

The approximate skewness and kurtosis then follows:

$$\begin{align*}
S(\mu_Y) & \approx \frac{(1 + \hat{p})}{\sqrt{\hat{r} \hat{p}}} \frac{1}{\sqrt{\hat{P}}} \\
K(\mu_Y) & \approx \left( \frac{6}{\hat{r}} + \frac{(1 + \hat{p})^2}{\sqrt{\hat{r} \hat{p}}} \right) \frac{1}{\hat{P}^2}
\end{align*}$$

Dividing by $\sqrt{\hat{P}}$ and $\hat{P}^2$ respectively is derived by the relationship for sum of identical distributions. Patients in hospitals are not identical but the adjustment is asymptotically valid because of Central-Limit Theorem. The skewness and kurtosis estimates are then substituted back into the Edgeworth Series to adjust the percentiles for the 95% CI’s. This study will use the first two expansions:

$$\begin{align*}
\zeta(\kappa) & \approx \frac{\sqrt{S(\mu_Y)}}{6} + \left( \frac{K(\mu_Y) - 3}{8} + 15 \frac{S(\mu_Y)}{72} \right) \kappa - \left( \frac{\sqrt{S(\mu_Y)}}{6} \right) \kappa^2 \\
& - \left( \frac{K(\mu_Y) - 3}{24} + 5 \frac{S(\mu_Y)}{36} \right) \kappa^3 - \frac{\sqrt{S(\mu_Y)}}{72} \kappa^5
\end{align*}$$

126
The confidence limits is then defined as follows,

\[
SIR_{lb} \approx SIR - \left( z_{1 - \frac{\alpha}{2}} + \zeta (1 - \frac{\alpha}{2}) \right) \sqrt{\text{Var}(SIR)}
\]

\[
SIR_{ub} \approx SIR + \left( z_{1 - \frac{\alpha}{2}} + \zeta (\frac{\alpha}{2}) \right) \sqrt{\text{Var}(SIR)}
\]

where \(\sqrt{\text{Var}(SIR)}\) is estimated in the same way as S3. Note: the sign in the percentile adjustment in the lower bound is changed from minus to plus to accommodate the fact that the skewness and kurtosis estimated is for the denominator, not the whole function.

### A.2 Estimating the sufficient number of repetitions for convergence in CI coverage statistics

Ensuring the number of resamples is sufficient for convergence to the true value is an important implementational detail in chapter 2. To estimate the number of repetitions deemed sufficient, the median hospital in the VAED dataset is repeatedly resampled to see when coverage rates start to converge. This median hospital had 5000-10000 separations that year and around 20-50 mortalities; exact numbers and year are not released for confidentiality reasons.

Results indicate that, for the high variance (DRG) scenario, the difference between methods are exaggerated enough that so that 100 resamples per hospital over the VAED sample is sufficient. More resamples would be preferred in the low variance case. However, the loss of precision for using a less accurate approximation method is also much less of a problem if the variance of expected incidents is smaller to begin with.
### Table A.1: Number of resamples, D1 MDC

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.3045</td>
<td>0.5739</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.97%</td>
<td>3.03%</td>
<td>0.3074</td>
<td>0.5755</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.91%</td>
<td>0.09%</td>
<td>0.3102</td>
<td>0.5788</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>93.33%</td>
<td>6.67%</td>
<td>0.3082</td>
<td>0.5751</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>90.54%</td>
<td>9.46%</td>
<td>0.3081</td>
<td>0.5754</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>90.61%</td>
<td>9.39%</td>
<td>0.3051</td>
<td>0.5730</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>89.58%</td>
<td>10.42%</td>
<td>0.3044</td>
<td>0.5706</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>88.12%</td>
<td>11.88%</td>
<td>0.3038</td>
<td>0.5693</td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.51%</td>
<td>12.49%</td>
<td>0.3047</td>
<td>0.5712</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>86.95%</td>
<td>13.05%</td>
<td>0.3022</td>
<td>0.5670</td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.05%</td>
<td>12.95%</td>
<td>0.3025</td>
<td>0.5675</td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.69%</td>
<td>12.31%</td>
<td>0.3014</td>
<td>0.5638</td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.61%</td>
<td>12.39%</td>
<td>0.3032</td>
<td>0.5687</td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.26%</td>
<td>12.74%</td>
<td>0.3026</td>
<td>0.5679</td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.22%</td>
<td>12.78%</td>
<td>0.3033</td>
<td>0.5699</td>
</tr>
</tbody>
</table>

The percentages in the first five columns correspond to the coverage categories as defined in the main text. The percentages indicate the number of resampled CIs that belong to a given category. For example, at N=90, 93.33% of the resampled CIs of the median hospital (84 out of 99) include the true SIR. This value slowly converges to around 87% as N exceeds 1000.

### Table A.2: Number of resamples, D1 DRG

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>10.00%</td>
<td>90.00%</td>
<td>0.3445</td>
<td>0.6279</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>6.06%</td>
<td>93.94%</td>
<td>0.2751</td>
<td>0.5025</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>9.09%</td>
<td>90.91%</td>
<td>0.2456</td>
<td>0.4494</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.89%</td>
<td>91.11%</td>
<td>0.2173</td>
<td>0.4043</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>10.81%</td>
<td>89.19%</td>
<td>0.2437</td>
<td>0.4560</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.98%</td>
<td>90.61%</td>
<td>0.3247</td>
<td>0.6067</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.68%</td>
<td>91.07%</td>
<td>0.2981</td>
<td>0.5416</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.97%</td>
<td>91.88%</td>
<td>0.2670</td>
<td>0.5019</td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.84%</td>
<td>91.80%</td>
<td>0.2503</td>
<td>0.4759</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.85%</td>
<td>91.70%</td>
<td>0.2387</td>
<td>0.4493</td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.98%</td>
<td>91.61%</td>
<td>0.2369</td>
<td>0.4451</td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.00%</td>
<td>91.64%</td>
<td>0.2331</td>
<td>0.4385</td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.68%</td>
<td>91.98%</td>
<td>0.2276</td>
<td>0.4275</td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.43%</td>
<td>92.23%</td>
<td>0.2262</td>
<td>0.4253</td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>7.44%</td>
<td>92.19%</td>
<td>0.2279</td>
<td>0.4279</td>
</tr>
</tbody>
</table>
Table A.3: Number of resamples, D2 MDC

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2954</td>
<td>0.5746</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.97%</td>
<td>3.03%</td>
<td>0.2985</td>
<td>0.5762</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>90.91%</td>
<td>9.09%</td>
<td>0.3013</td>
<td>0.5796</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>94.44%</td>
<td>5.56%</td>
<td>0.2994</td>
<td>0.5775</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>92.57%</td>
<td>7.43%</td>
<td>0.2992</td>
<td>0.5762</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>91.84%</td>
<td>8.16%</td>
<td>0.2981</td>
<td>0.5737</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>91.07%</td>
<td>8.93%</td>
<td>0.2983</td>
<td>0.5751</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>89.17%</td>
<td>10.83%</td>
<td>0.2979</td>
<td>0.5701</td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>88.88%</td>
<td>11.12%</td>
<td>0.2998</td>
<td>0.5720</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>88.22%</td>
<td>11.78%</td>
<td>0.2984</td>
<td>0.5677</td>
</tr>
</tbody>
</table>

Table A.4: Number of resamples, D2 DRG

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>15.00%</td>
<td>85.00%</td>
<td>0.3353</td>
<td>0.6288</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>15.15%</td>
<td>84.85%</td>
<td>0.2677</td>
<td>0.5032</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>16.36%</td>
<td>83.64%</td>
<td>0.2390</td>
<td>0.4501</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>18.92%</td>
<td>81.08%</td>
<td>0.2366</td>
<td>0.4566</td>
</tr>
<tr>
<td>148</td>
<td>0.25%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.84%</td>
<td>78.18%</td>
<td>0.2376</td>
<td>0.5423</td>
</tr>
<tr>
<td>245</td>
<td>0.41%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>22.56%</td>
<td>77.29%</td>
<td>0.2591</td>
<td>0.5026</td>
</tr>
<tr>
<td>403</td>
<td>0.15%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>665</td>
<td>0.15%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>1097</td>
<td>0.36%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>1808</td>
<td>0.44%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>2981</td>
<td>0.40%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>4915</td>
<td>0.35%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>8103</td>
<td>0.31%</td>
<td>0.03%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
<tr>
<td>13360</td>
<td>0.35%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>20.78%</td>
<td>78.18%</td>
<td>0.2429</td>
<td>0.4715</td>
</tr>
</tbody>
</table>

Table A.5: Number of resamples, D3 MDC

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>95.00%</td>
<td>5.00%</td>
<td>0.3008</td>
<td>0.5594</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>87.88%</td>
<td>12.12%</td>
<td>0.3051</td>
<td>0.5631</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>83.64%</td>
<td>16.36%</td>
<td>0.2983</td>
<td>0.5540</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>80.00%</td>
<td>20.00%</td>
<td>0.2990</td>
<td>0.5529</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>79.73%</td>
<td>20.27%</td>
<td>0.3008</td>
<td>0.5573</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>79.59%</td>
<td>20.41%</td>
<td>0.3005</td>
<td>0.5595</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>79.16%</td>
<td>20.84%</td>
<td>0.3013</td>
<td>0.5589</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>78.35%</td>
<td>21.65%</td>
<td>0.2980</td>
<td>0.5538</td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>78.49%</td>
<td>21.51%</td>
<td>0.2972</td>
<td>0.5535</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>77.27%</td>
<td>22.73%</td>
<td>0.2940</td>
<td>0.5485</td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>76.52%</td>
<td>23.48%</td>
<td>0.2936</td>
<td>0.5483</td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>75.99%</td>
<td>24.01%</td>
<td>0.2921</td>
<td>0.5460</td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>75.92%</td>
<td>24.08%</td>
<td>0.2932</td>
<td>0.5480</td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>75.93%</td>
<td>24.07%</td>
<td>0.2926</td>
<td>0.5471</td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>75.72%</td>
<td>24.28%</td>
<td>0.2927</td>
<td>0.5474</td>
</tr>
<tr>
<td>N</td>
<td>zero CI bound(s)</td>
<td>fail, solution</td>
<td>fail, negative</td>
<td>cover</td>
<td>not cover</td>
<td>Avg. lower bound</td>
<td>Avg. upper bound</td>
</tr>
<tr>
<td>----</td>
<td>------------------</td>
<td>----------------</td>
<td>----------------</td>
<td>-------</td>
<td>-----------</td>
<td>------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>5.00%</td>
<td>20.00%</td>
<td>75.00%</td>
<td>0.1894</td>
<td>0.6293</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.03%</td>
<td>27.27%</td>
<td>69.70%</td>
<td>0.1770</td>
<td>0.5128</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.82%</td>
<td>27.27%</td>
<td>70.91%</td>
<td>0.1870</td>
<td>0.5156</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.11%</td>
<td>23.33%</td>
<td>75.56%</td>
<td>0.1720</td>
<td>0.4594</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.35%</td>
<td>29.05%</td>
<td>69.59%</td>
<td>0.1872</td>
<td>0.5254</td>
</tr>
<tr>
<td>245</td>
<td>0.41%</td>
<td>0.00%</td>
<td>1.22%</td>
<td>28.57%</td>
<td>69.80%</td>
<td>0.1746</td>
<td>0.4729</td>
</tr>
<tr>
<td>403</td>
<td>0.25%</td>
<td>0.00%</td>
<td>1.24%</td>
<td>27.05%</td>
<td>71.46%</td>
<td>0.1759</td>
<td>0.5184</td>
</tr>
<tr>
<td>665</td>
<td>0.15%</td>
<td>0.00%</td>
<td>1.50%</td>
<td>28.57%</td>
<td>69.77%</td>
<td>0.1768</td>
<td>0.5013</td>
</tr>
<tr>
<td>1097</td>
<td>0.36%</td>
<td>0.00%</td>
<td>1.55%</td>
<td>26.16%</td>
<td>71.92%</td>
<td>0.1725</td>
<td>0.4901</td>
</tr>
<tr>
<td>1808</td>
<td>0.44%</td>
<td>0.00%</td>
<td>1.38%</td>
<td>26.88%</td>
<td>71.29%</td>
<td>0.1708</td>
<td>0.4711</td>
</tr>
<tr>
<td>2981</td>
<td>0.40%</td>
<td>0.00%</td>
<td>1.58%</td>
<td>26.90%</td>
<td>71.12%</td>
<td>0.1711</td>
<td>0.4685</td>
</tr>
<tr>
<td>4915</td>
<td>0.03%</td>
<td>0.00%</td>
<td>1.55%</td>
<td>26.16%</td>
<td>71.92%</td>
<td>0.1725</td>
<td>0.4901</td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.00%</td>
<td>1.52%</td>
<td>26.16%</td>
<td>71.92%</td>
<td>0.1725</td>
<td>0.4901</td>
</tr>
<tr>
<td>13360</td>
<td>0.39%</td>
<td>0.00%</td>
<td>1.56%</td>
<td>25.37%</td>
<td>72.40%</td>
<td>0.1679</td>
<td>0.4890</td>
</tr>
<tr>
<td>22026</td>
<td>0.35%</td>
<td>0.00%</td>
<td>1.56%</td>
<td>25.37%</td>
<td>72.71%</td>
<td>0.1677</td>
<td>0.4596</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>90.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2319</td>
<td>1.0121</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>87.88%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2351</td>
<td>1.0010</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>81.82%</td>
<td>0.00%</td>
<td>97.00%</td>
<td>10.91%</td>
<td>0.2369</td>
<td>0.9897</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>83.11%</td>
<td>0.00%</td>
<td>97.86%</td>
<td>10.14%</td>
<td>0.2343</td>
<td>0.9816</td>
</tr>
<tr>
<td>245</td>
<td>0.41%</td>
<td>82.04%</td>
<td>0.00%</td>
<td>98.76%</td>
<td>10.14%</td>
<td>0.2330</td>
<td>0.9777</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>80.89%</td>
<td>0.00%</td>
<td>98.56%</td>
<td>10.14%</td>
<td>0.2331</td>
<td>0.9863</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>80.51%</td>
<td>0.00%</td>
<td>98.58%</td>
<td>10.14%</td>
<td>0.2330</td>
<td>0.9900</td>
</tr>
<tr>
<td>1097</td>
<td>0.01%</td>
<td>80.06%</td>
<td>0.00%</td>
<td>98.49%</td>
<td>10.14%</td>
<td>0.2330</td>
<td>0.9900</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>79.99%</td>
<td>0.00%</td>
<td>98.49%</td>
<td>10.14%</td>
<td>0.2330</td>
<td>0.9900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>10.00%</td>
<td>90.00%</td>
<td>0.00%</td>
<td>0.0240</td>
<td>0.0708</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>87.88%</td>
<td>0.00%</td>
<td>6.06%</td>
<td>94.00%</td>
<td>0.0351</td>
<td>2.3631</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>81.82%</td>
<td>0.00%</td>
<td>7.27%</td>
<td>92.73%</td>
<td>0.0401</td>
<td>1.4718</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>81.11%</td>
<td>0.00%</td>
<td>7.51%</td>
<td>92.49%</td>
<td>0.0428</td>
<td>1.1120</td>
</tr>
<tr>
<td>245</td>
<td>0.41%</td>
<td>82.04%</td>
<td>0.00%</td>
<td>6.76%</td>
<td>93.24%</td>
<td>0.0433</td>
<td>1.1153</td>
</tr>
<tr>
<td>403</td>
<td>0.25%</td>
<td>80.89%</td>
<td>0.00%</td>
<td>6.70%</td>
<td>93.30%</td>
<td>0.0423</td>
<td>0.8527</td>
</tr>
<tr>
<td>665</td>
<td>0.15%</td>
<td>82.56%</td>
<td>0.00%</td>
<td>6.92%</td>
<td>93.08%</td>
<td>0.0439</td>
<td>1.5448</td>
</tr>
<tr>
<td>1097</td>
<td>0.31%</td>
<td>80.06%</td>
<td>0.00%</td>
<td>7.96%</td>
<td>92.04%</td>
<td>0.0453</td>
<td>1.4789</td>
</tr>
<tr>
<td>1808</td>
<td>0.44%</td>
<td>80.25%</td>
<td>0.00%</td>
<td>8.19%</td>
<td>91.81%</td>
<td>0.0448</td>
<td>1.3294</td>
</tr>
<tr>
<td>2981</td>
<td>0.40%</td>
<td>80.51%</td>
<td>0.00%</td>
<td>8.25%</td>
<td>91.75%</td>
<td>0.0445</td>
<td>1.3648</td>
</tr>
<tr>
<td>4915</td>
<td>0.35%</td>
<td>80.55%</td>
<td>0.00%</td>
<td>8.44%</td>
<td>91.56%</td>
<td>0.0453</td>
<td>1.3847</td>
</tr>
<tr>
<td>8103</td>
<td>0.32%</td>
<td>80.06%</td>
<td>0.00%</td>
<td>8.75%</td>
<td>90.25%</td>
<td>0.0453</td>
<td>1.4926</td>
</tr>
<tr>
<td>13360</td>
<td>0.31%</td>
<td>79.99%</td>
<td>0.00%</td>
<td>8.91%</td>
<td>90.09%</td>
<td>0.0456</td>
<td>1.4624</td>
</tr>
<tr>
<td>22026</td>
<td>0.35%</td>
<td>79.73%</td>
<td>0.00%</td>
<td>9.14%</td>
<td>89.86%</td>
<td>0.0456</td>
<td>1.5362</td>
</tr>
</tbody>
</table>
Table A.9: Number of resamples, N2 MDC

<table>
<thead>
<tr>
<th>N2</th>
<th>MDC</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2307</td>
<td>0.6920</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2340</td>
<td>0.6949</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.36%</td>
<td>3.64%</td>
<td>0.2358</td>
<td>0.6999</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.78%</td>
<td>2.22%</td>
<td>0.2352</td>
<td>0.6975</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.30%</td>
<td>2.70%</td>
<td>0.2346</td>
<td>0.6946</td>
<td></td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.14%</td>
<td>2.86%</td>
<td>0.2317</td>
<td>0.6911</td>
<td></td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.26%</td>
<td>1.74%</td>
<td>0.2319</td>
<td>0.6896</td>
<td></td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.35%</td>
<td>1.65%</td>
<td>0.2316</td>
<td>0.6871</td>
<td></td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.06%</td>
<td>1.94%</td>
<td>0.2303</td>
<td>0.6840</td>
<td></td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.06%</td>
<td>1.94%</td>
<td>0.2303</td>
<td>0.6840</td>
<td></td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.03%</td>
<td>0.00%</td>
<td>98.74%</td>
<td>1.25%</td>
<td>0.2297</td>
<td>0.6824</td>
<td></td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>98.70%</td>
<td>2.08%</td>
<td>0.2309</td>
<td>0.6861</td>
<td></td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>98.89%</td>
<td>1.90%</td>
<td>0.2307</td>
<td>0.6871</td>
<td></td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>97.93%</td>
<td>2.06%</td>
<td>0.2304</td>
<td>0.6850</td>
<td></td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.95%</td>
<td>2.04%</td>
<td>0.2308</td>
<td>0.6864</td>
<td></td>
</tr>
</tbody>
</table>

Table A.10: Number of resamples, N2 DRG

<table>
<thead>
<tr>
<th>N2</th>
<th>DRG</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>85.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>15.00%</td>
<td>0.0242</td>
<td>0.1078</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>84.85%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>15.15%</td>
<td>0.0291</td>
<td>0.1302</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>80.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.00%</td>
<td>0.0318</td>
<td>0.1344</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>80.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.00%</td>
<td>0.0383</td>
<td>0.1393</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>0.41%</td>
<td>82.43%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>17.57%</td>
<td>0.0391</td>
<td>0.1397</td>
<td></td>
</tr>
<tr>
<td>245</td>
<td>0.25%</td>
<td>80.65%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>19.11%</td>
<td>0.0391</td>
<td>0.1359</td>
<td></td>
</tr>
<tr>
<td>403</td>
<td>0.15%</td>
<td>82.41%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>17.44%</td>
<td>0.0406</td>
<td>0.1433</td>
<td></td>
</tr>
<tr>
<td>665</td>
<td>0.36%</td>
<td>81.31%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>18.32%</td>
<td>0.0401</td>
<td>0.1411</td>
<td></td>
</tr>
<tr>
<td>1097</td>
<td>0.44%</td>
<td>80.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>19.47%</td>
<td>0.0407</td>
<td>0.1437</td>
<td></td>
</tr>
<tr>
<td>1808</td>
<td>0.40%</td>
<td>80.41%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>19.19%</td>
<td>0.0405</td>
<td>0.1430</td>
<td></td>
</tr>
<tr>
<td>2981</td>
<td>0.35%</td>
<td>80.43%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>19.23%</td>
<td>0.0412</td>
<td>0.1457</td>
<td></td>
</tr>
<tr>
<td>4915</td>
<td>0.32%</td>
<td>79.97%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>19.71%</td>
<td>0.0413</td>
<td>0.1458</td>
<td></td>
</tr>
<tr>
<td>8103</td>
<td>0.31%</td>
<td>79.18%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>19.81%</td>
<td>0.0415</td>
<td>0.1468</td>
<td></td>
</tr>
<tr>
<td>13360</td>
<td>0.35%</td>
<td>79.64%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>20.01%</td>
<td>0.0415</td>
<td>0.1472</td>
<td></td>
</tr>
</tbody>
</table>

Table A.11: Number of resamples, S1 MDC

<table>
<thead>
<tr>
<th>N1</th>
<th>MDC</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2319</td>
<td>1.0118</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2351</td>
<td>1.0007</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.36%</td>
<td>3.64%</td>
<td>0.2370</td>
<td>1.0128</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.78%</td>
<td>2.22%</td>
<td>0.2343</td>
<td>0.9935</td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.30%</td>
<td>2.70%</td>
<td>0.2357</td>
<td>1.0020</td>
<td></td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.14%</td>
<td>2.86%</td>
<td>0.2329</td>
<td>1.0030</td>
<td></td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.76%</td>
<td>1.24%</td>
<td>0.2331</td>
<td>0.9952</td>
<td></td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.65%</td>
<td>1.35%</td>
<td>0.2328</td>
<td>0.9852</td>
<td></td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.72%</td>
<td>1.28%</td>
<td>0.2331</td>
<td>0.9975</td>
<td></td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.56%</td>
<td>1.44%</td>
<td>0.2314</td>
<td>0.9861</td>
<td></td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.03%</td>
<td>0.00%</td>
<td>98.19%</td>
<td>1.78%</td>
<td>0.2315</td>
<td>0.9880</td>
<td></td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>98.33%</td>
<td>1.65%</td>
<td>0.2309</td>
<td>0.9817</td>
<td></td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>98.62%</td>
<td>1.37%</td>
<td>0.2320</td>
<td>0.9897</td>
<td></td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>98.53%</td>
<td>1.47%</td>
<td>0.2316</td>
<td>0.9901</td>
<td></td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.49%</td>
<td>1.51%</td>
<td>0.2320</td>
<td>0.9947</td>
<td></td>
</tr>
</tbody>
</table>
### Table A.12: Number of resamples, S1 DRG

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>90.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>10.00%</td>
<td>0.0240</td>
<td>0.0708</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>87.88%</td>
<td>0.00%</td>
<td>6.06%</td>
<td>6.06%</td>
<td>0.0351</td>
<td>2.3298</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>81.82%</td>
<td>0.00%</td>
<td>7.27%</td>
<td>10.91%</td>
<td>0.0401</td>
<td>1.4560</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>81.11%</td>
<td>0.00%</td>
<td>7.28%</td>
<td>11.11%</td>
<td>0.0428</td>
<td>1.0223</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>83.11%</td>
<td>0.00%</td>
<td>6.76%</td>
<td>10.14%</td>
<td>0.0433</td>
<td>1.1062</td>
</tr>
<tr>
<td>245</td>
<td>0.41%</td>
<td>82.04%</td>
<td>0.00%</td>
<td>6.94%</td>
<td>10.61%</td>
<td>0.0441</td>
<td>2.3298</td>
</tr>
<tr>
<td>403</td>
<td>0.25%</td>
<td>80.89%</td>
<td>0.00%</td>
<td>6.70%</td>
<td>12.16%</td>
<td>0.0423</td>
<td>0.8473</td>
</tr>
<tr>
<td>665</td>
<td>0.15%</td>
<td>82.56%</td>
<td>0.00%</td>
<td>6.92%</td>
<td>10.38%</td>
<td>0.0446</td>
<td>1.5504</td>
</tr>
<tr>
<td>1097</td>
<td>0.36%</td>
<td>81.40%</td>
<td>0.00%</td>
<td>7.29%</td>
<td>10.94%</td>
<td>0.0439</td>
<td>1.4070</td>
</tr>
<tr>
<td>1808</td>
<td>0.44%</td>
<td>80.51%</td>
<td>0.00%</td>
<td>8.19%</td>
<td>10.78%</td>
<td>0.0456</td>
<td>1.5609</td>
</tr>
</tbody>
</table>

### Table A.13: Number of resamples, S2 MDC

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>10.00%</td>
<td>90.00%</td>
<td>0.4147</td>
<td>0.4214</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>6.06%</td>
<td>93.94%</td>
<td>0.4173</td>
<td>0.4240</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>4.44%</td>
<td>95.56%</td>
<td>0.4177</td>
<td>0.4243</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.38%</td>
<td>96.62%</td>
<td>0.4178</td>
<td>0.4244</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.23%</td>
<td>96.77%</td>
<td>0.4148</td>
<td>0.4201</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.71%</td>
<td>97.29%</td>
<td>0.4126</td>
<td>0.4192</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.83%</td>
<td>97.17%</td>
<td>0.4139</td>
<td>0.4205</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.37%</td>
<td>96.63%</td>
<td>0.4107</td>
<td>0.4172</td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.29%</td>
<td>96.71%</td>
<td>0.4111</td>
<td>0.4176</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.01%</td>
<td>96.99%</td>
<td>0.4097</td>
<td>0.4162</td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.98%</td>
<td>97.02%</td>
<td>0.4112</td>
<td>0.4178</td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.03%</td>
<td>96.97%</td>
<td>0.4121</td>
<td>0.4187</td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.01%</td>
<td>96.99%</td>
<td>0.4119</td>
<td>0.4185</td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>2.98%</td>
<td>97.02%</td>
<td>0.4112</td>
<td>0.4178</td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.03%</td>
<td>96.97%</td>
<td>0.4121</td>
<td>0.4187</td>
</tr>
</tbody>
</table>

### Table A.14: Number of resamples, S2 DRG

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>5.00%</td>
<td>95.00%</td>
<td>0.3972</td>
<td>0.5564</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>6.06%</td>
<td>93.94%</td>
<td>0.3285</td>
<td>0.4294</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>9.09%</td>
<td>90.91%</td>
<td>0.3024</td>
<td>0.3708</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>6.67%</td>
<td>93.33%</td>
<td>0.2748</td>
<td>0.3235</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>4.73%</td>
<td>95.27%</td>
<td>0.3046</td>
<td>0.3708</td>
</tr>
<tr>
<td>245</td>
<td>0.41%</td>
<td>0.41%</td>
<td>0.00%</td>
<td>3.27%</td>
<td>95.92%</td>
<td>0.2837</td>
<td>0.3328</td>
</tr>
<tr>
<td>403</td>
<td>0.25%</td>
<td>0.25%</td>
<td>0.00%</td>
<td>2.98%</td>
<td>96.53%</td>
<td>0.2869</td>
<td>0.3423</td>
</tr>
<tr>
<td>665</td>
<td>0.15%</td>
<td>0.15%</td>
<td>0.00%</td>
<td>2.86%</td>
<td>96.84%</td>
<td>0.2915</td>
<td>0.3454</td>
</tr>
<tr>
<td>1097</td>
<td>0.36%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>2.64%</td>
<td>96.96%</td>
<td>0.2856</td>
<td>0.3547</td>
</tr>
<tr>
<td>1808</td>
<td>0.44%</td>
<td>0.06%</td>
<td>0.00%</td>
<td>2.60%</td>
<td>96.90%</td>
<td>0.2822</td>
<td>0.3510</td>
</tr>
<tr>
<td>2981</td>
<td>0.40%</td>
<td>0.03%</td>
<td>0.00%</td>
<td>2.35%</td>
<td>97.22%</td>
<td>0.2850</td>
<td>0.3614</td>
</tr>
<tr>
<td>4915</td>
<td>0.35%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>2.32%</td>
<td>97.29%</td>
<td>0.2834</td>
<td>0.3737</td>
</tr>
<tr>
<td>8103</td>
<td>0.32%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>2.16%</td>
<td>97.54%</td>
<td>0.2807</td>
<td>0.3572</td>
</tr>
<tr>
<td>13360</td>
<td>0.31%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>2.26%</td>
<td>97.38%</td>
<td>0.2786</td>
<td>0.3541</td>
</tr>
<tr>
<td>22026</td>
<td>0.35%</td>
<td>0.05%</td>
<td>0.00%</td>
<td>2.43%</td>
<td>97.17%</td>
<td>0.2764</td>
<td>0.3470</td>
</tr>
</tbody>
</table>
Table A.15: Number of resamples, S3 MDC

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2022</td>
<td>0.7514</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.2058</td>
<td>0.7504</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.36%</td>
<td>3.64%</td>
<td>0.2071</td>
<td>0.7567</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.78%</td>
<td>2.22%</td>
<td>0.2072</td>
<td>0.7483</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.30%</td>
<td>2.70%</td>
<td>0.2064</td>
<td>0.7504</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.14%</td>
<td>2.86%</td>
<td>0.2036</td>
<td>0.7485</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.01%</td>
<td>1.99%</td>
<td>0.2044</td>
<td>0.7428</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>98.20%</td>
<td>1.80%</td>
<td>0.2043</td>
<td>0.7403</td>
</tr>
<tr>
<td>1360</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.93%</td>
<td>1.82%</td>
<td>0.2031</td>
<td>0.7376</td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.78%</td>
<td>2.22%</td>
<td>0.2034</td>
<td>0.7413</td>
</tr>
</tbody>
</table>

Table A.16: Number of resamples, S3 DRG

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>5.00%</td>
<td>20.00%</td>
<td>45.00%</td>
<td>30.00%</td>
<td>0.0220</td>
<td>0.7576</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>3.03%</td>
<td>15.15%</td>
<td>57.58%</td>
<td>24.24%</td>
<td>0.0233</td>
<td>0.7774</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>1.82%</td>
<td>17.88%</td>
<td>54.44%</td>
<td>26.67%</td>
<td>0.0229</td>
<td>0.7401</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>1.11%</td>
<td>20.00%</td>
<td>54.29%</td>
<td>26.12%</td>
<td>0.0235</td>
<td>0.7433</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>1.81%</td>
<td>17.78%</td>
<td>53.35%</td>
<td>27.54%</td>
<td>0.0208</td>
<td>0.8728</td>
</tr>
<tr>
<td>245</td>
<td>0.25%</td>
<td>2.48%</td>
<td>16.38%</td>
<td>54.59%</td>
<td>26.32%</td>
<td>0.0207</td>
<td>0.9186</td>
</tr>
<tr>
<td>403</td>
<td>0.15%</td>
<td>2.41%</td>
<td>16.54%</td>
<td>54.20%</td>
<td>26.12%</td>
<td>0.0207</td>
<td>0.9186</td>
</tr>
<tr>
<td>665</td>
<td>0.44%</td>
<td>2.05%</td>
<td>15.76%</td>
<td>54.42%</td>
<td>27.32%</td>
<td>0.0212</td>
<td>0.8755</td>
</tr>
<tr>
<td>1097</td>
<td>0.40%</td>
<td>2.06%</td>
<td>16.34%</td>
<td>53.35%</td>
<td>27.81%</td>
<td>0.0212</td>
<td>0.8655</td>
</tr>
<tr>
<td>1808</td>
<td>0.32%</td>
<td>1.95%</td>
<td>15.94%</td>
<td>53.62%</td>
<td>28.16%</td>
<td>0.0218</td>
<td>0.8452</td>
</tr>
<tr>
<td>3360</td>
<td>0.31%</td>
<td>1.92%</td>
<td>15.64%</td>
<td>53.81%</td>
<td>28.32%</td>
<td>0.0218</td>
<td>0.8419</td>
</tr>
<tr>
<td>22026</td>
<td>0.35%</td>
<td>1.87%</td>
<td>15.78%</td>
<td>53.27%</td>
<td>28.73%</td>
<td>0.0220</td>
<td>0.8331</td>
</tr>
</tbody>
</table>

Table A.17: Number of resamples, S4 MDC

<table>
<thead>
<tr>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail solution</th>
<th>fail negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.1280</td>
<td>0.7581</td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>0.00%</td>
<td>0.1328</td>
<td>0.7084</td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.36%</td>
<td>3.64%</td>
<td>0.1334</td>
<td>0.7142</td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.78%</td>
<td>2.22%</td>
<td>0.1349</td>
<td>0.7071</td>
</tr>
<tr>
<td>148</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.62%</td>
<td>3.38%</td>
<td>0.1346</td>
<td>0.7086</td>
</tr>
<tr>
<td>245</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.73%</td>
<td>3.27%</td>
<td>0.1303</td>
<td>0.7060</td>
</tr>
<tr>
<td>403</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.52%</td>
<td>2.48%</td>
<td>0.1322</td>
<td>0.7013</td>
</tr>
<tr>
<td>665</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.44%</td>
<td>2.56%</td>
<td>0.1326</td>
<td>0.6992</td>
</tr>
<tr>
<td>1097</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.45%</td>
<td>2.55%</td>
<td>0.1318</td>
<td>0.7026</td>
</tr>
<tr>
<td>1808</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.90%</td>
<td>3.10%</td>
<td>0.1314</td>
<td>0.6965</td>
</tr>
<tr>
<td>2981</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.85%</td>
<td>3.12%</td>
<td>0.1313</td>
<td>0.6971</td>
</tr>
<tr>
<td>4915</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>96.99%</td>
<td>2.99%</td>
<td>0.1313</td>
<td>0.6944</td>
</tr>
<tr>
<td>8103</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.28%</td>
<td>2.70%</td>
<td>0.1316</td>
<td>0.6988</td>
</tr>
<tr>
<td>13360</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.19%</td>
<td>2.80%</td>
<td>0.1312</td>
<td>0.6996</td>
</tr>
<tr>
<td>22026</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>97.19%</td>
<td>2.80%</td>
<td>0.1312</td>
<td>0.6996</td>
</tr>
</tbody>
</table>
### Table A.18: Number of resamples, S4 DRG

<table>
<thead>
<tr>
<th>S4</th>
<th>DRG</th>
<th>N</th>
<th>zero CI bound(s)</th>
<th>fail, solution</th>
<th>fail, negative</th>
<th>cover</th>
<th>not cover</th>
<th>Avg. lower bound</th>
<th>Avg. upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.00%</td>
<td>5.00%</td>
<td>85.00%</td>
<td>0.00%</td>
<td>10.00%</td>
<td>0.0185</td>
<td>0.0627</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>0.00%</td>
<td>3.03%</td>
<td>87.88%</td>
<td>0.00%</td>
<td>12.73%</td>
<td>0.0195</td>
<td>0.1135</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>0.00%</td>
<td>1.82%</td>
<td>85.45%</td>
<td>0.00%</td>
<td>15.56%</td>
<td>0.0171</td>
<td>0.1402</td>
<td></td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>0.00%</td>
<td>1.11%</td>
<td>83.33%</td>
<td>0.00%</td>
<td>16.19%</td>
<td>0.0214</td>
<td>0.2558</td>
<td></td>
<td></td>
</tr>
<tr>
<td>148</td>
<td>0.41%</td>
<td>2.45%</td>
<td>81.22%</td>
<td>0.41%</td>
<td>15.51%</td>
<td>0.0188</td>
<td>0.2122</td>
<td></td>
<td></td>
</tr>
<tr>
<td>403</td>
<td>0.25%</td>
<td>2.48%</td>
<td>80.89%</td>
<td>0.25%</td>
<td>16.13%</td>
<td>0.0185</td>
<td>0.1759</td>
<td></td>
<td></td>
</tr>
<tr>
<td>665</td>
<td>0.15%</td>
<td>2.41%</td>
<td>83.01%</td>
<td>0.15%</td>
<td>14.29%</td>
<td>0.0183</td>
<td>0.1692</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1097</td>
<td>0.36%</td>
<td>2.37%</td>
<td>81.86%</td>
<td>0.18%</td>
<td>15.22%</td>
<td>0.0197</td>
<td>0.1667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1808</td>
<td>0.44%</td>
<td>2.21%</td>
<td>80.97%</td>
<td>0.28%</td>
<td>16.10%</td>
<td>0.0206</td>
<td>0.1852</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2991</td>
<td>0.40%</td>
<td>2.28%</td>
<td>81.21%</td>
<td>0.17%</td>
<td>15.93%</td>
<td>0.0267</td>
<td>0.1731</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4915</td>
<td>0.35%</td>
<td>2.10%</td>
<td>81.44%</td>
<td>0.16%</td>
<td>16.05%</td>
<td>0.0287</td>
<td>0.1787</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8103</td>
<td>0.32%</td>
<td>2.15%</td>
<td>80.82%</td>
<td>0.14%</td>
<td>16.57%</td>
<td>0.0249</td>
<td>0.1722</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13360</td>
<td>0.31%</td>
<td>2.15%</td>
<td>80.82%</td>
<td>0.13%</td>
<td>16.59%</td>
<td>0.0253</td>
<td>0.1755</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22026</td>
<td>0.35%</td>
<td>2.07%</td>
<td>80.73%</td>
<td>0.12%</td>
<td>16.73%</td>
<td>0.0250</td>
<td>0.1745</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A.3 Comparing the location of analytical CIs versus benchmark results

The accuracy of an analytical method is measured by how close their CI bounds are to benchmark ones that were estimated using parametric simulation. A sample of 5 hospitals are chosen, each with 5000 generated resamples. CIs are computed for every resample and then compared against benchmark values. A good fit would see most of the analytical CIs in categories 1-3, with similar percentages for categories 2 and 3.

Table A.19: Location of analytical CIs relative to benchmark simulation, D1

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>27.62%</td>
<td>35.50%</td>
<td>30.88%</td>
<td>48.88%</td>
<td>82.14%</td>
<td>70.74%</td>
<td>63.34%</td>
<td>82.22%</td>
<td>85.82%</td>
<td>91.24%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>27.00%</td>
<td>33.34%</td>
<td>30.84%</td>
<td>25.94%</td>
<td>8.40%</td>
<td>19.36%</td>
<td>19.74%</td>
<td>11.04%</td>
<td>7.86%</td>
<td>3.36%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>44.90%</td>
<td>31.14%</td>
<td>37.88%</td>
<td>25.06%</td>
<td>8.24%</td>
<td>8.48%</td>
<td>16.46%</td>
<td>5.42%</td>
<td>4.74%</td>
<td>3.02%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.38%</td>
<td>0.02%</td>
<td>0.52%</td>
<td>0.00%</td>
<td>0.29%</td>
<td>0.44%</td>
<td>0.42%</td>
<td>1.16%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.22%</td>
<td>0.16%</td>
<td>0.76%</td>
<td>0.54%</td>
<td>0.20%</td>
<td>0.86%</td>
<td>1.16%</td>
<td>1.22%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>0.44%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

The sample hospitals are labelled 1-5 in ascending order by number of patients. Read down the first column, then across. For example, the value of 27.62% in the upper-left corner indicates that out of the 5000 CIs estimated from low-variance (MDC) resamples, 1381 of them are contained within the benchmark CI bounds. Refer to main text for definitions of location categories and approximate statistics of sampled hospitals.

Table A.20: Location of analytical CIs relative to benchmark simulation, D2

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>7.42%</td>
<td>27.66%</td>
<td>24.86%</td>
<td>45.86%</td>
<td>81.58%</td>
<td>57.90%</td>
<td>58.26%</td>
<td>80.72%</td>
<td>85.36%</td>
<td>91.20%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>35.60%</td>
<td>39.36%</td>
<td>36.04%</td>
<td>28.34%</td>
<td>8.78%</td>
<td>30.16%</td>
<td>23.84%</td>
<td>12.48%</td>
<td>8.30%</td>
<td>3.36%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>51.16%</td>
<td>32.96%</td>
<td>38.78%</td>
<td>25.72%</td>
<td>8.44%</td>
<td>10.62%</td>
<td>17.48%</td>
<td>5.56%</td>
<td>4.82%</td>
<td>3.06%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.38%</td>
<td>0.02%</td>
<td>0.52%</td>
<td>0.00%</td>
<td>0.18%</td>
<td>0.42%</td>
<td>0.40%</td>
<td>1.16%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.68%</td>
<td>0.36%</td>
<td>0.22%</td>
<td>0.80%</td>
<td>1.12%</td>
<td>1.22%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>5.78%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.86%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Table A.21: Location of analytical CIs relative to benchmark simulation, D3

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>14.76%</td>
<td>29.94%</td>
<td>28.98%</td>
<td>44.72%</td>
<td>80.58%</td>
<td>88.34%</td>
<td>88.34%</td>
<td>88.34%</td>
<td>88.34%</td>
<td>88.34%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>30.96%</td>
<td>42.34%</td>
<td>37.74%</td>
<td>38.76%</td>
<td>9.06%</td>
<td>23.04%</td>
<td>30.52%</td>
<td>17.34%</td>
<td>10.10%</td>
<td>4.22%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>35.22%</td>
<td>30.14%</td>
<td>32.34%</td>
<td>23.86%</td>
<td>8.66%</td>
<td>11.56%</td>
<td>19.40%</td>
<td>12.48%</td>
<td>7.30%</td>
<td>3.98%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.34%</td>
<td>0.64%</td>
<td>0.34%</td>
<td>0.84%</td>
<td>0.00%</td>
<td>0.38%</td>
<td>0.68%</td>
<td>0.64%</td>
<td>1.20%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.20%</td>
<td>0.24%</td>
<td>0.36%</td>
<td>0.32%</td>
<td>0.86%</td>
<td>0.00%</td>
<td>0.34%</td>
<td>0.42%</td>
<td>0.88%</td>
<td>1.18%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>3.24%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.92%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>15.68%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>30.14%</td>
<td>0.00%</td>
<td>0.32%</td>
<td>0.02%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Table A.22: Location of analytical CIs relative to benchmark simulation, N1

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>0.38%</td>
<td>21.28%</td>
<td>0.00%</td>
<td>0.56%</td>
<td>64.46%</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>29.40%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>15.38%</td>
<td>46.94%</td>
<td>1.74%</td>
<td>41.02%</td>
<td>15.42%</td>
<td>0.26%</td>
<td>12.18%</td>
<td>0.86%</td>
<td>8.32%</td>
<td>24.80%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>30.62%</td>
<td>31.68%</td>
<td>8.98%</td>
<td>46.74%</td>
<td>19.72%</td>
<td>0.00%</td>
<td>22.67%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>45.64%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.24%</td>
<td>0.00%</td>
<td>0.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.16%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.16%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>37.62%</td>
<td>0.00%</td>
<td>0.07%</td>
<td>89.60%</td>
<td>11.68%</td>
<td>0.00%</td>
<td>0.28%</td>
<td>51.02%</td>
<td>0.96%</td>
<td>18.00%</td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>16.00%</td>
<td>0.00%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.46%</td>
<td>14.08%</td>
<td>98.14%</td>
<td>73.66%</td>
<td></td>
</tr>
</tbody>
</table>

Table A.23: Location of analytical CIs relative to benchmark simulation, N2

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>1.30%</td>
<td>21.06%</td>
<td>0.00%</td>
<td>0.52%</td>
<td>64.34%</td>
<td>0.00%</td>
<td>2.46%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>49.78%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>35.50%</td>
<td>47.12%</td>
<td>19.78%</td>
<td>40.72%</td>
<td>15.44%</td>
<td>0.60%</td>
<td>59.82%</td>
<td>1.78%</td>
<td>26.48%</td>
<td>25.70%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>29.10%</td>
<td>31.72%</td>
<td>6.82%</td>
<td>46.42%</td>
<td>19.82%</td>
<td>0.00%</td>
<td>13.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>24.36%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.24%</td>
<td>0.00%</td>
<td>0.06%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.16%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>17.82%</td>
<td>0.00%</td>
<td>73.14%</td>
<td>12.34%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>11.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>16.28%</td>
<td>0.00%</td>
<td>0.12%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.40%</td>
<td>13.70%</td>
<td>98.14%</td>
<td>73.50%</td>
<td></td>
</tr>
</tbody>
</table>

Table A.24: Location of analytical CIs relative to benchmark simulation, S1

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>0.56%</td>
<td>21.62%</td>
<td>0.00%</td>
<td>0.60%</td>
<td>64.46%</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>29.40%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>16.36%</td>
<td>46.64%</td>
<td>1.74%</td>
<td>41.02%</td>
<td>15.42%</td>
<td>0.26%</td>
<td>12.26%</td>
<td>0.86%</td>
<td>8.32%</td>
<td>24.80%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>33.16%</td>
<td>31.64%</td>
<td>9.68%</td>
<td>46.74%</td>
<td>19.72%</td>
<td>0.00%</td>
<td>22.54%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>45.64%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.24%</td>
<td>0.00%</td>
<td>0.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.16%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.16%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>35.28%</td>
<td>0.00%</td>
<td>88.90%</td>
<td>11.64%</td>
<td>0.00%</td>
<td>0.34%</td>
<td>50.84%</td>
<td>0.98%</td>
<td>18.04%</td>
<td></td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>14.64%</td>
<td>0.00%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.38%</td>
<td>13.92%</td>
<td>98.14%</td>
<td>73.62%</td>
<td></td>
</tr>
</tbody>
</table>

Table A.25: Location of analytical CIs relative to benchmark simulation, S2

<table>
<thead>
<tr>
<th>Category</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
<th>Hosp 1</th>
<th>Hosp 2</th>
<th>Hosp 3</th>
<th>Hosp 4</th>
<th>Hosp 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contained within benchmark (1)</td>
<td>2.76%</td>
<td>25.72%</td>
<td>0.00%</td>
<td>0.72%</td>
<td>64.84%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>29.42%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>31.86%</td>
<td>40.58%</td>
<td>8.44%</td>
<td>47.40%</td>
<td>15.92%</td>
<td>33.50%</td>
<td>37.54%</td>
<td>24.70%</td>
<td>41.86%</td>
<td></td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>36.32%</td>
<td>33.66%</td>
<td>12.42%</td>
<td>41.16%</td>
<td>18.80%</td>
<td>0.00%</td>
<td>17.86%</td>
<td>0.00%</td>
<td>32.14%</td>
<td></td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.20%</td>
<td>0.00%</td>
<td>0.24%</td>
<td>0.00%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.24%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>29.06%</td>
<td>0.04%</td>
<td>91.40%</td>
<td>19.78%</td>
<td>0.00%</td>
<td>18.16%</td>
<td>50.84%</td>
<td>49.40%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>0.06%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>48.34%</td>
<td>0.00%</td>
<td>25.88%</td>
<td>3.64%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table A.26: Location of analytical CIs relative to benchmark simulation, S3
Table A.27: Location of analytical CIs relative to benchmark simulation, S4

<table>
<thead>
<tr>
<th>Category</th>
<th>S4, low</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hoop 1</td>
<td>Hoop 2</td>
<td>Hoop 3</td>
<td>Hoop 4</td>
<td>Hoop 5</td>
<td>Hoop 1</td>
<td>Hoop 2</td>
<td>Hoop 3</td>
<td>Hoop 4</td>
<td>Hoop 5</td>
</tr>
<tr>
<td>Contained within benchmark (1)</td>
<td>0.36%</td>
<td>9.94%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>60.14%</td>
<td>7.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.12%</td>
<td>33.66%</td>
</tr>
<tr>
<td>Left of benchmark (2)</td>
<td>27.60%</td>
<td>57.72%</td>
<td>13.96%</td>
<td>48.54%</td>
<td>19.36%</td>
<td>9.84%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.14%</td>
<td>1.62%</td>
</tr>
<tr>
<td>Right of benchmark (3)</td>
<td>4.54%</td>
<td>31.18%</td>
<td>0.00%</td>
<td>20.26%</td>
<td>20.16%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>18.24%</td>
</tr>
<tr>
<td>Below benchmark (4)</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.24%</td>
<td>0.00%</td>
<td>0.24%</td>
<td>0.02%</td>
<td>0.10%</td>
<td>0.04%</td>
<td>0.02%</td>
<td>0.24%</td>
</tr>
<tr>
<td>Above benchmark (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.10%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Covers benchmark (6)</td>
<td>26.14%</td>
<td>1.14%</td>
<td>85.20%</td>
<td>31.16%</td>
<td>0.00%</td>
<td>0.12%</td>
<td>33.66%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>12.20%</td>
</tr>
<tr>
<td>Solution fail (7)</td>
<td>41.36%</td>
<td>0.00%</td>
<td>0.76%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>88.96%</td>
<td>16.66%</td>
<td>97.84%</td>
<td>76.44%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

137
A.4 Kernel density plots of the location of analytical CIs

This section contains the location of analytical CIs from resampled data for the 5 sample hospitals. Results for all 5000 resamples are summarised in kernel density curves for both the low-variance (MDC) and high-variance (DRG) scenarios.
Table A.28: Kernel density plots of the location of analytical CI bounds: D1

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image1" alt="Kernel Density" /></td>
<td><img src="image2" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image3" alt="Kernel Density" /></td>
<td><img src="image4" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image5" alt="Kernel Density" /></td>
<td><img src="image6" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image7" alt="Kernel Density" /></td>
<td><img src="image8" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image9" alt="Kernel Density" /></td>
<td><img src="image10" alt="Kernel Density" /></td>
</tr>
</tbody>
</table>

To read: horizontal axis is log-scaled. Vertical lines cutting across kernel density curves are the simulated benchmarks. The solid curve represents the kernel density of the lower bound and the dashed line for the upper bound. A method is considered more accurate if the vertical benchmark lines divide through the curve where the area underneath is approximately separated into two halves.
Table A.29: Kernel density plots of the location of analytical CI bounds: D2

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image1" alt="Kernel density plot for Hosp1" /></td>
<td><img src="image2" alt="Kernel density plot for Hosp1" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image3" alt="Kernel density plot for Hosp2" /></td>
<td><img src="image4" alt="Kernel density plot for Hosp2" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image5" alt="Kernel density plot for Hosp3" /></td>
<td><img src="image6" alt="Kernel density plot for Hosp3" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image7" alt="Kernel density plot for Hosp4" /></td>
<td><img src="image8" alt="Kernel density plot for Hosp4" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image9" alt="Kernel density plot for Hosp5" /></td>
<td><img src="image10" alt="Kernel density plot for Hosp5" /></td>
</tr>
</tbody>
</table>
Table A.30: Kernel density plots of the location of analytical CI bounds: D3

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosp2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosp3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosp4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hosp5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table A.31: Kernel density plots of the location of analytical CI bounds: N1

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image1" alt="Kernel density plot" /></td>
<td><img src="image2" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image3" alt="Kernel density plot" /></td>
<td><img src="image4" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image5" alt="Kernel density plot" /></td>
<td><img src="image6" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image7" alt="Kernel density plot" /></td>
<td><img src="image8" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image9" alt="Kernel density plot" /></td>
<td><img src="image10" alt="Kernel density plot" /></td>
</tr>
</tbody>
</table>
Table A.32: Kernel density plots of the location of analytical CI bounds: N2

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image1" alt="Kernel density plot for Hosp1" /></td>
<td><img src="image2" alt="Kernel density plot for Hosp1" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image3" alt="Kernel density plot for Hosp2" /></td>
<td><img src="image4" alt="Kernel density plot for Hosp2" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image5" alt="Kernel density plot for Hosp3" /></td>
<td><img src="image6" alt="Kernel density plot for Hosp3" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image7" alt="Kernel density plot for Hosp4" /></td>
<td><img src="image8" alt="Kernel density plot for Hosp4" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image9" alt="Kernel density plot for Hosp5" /></td>
<td><img src="image10" alt="Kernel density plot for Hosp5" /></td>
</tr>
</tbody>
</table>
Table A.33: Kernel density plots of the location of analytical CI bounds: S1

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image" alt="Kernel density plot for Hosp1 Low variance" /></td>
<td><img src="image" alt="Kernel density plot for Hosp1 High variance" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image" alt="Kernel density plot for Hosp2 Low variance" /></td>
<td><img src="image" alt="Kernel density plot for Hosp2 High variance" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image" alt="Kernel density plot for Hosp3 Low variance" /></td>
<td><img src="image" alt="Kernel density plot for Hosp3 High variance" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image" alt="Kernel density plot for Hosp4 Low variance" /></td>
<td><img src="image" alt="Kernel density plot for Hosp4 High variance" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image" alt="Kernel density plot for Hosp5 Low variance" /></td>
<td><img src="image" alt="Kernel density plot for Hosp5 High variance" /></td>
</tr>
</tbody>
</table>
Table A.34: Kernel density plots of the location of analytical CI bounds: S2

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image" alt="Kernel density plot" /></td>
<td><img src="image" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image" alt="Kernel density plot" /></td>
<td><img src="image" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image" alt="Kernel density plot" /></td>
<td><img src="image" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image" alt="Kernel density plot" /></td>
<td><img src="image" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image" alt="Kernel density plot" /></td>
<td><img src="image" alt="Kernel density plot" /></td>
</tr>
</tbody>
</table>
Table A.35: Kernel density plots of the location of analytical CI bounds: S3

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image1" alt="Kernel Density" /></td>
<td><img src="image2" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image3" alt="Kernel Density" /></td>
<td><img src="image4" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image5" alt="Kernel Density" /></td>
<td><img src="image6" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image7" alt="Kernel Density" /></td>
<td><img src="image8" alt="Kernel Density" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image9" alt="Kernel Density" /></td>
<td><img src="image10" alt="Kernel Density" /></td>
</tr>
</tbody>
</table>
Table A.36: Kernel density plots of the location of analytical CI bounds: S4

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Low variance (MDC)</th>
<th>High variance (DRG)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosp1</td>
<td><img src="image1" alt="Kernel density plot" /></td>
<td><img src="image2" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp2</td>
<td><img src="image3" alt="Kernel density plot" /></td>
<td><img src="image4" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp3</td>
<td><img src="image5" alt="Kernel density plot" /></td>
<td><img src="image6" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp4</td>
<td><img src="image7" alt="Kernel density plot" /></td>
<td><img src="image8" alt="Kernel density plot" /></td>
</tr>
<tr>
<td>Hosp5</td>
<td><img src="image9" alt="Kernel density plot" /></td>
<td><img src="image10" alt="Kernel density plot" /></td>
</tr>
</tbody>
</table>
A.5 Comparison of CI positioning between analytic methods using observed hospital data

This section contains tables of results from comparisons between the locations of analytical CIs against each other. The 1123 observed hospitals are used rather than simulated ones, see main text for explanation. Categories 1 to 6 are defined as before. If at least one of the CIs failed to produce a solution, that pair is placed in category 7.

Table A.37: Pairwise comparison between analytic methods, D1, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 contained within (1)</td>
<td>90.29%</td>
<td>2.14%</td>
<td>82.10%</td>
<td>78.90%</td>
<td>82.28%</td>
<td>0.00%</td>
<td>88.16%</td>
<td>72.66%</td>
<td></td>
</tr>
<tr>
<td>D1 left of (2)</td>
<td>0.00%</td>
<td>0.18%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D1 right of (3)</td>
<td>9.71%</td>
<td>74.53%</td>
<td>6.23%</td>
<td>8.90%</td>
<td>6.23%</td>
<td>0.00%</td>
<td>11.84%</td>
<td>9.17%</td>
<td></td>
</tr>
<tr>
<td>D1 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D1 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D1 covers (6)</td>
<td>0.00%</td>
<td>14.07%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.91%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>0.00%</td>
<td>9.08%</td>
<td>11.67%</td>
<td>12.20%</td>
<td>11.49%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>18.17%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

The CI of each hospital is computed using each of the analytical methods. The location of their lower and upper bounds are compared against each other. Each percentage represents the position of the analytic method being examined, in this case D1, relative to the other methods. Categories add up to 100%. Read down the first column, then across. For example, the value 82.10% means that for 922 out of 1123 hospitals in the dataset, the bounds of the CI computed using method D1 are completely contained within the CI computed using N2.

Table A.38: Pairwise comparison between analytic methods, D2, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2 contained within (1)</td>
<td>0.00%</td>
<td>- 0.45%</td>
<td>81.83%</td>
<td>77.83%</td>
<td>81.92%</td>
<td>0.00%</td>
<td>87.53%</td>
<td>71.24%</td>
<td></td>
</tr>
<tr>
<td>D2 left of (2)</td>
<td>9.71%</td>
<td>- 0.62%</td>
<td>6.50%</td>
<td>9.97%</td>
<td>6.59%</td>
<td>0.00%</td>
<td>3.92%</td>
<td>10.60%</td>
<td></td>
</tr>
<tr>
<td>D2 right of (3)</td>
<td>0.00%</td>
<td>- 42.48%</td>
<td>5.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D2 below (4)</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D2 above (5)</td>
<td>90.29%</td>
<td>- 47.37%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>8.55%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>0.00%</td>
<td>- 9.08%</td>
<td>11.67%</td>
<td>12.20%</td>
<td>11.49%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>18.17%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Table A.39: Pairwise comparison between analytic methods, D3, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3 contained within (1)</td>
<td>14.07%</td>
<td>47.37%</td>
<td>- 88.25%</td>
<td>87.71%</td>
<td>88.42%</td>
<td>0.00%</td>
<td>82.19%</td>
<td>81.83%</td>
<td></td>
</tr>
<tr>
<td>D3 left of (2)</td>
<td>74.53%</td>
<td>42.48%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.73%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D3 right of (3)</td>
<td>0.18%</td>
<td>0.62%</td>
<td>- 0.09%</td>
<td>0.09%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D3 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>D3 above (5)</td>
<td>2.14%</td>
<td>0.45%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>90.92%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>9.08%</td>
<td>9.08%</td>
<td>- 11.67%</td>
<td>12.20%</td>
<td>11.49%</td>
<td>9.08%</td>
<td>9.08%</td>
<td>18.17%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>
Table A.40: Pairwise comparison between analytic methods, N1, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>52.18%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>17.36%</td>
</tr>
<tr>
<td>N1 left of (2)</td>
<td>6.23%</td>
<td>6.50%</td>
<td>0.09%</td>
<td>-</td>
<td>0.18%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>12.02%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N1 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>31.79%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>44.97%</td>
<td>64.38%</td>
</tr>
<tr>
<td>N1 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N1 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N1 covers (6)</td>
<td>82.10%</td>
<td>81.83%</td>
<td>88.25%</td>
<td>-</td>
<td>3.38%</td>
<td>88.33%</td>
<td>88.33%</td>
<td>31.34%</td>
<td>31.34%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>11.67%</td>
<td>11.67%</td>
<td>11.67%</td>
<td>-</td>
<td>12.47%</td>
<td>11.67%</td>
<td>11.67%</td>
<td>11.67%</td>
<td>18.25%</td>
</tr>
</tbody>
</table>

Table A.41: Pairwise comparison between analytic methods, N2, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>3.38%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>6.59%</td>
<td>19.86%</td>
</tr>
<tr>
<td>N2 left of (2)</td>
<td>8.90%</td>
<td>9.97%</td>
<td>0.09%</td>
<td>31.79%</td>
<td>-</td>
<td>34.37%</td>
<td>0.00%</td>
<td>18.97%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N2 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.18%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>35.00%</td>
<td>61.98%</td>
</tr>
<tr>
<td>N2 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N2 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N2 covers (6)</td>
<td>78.90%</td>
<td>77.83%</td>
<td>87.71%</td>
<td>52.18%</td>
<td>-</td>
<td>53.34%</td>
<td>87.80%</td>
<td>27.25%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>12.20%</td>
<td>12.20%</td>
<td>12.20%</td>
<td>12.47%</td>
<td>-</td>
<td>12.29%</td>
<td>12.20%</td>
<td>12.20%</td>
<td>18.17%</td>
</tr>
</tbody>
</table>

Table A.42: Pairwise comparison between analytic methods, S1, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>88.33%</td>
<td>53.34%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>17.54%</td>
</tr>
<tr>
<td>S1 left of (2)</td>
<td>6.23%</td>
<td>6.59%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>12.02%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S1 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>34.37%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>45.41%</td>
<td>64.29%</td>
</tr>
<tr>
<td>S1 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S1 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S1 covers (6)</td>
<td>82.28%</td>
<td>81.92%</td>
<td>88.42%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>88.51%</td>
<td>31.08%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>11.49%</td>
<td>11.49%</td>
<td>11.49%</td>
<td>12.47%</td>
<td>-</td>
<td>11.49%</td>
<td>11.49%</td>
<td>11.49%</td>
<td>18.17%</td>
</tr>
</tbody>
</table>

Table A.43: Pairwise comparison between analytic methods, S2, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2 contained within (1)</td>
<td>99.91%</td>
<td>100.00%</td>
<td>90.92%</td>
<td>88.33%</td>
<td>87.80%</td>
<td>88.51%</td>
<td>-</td>
<td>100.00%</td>
<td>81.83%</td>
</tr>
<tr>
<td>S2 left of (2)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 covers (6)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>9.08%</td>
<td>11.67%</td>
<td>12.20%</td>
<td>11.49%</td>
<td>-</td>
<td>0.00%</td>
<td>18.17%</td>
</tr>
</tbody>
</table>

Table A.44: Pairwise comparison between analytic methods, S3, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3 contained within (1)</td>
<td>0.00%</td>
<td>8.55%</td>
<td>0.00%</td>
<td>31.34%</td>
<td>27.25%</td>
<td>31.08%</td>
<td>0.00%</td>
<td>-</td>
<td>23.24%</td>
</tr>
<tr>
<td>S3 left of (2)</td>
<td>11.84%</td>
<td>3.92%</td>
<td>0.00%</td>
<td>44.97%</td>
<td>35.00%</td>
<td>45.41%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
</tr>
<tr>
<td>S3 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.73%</td>
<td>12.02%</td>
<td>18.97%</td>
<td>12.02%</td>
<td>0.00%</td>
<td>-</td>
<td>58.59%</td>
</tr>
<tr>
<td>S3 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S3 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S3 covers (6)</td>
<td>88.16%</td>
<td>87.53%</td>
<td>82.19%</td>
<td>0.00%</td>
<td>6.59%</td>
<td>0.00%</td>
<td>100.00%</td>
<td>-</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>9.08%</td>
<td>11.67%</td>
<td>12.20%</td>
<td>11.49%</td>
<td>0.00%</td>
<td>-</td>
<td>18.17%</td>
</tr>
</tbody>
</table>
### Table A.45: Pairwise comparison between analytic methods, S4, low-variance (MDC) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S4 left of (2)</td>
<td>9.17%</td>
<td>10.60%</td>
<td>0.00%</td>
<td>64.38%</td>
<td>61.98%</td>
<td>64.29%</td>
<td>0.00%</td>
<td>58.59%</td>
<td>-</td>
</tr>
<tr>
<td>S4 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
</tr>
<tr>
<td>S4 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
</tr>
<tr>
<td>S4 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
</tr>
<tr>
<td>S4 covers (6)</td>
<td>72.66%</td>
<td>71.24%</td>
<td>81.83%</td>
<td>17.36%</td>
<td>17.54%</td>
<td>23.24%</td>
<td>18.17%</td>
<td>18.17%</td>
<td>-</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>18.17%</td>
<td>18.17%</td>
<td>18.17%</td>
<td>18.25%</td>
<td>18.17%</td>
<td>18.17%</td>
<td>18.17%</td>
<td>18.17%</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table A.46: Pairwise comparison between analytic methods, D1, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1 contained within (1)</td>
<td>- 90.29%</td>
<td>0.36%</td>
<td>64.92%</td>
<td>64.29%</td>
<td>64.92%</td>
<td>0.18%</td>
<td>95.81%</td>
<td>58.95%</td>
<td></td>
</tr>
<tr>
<td>D1 left of (2)</td>
<td>- 0.00%</td>
<td>0.89%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>D1 right of (3)</td>
<td>- 9.71%</td>
<td>47.55%</td>
<td>0.09%</td>
<td>1.16%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>2.58%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>D1 below (4)</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>D1 above (5)</td>
<td>- 0.00%</td>
<td>42.03%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.82%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>No solution (7)</td>
<td>- 0.00%</td>
<td>9.17%</td>
<td>35.00%</td>
<td>34.55%</td>
<td>35.00%</td>
<td>0.00%</td>
<td>1.60%</td>
<td>40.87%</td>
<td></td>
</tr>
</tbody>
</table>

### Table A.47: Pairwise comparison between analytic methods, D2, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D2 contained within (1)</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>64.92%</td>
<td>64.20%</td>
<td>64.92%</td>
<td>0.00%</td>
<td>95.99%</td>
<td>58.95%</td>
<td></td>
</tr>
<tr>
<td>D2 left of (2)</td>
<td>9.71%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>D2 right of (3)</td>
<td>0.00%</td>
<td>- 20.48%</td>
<td>0.09%</td>
<td>1.25%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>1.25%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>D2 below (4)</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>D2 above (5)</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>D2 covers (6)</td>
<td>90.29%</td>
<td>- 70.35%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>99.91%</td>
<td>1.16%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>No solution (7)</td>
<td>0.00%</td>
<td>- 9.17%</td>
<td>35.00%</td>
<td>34.55%</td>
<td>35.00%</td>
<td>0.00%</td>
<td>1.60%</td>
<td>40.87%</td>
<td></td>
</tr>
</tbody>
</table>

### Table A.48: Pairwise comparison between analytic methods, D3, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D3 contained within (1)</td>
<td>42.03%</td>
<td>70.35%</td>
<td>- 64.92%</td>
<td>65.45%</td>
<td>64.92%</td>
<td>0.27%</td>
<td>90.03%</td>
<td>58.95%</td>
<td></td>
</tr>
<tr>
<td>D3 left of (2)</td>
<td>47.55%</td>
<td>20.48%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>D3 right of (3)</td>
<td>0.89%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>D3 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>D3 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.09%</td>
<td></td>
</tr>
<tr>
<td>D3 covers (6)</td>
<td>0.36%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>90.20%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>No solution (7)</td>
<td>9.17%</td>
<td>9.17%</td>
<td>- 35.08%</td>
<td>34.55%</td>
<td>35.08%</td>
<td>9.17%</td>
<td>9.53%</td>
<td>40.96%</td>
<td></td>
</tr>
</tbody>
</table>

### Table A.49: Pairwise comparison between analytic methods, N1, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N1 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 7.66%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.36%</td>
<td></td>
</tr>
<tr>
<td>N1 left of (2)</td>
<td>0.09%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.09%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>N1 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 57.17%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>62.96%</td>
<td>58.95%</td>
<td></td>
</tr>
<tr>
<td>N1 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>N1 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>- 0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>N1 covers (6)</td>
<td>64.92%</td>
<td>64.92%</td>
<td>64.92%</td>
<td>- 0.00%</td>
<td>65.00%</td>
<td>65.00%</td>
<td>1.96%</td>
<td>0.00%</td>
<td></td>
</tr>
<tr>
<td>No solution (7)</td>
<td>35.00%</td>
<td>35.00%</td>
<td>35.08%</td>
<td>- 35.17%</td>
<td>35.00%</td>
<td>35.00%</td>
<td>35.00%</td>
<td>41.05%</td>
<td></td>
</tr>
</tbody>
</table>
Table A.50: Pairwise comparison between analytic methods, N2, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>N2 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>37.04%</td>
<td>28.23%</td>
</tr>
<tr>
<td>N2 left of (2)</td>
<td>1.16%</td>
<td>1.25%</td>
<td>0.00%</td>
<td>57.17%</td>
<td>-</td>
<td>57.17%</td>
<td>0.00%</td>
<td>3.56%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N2 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>24.22%</td>
<td>30.72%</td>
</tr>
<tr>
<td>N2 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N2 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>N2 covers (6)</td>
<td>64.29%</td>
<td>64.20%</td>
<td>65.45%</td>
<td>7.66%</td>
<td>-</td>
<td>65.45%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>34.55%</td>
<td>34.55%</td>
<td>34.55%</td>
<td>35.17%</td>
<td>-</td>
<td>34.55%</td>
<td>34.55%</td>
<td>41.05%</td>
<td></td>
</tr>
</tbody>
</table>

Table A.51: Pairwise comparison between analytic methods, S1, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1 contained within (1)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>65.00%</td>
<td>7.66%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.36%</td>
</tr>
<tr>
<td>S1 left of (2)</td>
<td>0.09%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.09%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S1 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>57.17%</td>
<td>-</td>
<td>0.00%</td>
<td>62.96%</td>
<td>58.59%</td>
</tr>
<tr>
<td>S1 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S1 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S1 covers (6)</td>
<td>64.92%</td>
<td>64.92%</td>
<td>64.92%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>65.00%</td>
<td>1.96%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>35.00%</td>
<td>35.00%</td>
<td>35.08%</td>
<td>35.00%</td>
<td>35.00%</td>
<td>-</td>
<td>35.00%</td>
<td>35.00%</td>
<td>41.05%</td>
</tr>
</tbody>
</table>

Table A.52: Pairwise comparison between analytic methods, S2, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2 contained within (1)</td>
<td>99.82%</td>
<td>99.91%</td>
<td>90.20%</td>
<td>65.00%</td>
<td>65.45%</td>
<td>65.00%</td>
<td>-</td>
<td>98.40%</td>
<td>58.95%</td>
</tr>
<tr>
<td>S2 left of (2)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.36%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.18%</td>
</tr>
<tr>
<td>S2 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>S2 covers (6)</td>
<td>0.18%</td>
<td>0.00%</td>
<td>0.27%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>9.17%</td>
<td>35.00%</td>
<td>35.00%</td>
<td>-</td>
<td>35.00%</td>
<td>35.00%</td>
<td>40.87%</td>
</tr>
</tbody>
</table>

Table A.53: Pairwise comparison between analytic methods, S3, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S3 contained within (1)</td>
<td>0.00%</td>
<td>1.16%</td>
<td>0.00%</td>
<td>1.96%</td>
<td>0.62%</td>
<td>1.96%</td>
<td>0.00%</td>
<td>-</td>
<td>9.44%</td>
</tr>
<tr>
<td>S3 left of (2)</td>
<td>2.58%</td>
<td>1.25%</td>
<td>0.00%</td>
<td>62.96%</td>
<td>24.22%</td>
<td>62.96%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
</tr>
<tr>
<td>S3 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.45%</td>
<td>0.00%</td>
<td>3.56%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>-</td>
<td>49.51%</td>
</tr>
<tr>
<td>S3 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
</tr>
<tr>
<td>S3 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>0.00%</td>
</tr>
<tr>
<td>S3 covers (6)</td>
<td>95.81%</td>
<td>95.99%</td>
<td>90.03%</td>
<td>0.00%</td>
<td>37.04%</td>
<td>0.00%</td>
<td>98.40%</td>
<td>-</td>
<td>0.00%</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>1.60%</td>
<td>1.60%</td>
<td>9.53%</td>
<td>35.00%</td>
<td>35.05%</td>
<td>-</td>
<td>35.00%</td>
<td>35.00%</td>
<td>41.05%</td>
</tr>
</tbody>
</table>

Table A.54: Pairwise comparison between analytic methods, S4, low-variance (DRG) case

<table>
<thead>
<tr>
<th>Compared against</th>
<th>D1</th>
<th>D2</th>
<th>D3</th>
<th>N1</th>
<th>N2</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4 contained within (1)</td>
<td>0.09%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S4 left of (2)</td>
<td>0.09%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>58.59%</td>
<td>30.72%</td>
<td>58.59%</td>
<td>0.18%</td>
<td>49.51%</td>
<td>-</td>
</tr>
<tr>
<td>S4 right of (3)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S4 below (4)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S4 above (5)</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>S4 covers (6)</td>
<td>58.95%</td>
<td>58.95%</td>
<td>58.95%</td>
<td>0.36%</td>
<td>28.23%</td>
<td>0.36%</td>
<td>58.95%</td>
<td>9.44%</td>
<td>-</td>
</tr>
<tr>
<td>No solution (7)</td>
<td>40.87%</td>
<td>40.87%</td>
<td>40.96%</td>
<td>41.05%</td>
<td>41.05%</td>
<td>41.05%</td>
<td>40.87%</td>
<td>41.05%</td>
<td>-</td>
</tr>
</tbody>
</table>
Appendix B

Appendix to Chapter 3

This section contains tables listing raw parameter estimates used to compute elasticities and other statistics from the translog cost functions in chapter 3. The column on the left is the point estimate and the column on the right is the standard error.
Table B.1: Parameter estimates without bootstrapping (N = 214)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Both (A4)</th>
<th>Mortality only (A2)</th>
<th>Readmissions only (A3)</th>
<th>(no quality) (A1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point estimate</td>
<td>Standard Error</td>
<td>Point estimate</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$\alpha_k$</td>
<td>0.273598</td>
<td>0.009180</td>
<td>0.273667</td>
<td>0.009158</td>
</tr>
<tr>
<td>$\alpha_l$</td>
<td>0.371844</td>
<td>0.009390</td>
<td>0.371793</td>
<td>0.009378</td>
</tr>
<tr>
<td>$\alpha_m$</td>
<td>0.354558</td>
<td>0.005100</td>
<td>0.354540</td>
<td>0.005097</td>
</tr>
<tr>
<td>$\alpha_t$</td>
<td>0.992668</td>
<td>0.010154</td>
<td>0.993140</td>
<td>0.009902</td>
</tr>
<tr>
<td>$\beta_{kk}$</td>
<td>0.099592</td>
<td>0.001024</td>
<td>0.099556</td>
<td>0.001023</td>
</tr>
<tr>
<td>$\beta_{kl}$</td>
<td>-0.083115</td>
<td>0.001842</td>
<td>-0.083084</td>
<td>0.001840</td>
</tr>
<tr>
<td>$\beta_{km}$</td>
<td>-0.016477</td>
<td>0.000546</td>
<td>-0.016472</td>
<td>0.000546</td>
</tr>
<tr>
<td>$\beta_{kt}$</td>
<td>0.003539</td>
<td>0.001154</td>
<td>0.003520</td>
<td>0.001149</td>
</tr>
<tr>
<td>$\beta_{ll}$</td>
<td>0.211639</td>
<td>0.001236</td>
<td>0.211609</td>
<td>0.001235</td>
</tr>
<tr>
<td>$\beta_{lm}$</td>
<td>-0.128524</td>
<td>0.001761</td>
<td>-0.128525</td>
<td>0.001760</td>
</tr>
<tr>
<td>$\beta_{lt}$</td>
<td>-0.002835</td>
<td>0.001040</td>
<td>-0.002820</td>
<td>0.001036</td>
</tr>
<tr>
<td>$\beta_{mm}$</td>
<td>0.145000</td>
<td>0.000846</td>
<td>0.144996</td>
<td>0.000846</td>
</tr>
<tr>
<td>$\beta_{mt}$</td>
<td>-0.000704</td>
<td>0.000355</td>
<td>-0.000701</td>
<td>0.000354</td>
</tr>
<tr>
<td>$\beta_{tt}$</td>
<td>0.001746</td>
<td>0.000556</td>
<td>0.001694</td>
<td>0.000541</td>
</tr>
<tr>
<td>$\alpha_d$</td>
<td>0.003367</td>
<td>0.003334</td>
<td>0.003346</td>
<td>0.003331</td>
</tr>
<tr>
<td>$\alpha_r$</td>
<td>-0.000328</td>
<td>0.001681</td>
<td>-0.000265</td>
<td>0.001683</td>
</tr>
<tr>
<td>(cons)</td>
<td>1.128470</td>
<td>0.049306</td>
<td>1.126336</td>
<td>0.048308</td>
</tr>
<tr>
<td>Variable</td>
<td>Both (B3)</td>
<td>Mortality only (B1)</td>
<td>Readmissions only (B2)</td>
<td></td>
</tr>
<tr>
<td>----------</td>
<td>-----------</td>
<td>---------------------</td>
<td>------------------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Point</td>
<td>Standard Error</td>
<td>Point</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$\alpha_k$</td>
<td>0.276485</td>
<td>0.009581</td>
<td>0.275721</td>
<td>0.009311</td>
</tr>
<tr>
<td>$\alpha_l$</td>
<td>0.368829</td>
<td>0.009721</td>
<td>0.369729</td>
<td>0.009497</td>
</tr>
<tr>
<td>$\alpha_m$</td>
<td>0.354709</td>
<td>0.005138</td>
<td>0.354542</td>
<td>0.005119</td>
</tr>
<tr>
<td>$\alpha_r$</td>
<td>0.985956</td>
<td>0.011358</td>
<td>0.987051</td>
<td>0.010459</td>
</tr>
<tr>
<td>$\beta_{kk}$</td>
<td>0.049603</td>
<td>0.001044</td>
<td>0.049643</td>
<td>0.001025</td>
</tr>
<tr>
<td>$\beta_{kl}$</td>
<td>-0.082800</td>
<td>0.001879</td>
<td>-0.082844</td>
<td>0.001841</td>
</tr>
<tr>
<td>$\beta_{km}$</td>
<td>-0.016401</td>
<td>0.000549</td>
<td>-0.016437</td>
<td>0.000548</td>
</tr>
<tr>
<td>$\beta_{kt}$</td>
<td>0.003187</td>
<td>0.001196</td>
<td>0.003290</td>
<td>0.001162</td>
</tr>
<tr>
<td>$\beta_{ll}$</td>
<td>0.106012</td>
<td>0.001249</td>
<td>0.105912</td>
<td>0.001237</td>
</tr>
<tr>
<td>$\beta_{lm}$</td>
<td>-0.129230</td>
<td>0.001769</td>
<td>-0.128972</td>
<td>0.001764</td>
</tr>
<tr>
<td>$\beta_{lt}$</td>
<td>-0.002572</td>
<td>0.001082</td>
<td>-0.002626</td>
<td>0.001049</td>
</tr>
<tr>
<td>$\beta_{mm}$</td>
<td>0.072816</td>
<td>0.000850</td>
<td>0.072703</td>
<td>0.000847</td>
</tr>
<tr>
<td>$\beta_{mt}$</td>
<td>-0.000619</td>
<td>0.000362</td>
<td>-0.000638</td>
<td>0.000356</td>
</tr>
<tr>
<td>$\beta_{tt}$</td>
<td>0.001148</td>
<td>0.000605</td>
<td>0.001106</td>
<td>0.000564</td>
</tr>
<tr>
<td>$\alpha_d$</td>
<td>-0.001602</td>
<td>0.003376</td>
<td>-0.003057</td>
<td>0.002354</td>
</tr>
<tr>
<td>$\alpha_r$</td>
<td>-0.001920</td>
<td>0.003606</td>
<td></td>
<td>-0.000455</td>
</tr>
<tr>
<td>(cons)</td>
<td>1.155028</td>
<td>0.053482</td>
<td>1.150111</td>
<td>0.049540</td>
</tr>
</tbody>
</table>
Table B.3: Parameter estimates when using multiple outputs

<table>
<thead>
<tr>
<th>Variable</th>
<th>Point estimate</th>
<th>Standard Error</th>
<th>Variable</th>
<th>Point estimate</th>
<th>Standard Error</th>
<th>Variable</th>
<th>Point estimate</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>One output (A1)</td>
<td></td>
<td></td>
<td>Two outputs (C2)</td>
<td></td>
<td></td>
<td>Three outputs (C3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \alpha_k )</td>
<td>0.273598</td>
<td>0.009180</td>
<td>( \alpha_k )</td>
<td>0.250740</td>
<td>0.011567</td>
<td>( \alpha_k )</td>
<td>0.279348</td>
<td>0.018822</td>
</tr>
<tr>
<td>( \alpha_l )</td>
<td>0.354558</td>
<td>0.005100</td>
<td>( \alpha_l )</td>
<td>0.358556</td>
<td>0.005547</td>
<td>( \alpha_l )</td>
<td>0.348343</td>
<td>0.006296</td>
</tr>
<tr>
<td>( \alpha_m )</td>
<td>0.992668</td>
<td>0.010154</td>
<td>( \alpha_m )</td>
<td>0.048984</td>
<td>0.001136</td>
<td>( \alpha_m )</td>
<td>0.049426</td>
<td>0.001514</td>
</tr>
<tr>
<td>( \alpha_t )</td>
<td>0.992668</td>
<td>0.010154</td>
<td>( \alpha_t )</td>
<td>-0.081713</td>
<td>0.002028</td>
<td>( \alpha_t )</td>
<td>-0.082309</td>
<td>0.002658</td>
</tr>
<tr>
<td>( \alpha_u )</td>
<td>0.048984</td>
<td>0.001136</td>
<td>( \alpha_u )</td>
<td>0.049426</td>
<td>0.001514</td>
<td>( \alpha_u )</td>
<td>0.041644</td>
<td>0.006568</td>
</tr>
<tr>
<td>( \alpha_v )</td>
<td>0.048984</td>
<td>0.001136</td>
<td>( \alpha_v )</td>
<td>-0.081713</td>
<td>0.002028</td>
<td>( \alpha_v )</td>
<td>-0.082309</td>
<td>0.002658</td>
</tr>
<tr>
<td>( \beta_k )</td>
<td>0.371844</td>
<td>0.009390</td>
<td>( \beta_k )</td>
<td>0.390704</td>
<td>0.011567</td>
<td>( \beta_k )</td>
<td>0.372009</td>
<td>0.017265</td>
</tr>
<tr>
<td>( \beta_l )</td>
<td>0.390704</td>
<td>0.011567</td>
<td>( \beta_l )</td>
<td>0.358556</td>
<td>0.005547</td>
<td>( \beta_l )</td>
<td>0.348343</td>
<td>0.006296</td>
</tr>
<tr>
<td>( \beta_m )</td>
<td>0.372009</td>
<td>0.017265</td>
<td>( \beta_m )</td>
<td>0.358556</td>
<td>0.005547</td>
<td>( \beta_m )</td>
<td>0.348343</td>
<td>0.006296</td>
</tr>
<tr>
<td>( \beta_t )</td>
<td>0.992668</td>
<td>0.010154</td>
<td>( \beta_t )</td>
<td>-0.081713</td>
<td>0.002028</td>
<td>( \beta_t )</td>
<td>-0.082309</td>
<td>0.002658</td>
</tr>
<tr>
<td>( \beta_u )</td>
<td>0.048984</td>
<td>0.001136</td>
<td>( \beta_u )</td>
<td>0.049426</td>
<td>0.001514</td>
<td>( \beta_u )</td>
<td>0.041644</td>
<td>0.006568</td>
</tr>
<tr>
<td>( \beta_v )</td>
<td>0.048984</td>
<td>0.001136</td>
<td>( \beta_v )</td>
<td>-0.081713</td>
<td>0.002028</td>
<td>( \beta_v )</td>
<td>-0.082309</td>
<td>0.002658</td>
</tr>
</tbody>
</table>

155
Table B.4: Parameter estimates with interaction terms included

<table>
<thead>
<tr>
<th>Variable</th>
<th>Bootstrapped version (E2)</th>
<th>Non-bootstrapped version (E1)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point estimate</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$\alpha_k$</td>
<td>0.278107</td>
<td>0.010540</td>
</tr>
<tr>
<td>$\alpha_l$</td>
<td>0.363610</td>
<td>0.010611</td>
</tr>
<tr>
<td>$\alpha_m$</td>
<td>0.358205</td>
<td>0.005495</td>
</tr>
<tr>
<td>$\alpha_t$</td>
<td>0.983512</td>
<td>0.012609</td>
</tr>
<tr>
<td>$\beta_{kk}$</td>
<td>0.049799</td>
<td>0.001063</td>
</tr>
<tr>
<td>$\beta_{kl}$</td>
<td>-0.083061</td>
<td>0.001908</td>
</tr>
<tr>
<td>$\beta_{km}$</td>
<td>-0.016534</td>
<td>0.000552</td>
</tr>
<tr>
<td>$\beta_{kt}$</td>
<td>0.003148</td>
<td>0.001257</td>
</tr>
<tr>
<td>$\beta_{ll}$</td>
<td>0.106295</td>
<td>0.001300</td>
</tr>
<tr>
<td>$\beta_{lm}$</td>
<td>-0.129555</td>
<td>0.001769</td>
</tr>
<tr>
<td>$\beta_{lt}$</td>
<td>-0.002559</td>
<td>0.001170</td>
</tr>
<tr>
<td>$\beta_{mm}$</td>
<td>0.073036</td>
<td>0.000844</td>
</tr>
<tr>
<td>$\beta_{mt}$</td>
<td>-0.000585</td>
<td>0.000371</td>
</tr>
<tr>
<td>$\beta_{tt}$</td>
<td>0.001242</td>
<td>0.000658</td>
</tr>
<tr>
<td>$\alpha_d$</td>
<td>-0.002412</td>
<td>0.004290</td>
</tr>
<tr>
<td>$\alpha_r$</td>
<td>-0.002063</td>
<td>0.004461</td>
</tr>
<tr>
<td>$\alpha_{dk}$</td>
<td>0.003297</td>
<td>0.005422</td>
</tr>
<tr>
<td>$\alpha_{dl}$</td>
<td>-0.006468</td>
<td>0.012837</td>
</tr>
<tr>
<td>$\alpha_{dm}$</td>
<td>-0.010590</td>
<td>0.008087</td>
</tr>
<tr>
<td>$\alpha_{dr}$</td>
<td>-0.000017</td>
<td>0.002431</td>
</tr>
<tr>
<td>$\alpha_{rk}$</td>
<td>-0.000167</td>
<td>0.006075</td>
</tr>
<tr>
<td>$\alpha_{rl}$</td>
<td>-0.004553</td>
<td>0.004593</td>
</tr>
<tr>
<td>$\alpha_{rm}$</td>
<td>-0.000163</td>
<td>0.001947</td>
</tr>
<tr>
<td>$\alpha_{rt}$</td>
<td>-0.000755</td>
<td>0.001044</td>
</tr>
<tr>
<td>(cons)</td>
<td>1.173767</td>
<td>0.060321</td>
</tr>
</tbody>
</table>
Appendix C

Appendix to Chapter 4

C.1 Parameter estimates of second-stage DEA/Tobit regression

This section contains point estimates from the second-stage DEA/Tobit hospital efficiency regression in chapter 4, and their corresponding bootstrapped 95% confidence intervals. An asterisk indicates statistical significance at 95%. Teaching hospitals are located in major urban centres and classified as metropolitan. Bootstrapped CIs are preferred over standard error because the efficiency values were generated by DEA and the distribution may not be symmetric. It is possible that the point estimate is not contained in the CI, indicating a very skewed implied EDF.

To read: column one lists the aggregation tier used and column two states which hospitals are included. The row in the top is the point estimate and the row in the bottom is the 95% two-sided confidence interval.
Table C.1: Parameter estimates of independent variables driving hospital efficiency

<table>
<thead>
<tr>
<th></th>
<th>(constant)</th>
<th>quality</th>
<th>HHI</th>
<th>teaching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group - SIR Pooled</td>
<td>1.2959</td>
<td>0.335</td>
<td>0.6864</td>
<td>-0.2148</td>
</tr>
<tr>
<td>Metro</td>
<td>1.0138</td>
<td>*0.2683</td>
<td>-1.8773</td>
<td>0.0283</td>
</tr>
<tr>
<td>Rural</td>
<td>1.2441</td>
<td>*0.2039</td>
<td>*0.4826</td>
<td></td>
</tr>
<tr>
<td>[1.3999, 1.6159]</td>
<td>[0.2977, 0.3895]</td>
<td>[-0.1250, 0.2804]</td>
<td>[-0.1685, 0.0034]</td>
<td></td>
</tr>
<tr>
<td>MDC - SIR Pooled</td>
<td>1.3502</td>
<td>*0.3726</td>
<td>0.7885</td>
<td>-0.3067</td>
</tr>
<tr>
<td>Metro</td>
<td>1.1619</td>
<td>*0.2194</td>
<td>-3.2978</td>
<td>-0.0809</td>
</tr>
<tr>
<td>Rural</td>
<td>1.2745</td>
<td>*0.2045</td>
<td>*0.5668</td>
<td></td>
</tr>
<tr>
<td>[1.4779, 1.7373]</td>
<td>[0.3120, 0.4515]</td>
<td>[-0.1868, 0.2603]</td>
<td>[-0.1428, -0.2029]</td>
<td></td>
</tr>
<tr>
<td>DRG - SIR Pooled</td>
<td>1.2181</td>
<td>*0.5806</td>
<td>0.2955</td>
<td>-0.3472</td>
</tr>
<tr>
<td>Metro</td>
<td>1.1951</td>
<td>*0.2345</td>
<td>-3.4066</td>
<td>-0.1277</td>
</tr>
<tr>
<td>Rural</td>
<td>1.1876</td>
<td>*0.2938</td>
<td>*0.3346</td>
<td></td>
</tr>
<tr>
<td>[1.3374, 1.6277]</td>
<td>[0.4829, 0.6933]</td>
<td>[-0.2325, 0.1889]</td>
<td>[-0.1859, -0.2941]</td>
<td></td>
</tr>
</tbody>
</table>

Parameter estimates and their approximate 95% CIs of the second stage analysis. Results with * indicates statistical significance at the 95% level. The dependent variable is the inefficiency score of hospitals, defined as the reciprocal of efficiency, to simplify the Tobit regression.

Table C.2: Robustness check: parameter estimates using alternative quality statistic

<table>
<thead>
<tr>
<th></th>
<th>(constant)</th>
<th>quality</th>
<th>HHI</th>
<th>teaching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group - Diff Pooled</td>
<td>3.1609</td>
<td>-1.3039</td>
<td>0.5580</td>
<td>*-0.6343</td>
</tr>
<tr>
<td>Metro</td>
<td>1.8907</td>
<td>0.2060</td>
<td>-7.7131</td>
<td>0.3436</td>
</tr>
<tr>
<td>Rural</td>
<td>1.7116</td>
<td>-2.9200</td>
<td>*1.1003</td>
<td></td>
</tr>
<tr>
<td>[3.4844, 4.0181]</td>
<td>[-1.7837, 3.8426]</td>
<td>[-0.5723, 0.7920]</td>
<td>[-0.6674, -0.3298]</td>
<td></td>
</tr>
<tr>
<td>MDC - Diff Pooled</td>
<td>3.3686</td>
<td>-1.8833</td>
<td>0.3007</td>
<td>-0.4664</td>
</tr>
<tr>
<td>Metro</td>
<td>1.9894</td>
<td>0.0615</td>
<td>-7.8662</td>
<td>0.5542</td>
</tr>
<tr>
<td>Rural</td>
<td>1.7177</td>
<td>-3.0200</td>
<td>*1.0983</td>
<td></td>
</tr>
<tr>
<td>[3.7057, 4.2528]</td>
<td>[-2.2206, 2.9669]</td>
<td>[-0.4978, 0.8673]</td>
<td>[-0.6099, 0.2206]</td>
<td></td>
</tr>
<tr>
<td>DRG - Diff Pooled</td>
<td>3.5109</td>
<td>-3.0638</td>
<td>0.0829</td>
<td>0.0417</td>
</tr>
<tr>
<td>Metro</td>
<td>2.2221</td>
<td>-0.9894</td>
<td>-5.6554</td>
<td>0.9575</td>
</tr>
<tr>
<td>Rural</td>
<td>1.7183</td>
<td>-3.6003</td>
<td>*1.095</td>
<td></td>
</tr>
<tr>
<td>[3.8214, 4.4630]</td>
<td>[-3.2517, 3.6468]</td>
<td>[-0.2011, 1.5174]</td>
<td>[0.0684, -0.2011]</td>
<td></td>
</tr>
</tbody>
</table>

158
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Li, Chun Lok K.

Title:
Quality-adjusted efficiency measures for public hospitals

Date:
2014

Citation:

Persistent Link:
http://hdl.handle.net/11343/38631

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.