Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP01/2021

Performance Analysis of Hospitals in Australia and its Peers: A Systematic Review

Zhichao Wang and Valentin Zelenyuk

Date: January 2021

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Performance Analysis of Hospitals in Australia and its Peers: A Systematic Review

Zhichao Wang[*], Valentin Zelenyuk[†]

January 15, 2021

Abstract

Research about the productivity and efficiency of hospitals in providing healthcare services has developed substantially in the last few decades. How does this topic proceed in Australia, its peer countries and regions who share a similar healthcare system? In this article, we conduct a systematic review and a series of bibliometric analyses of the research about the efficiency of hospitals, which are the core organizations in the the healthcare system, in order to obtain a broad perspective of this topic in Australia and its peers. Among others, a random forests model was trained to evaluate the impact of features of an article on the scientific influence of the research. We used bibliometric data in Scopus from 1970 to 2020 and extracted the review pool by a peer-review process. Besides identifying the productive authors and most cited publication sources, the bibliometric analysis also indicated a shifting of topics over time. Through the training process of the random forests classification model, the most influential features of an article were also identified.

Keywords: Performance analysis, efficiency, Australia, hospital, systematic review, bibliometric analysis, random forests, machine learning

JEL Codes: C14, C61, D24, I11

[*] School of Economics, University of Queensland, Brisbane, Qld 4072, Australia

[†] School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Qld 4072, Australia

1 Introduction

Healthcare costs in most developed countries have grown dramatically in the last few decades. According to records from the Organization for Economic Co-Operation and Development (OECD), the proportion of Gross Domestic Product (GDP) devoted to healthcare systems has increased from an average among all member countries of 4.6% in 1970, when the OECD was founded, to 6.5% in 1990 and reached 8.7% in 2010 (Organization for Economic Co-operation and Development, 2020; Eckermann and Coelli, 2013). Moreover, the average proportion of developed country members continued to increase reaching an average of around 10% in 2019. As for Australia, the proportion of health care expenditure to GDP shows a consistent and upward trend, which in 2019 was estimated at 9.3%, which locates at around the 40 percentile among all OECD members (Organization for Economic Co-operation and Development, 2020). While compared with other member countries sharing a similar healthcare system, such as the United Kingdom (UK), Canada, and New Zealand, the proportions of GDP spent in healthcare show a high degree of homogeneity over the last 50 years.

In another view of national expenditure, when opening the budget of the Australian government in 2019-2020, the expense in the "Health" sector ($87,023 million, 15.04% of total expenses, 4.78% of GDP) is second only to the "Social security and welfare" sector ($196,119 million, 33.90% of the total expense, 10.78% of GDP), and is followed by "Education" and "Defense". In recent years the budget in the health category is even more than that of education together with defense (Commonwealth of Australia 2020, 2020). Meanwhile, considering the huge expenditure from the consumer side, the expenditure of healthcare greatly exceeds the budget expenses. According to the report from the Australian Institute of Health and Welfare (AIHW), the nation-wide healthcare spending on health goods and services in 2018-19 was $195.7 billion, equating to $7,772 per person, and accounting for about 10% of overall economic activity (Australian Institute of Health and Welfare, 2020).

For such tremendous and still rapidly increasing costs, due to both public pressure and the executive interest of cost containment, much research has been published for policymakers to solve these crucial problems (O'Neill et al., 2008). It is widely believed that the inefficiency of health care institutions has contributed to some degree to this global continuous increase (Worthington, 2004). Therefore, efficiency measurement and improvement may be the first step in the evaluation of a coordinated health care system (O'Neill et al., 2008), whether for a better healthcare service outcome or a more controllable budget. Hospitals, as the core institutions of the healthcare system, which also account for the majority of 40% of Australian healthcare spending in 2017-18, have become the most popular research target (Australian Institute of Health and Welfare, 2020). If we simply search the keywords "efficiency" and "hospital" in Google Scholar, the results were around 705,000 rows dating from 1979 to 1999, but increased to 2,170,000 between 2000 and 2020. Along with the increasing global jump of research, what have we discovered about the hospital efficiency of Australia and what could be inspired to achieve for the future? To obtain a broader scope of papers for review and further analysis, we extended the topic about Australia to several other countries and regions, which share a similar healthcare system with high-level quality healthcare services and similar traditions in British culture, i.e. the UK, Canada, New Zealand and Hong Kong.

There are some good reviews of similar topics worldwide. For example, Hollingsworth et al. (1999) and Hollingsworth (2003) and (2008) broadly reviewed the studies of healthcare delivery efficiency, focusing on the application and development of methods, the main findings and the indicators of output and quality, etc. The measurement of efficiency, including the indicators of input and output of a facility and the approach to evaluate the utilization were also among the main concerns of the reviews by Worthington (2004), O'Neill (2008) and Hussey (2009).

The novelty of our review could be expressed in three aspects. Firstly, we conducted a systematic review instead of the commonly used traditional approach to avoid the selection bias in the paper selection. The articles chosen by the authors, though mostly by field experts, might not be representative of the existing knowledge (Linnenluecke et al., 2020). On the contrary, a systematic review collects a comprehensive set of available research and selects papers by predetermined criteria for further analysis (Linnenluecke et al., 2020; Tranfield et al., 2003). Secondly, we introduced bibliometric tools and analyzing methods for visualization and network analysis, which have been widely used in some reviews about other topics (e.g. Hyun-do Choi (2019)), into the review of efficiency studies of hospitals. Consequently, the dynamic patterns of the most researched topics and productive authors were revealed. Furthermore, we constructed a random forests model to classify the research by scientific influence, which is represented by the average citations per year and we identified the most influential features of an article on the classification, which could also be an indication of the preference of the peer researchers under this topic.

The remainder of this paper is organized as follows. After this introduction section, selected previous reviews on the relative topic are presented in Section 2. Methodologies about paper collection, processing procedure, bibliometric analysis and random forests are discussed in Section 3. Section 4 presents detailed analysis results and the concluding remarks are summarized in Section 5.

2 Related Works

Along with the continuous growing of demand from policymakers and the publishing of relative topics, review studies have been conducted during the development of the research field. As a pioneer, Hollingsworth et al. (1999) reviewed the global studies of non-parametric methods and applications in healthcare efficiency published before 1997 and found that more than two-thirds of the applications used data envelopment analysis (DEA) or DEA-related techniques, and more than two-thirds of the research is about hospitals and nursing homes in the US. Besides, in most evidence from the US and the EU, public provision performs better in efficiency. Years later, when parametric methods, represented by stochastic frontier analysis (SFA), had been further developed, Hollingsworth (2003) reviewed this topic again. By reviewing 188 papers on frontier techniques in healthcare efficiency studies, similar conclusions to those in the previous review were obtained by the author that public provision potentially performs with less variability than private. Parametric methods, such as SFA, have been more widely used, however, the dominant methods are still DEA and DEA related techniques. In the meantime, under an overall perspective of frontier techniques, about three-quarters of the research is based on DEA, SFA, and their variations. In another review of the hospital efficiency studies with DEA-based methods, O'Neil et al. (2008) reviewed 79 studies from 1984 to 2004, which are also mainly focused on the samples from the US and Europe. Besides the selection of inputs and outputs, this cross-national comparison study also revealed the difference in preference of the research topic and model selection. For example, they concluded that the European researchers pay more attention to the allocative efficiency than to the technical efficiency, compared with those in the US.

In a nutshell, basically three measures of efficiency have been developed to satisfy the requirements of researchers and policy makers. These are technical efficiency, allocative efficiency, and the combination of both, which are focused on maximizing output from a given input, minimizing input with expected output and their dual counterparts[1] (Worthington, 2004; Färe et al., 2019). Worthington (2004) reviewed the efficiency measurements applied in healthcare topics, especially in-depth of the frontier techniques. The author checked the main approaches, such as DEA, SFA, the Malmquist index (MI) and their combinations, and the implications, including the input and output indicators and explanatory approaches of the difference in efficiency. In conclusion, although the efficiency measurement has attracted more and more attention in the early 2000s, the applications of advanced frontier techniques are still in an early stage.

More recently, Hollingsworth (2008) reviewed a broader scope of 317 papers. The popular trend of methods is the same as years ago. Besides, sharing the same viewpoint with other discussions and reviews on this topic, output indicators are mostly for physical performance, such as inpatient days, without considering the quality of the treatment. Only 9% of studies included outcome measures, such as the mortality rate and changes in health status. Another weakness is that only a few studies tested methods with statistical or sensitivity analysis. In conclusion, technique efficiency is mostly analyzed, while only a few of the studies focused on allocation efficiency.

[1] There is also a measure of scale efficiency, which can be measured both in primal and dual contexts (e.g., see Sickles and Zelenyuk (2019) and references therein).

Hussey et al. (2009) stated similar opinions of the deficiency of current research, especially in the aspects of output and statistical tests. By reviewing articles in Medline[2] and EconLit[3] from 1990 to 2008, the authors identified 265 measures in peer-reviewed articles and 8 in so-called 'gray literature'[4]. Their systematic review focused on the efficiency measures and tried to create a mutual understanding of the adequacy of these approaches. Following McGlynn et al. (2008), the measures were classified into three branches: "perspective", "inputs" and "outputs". Among the 265 measures abstracted from the 172 reviewed articles, the production of hospital service, such as length of stay and cost per discharge, was the most commonly used indicator. Half of the measurements used physical resources to reflect the input, while one third used costs and one quarter used both as input indicators. For output, most measurements count the healthcare service, for example, discharges, procedures and physician visits. In rare cases was quality integrated into output, which is the most concerning issue in the review and also an enduring focus of discussion in the field. Meanwhile, another empirical issue is that only about 2.3% of articles included tests of reliability or validity, while sensitivity analysis was considered in about one-quarter of the articles, even though it is commonly used in multivariate statistical models.

Consequently, according to the conclusive reviews by the field experts, the efficiency studies on healthcare show a prosperous atmosphere, especially in the US and the EU. However, there is a lack of review of similar topics in Australia, especially about hospitals.

3 Methodologies

3.1 Paper collection

Similar to the paper collection used in Hussey et al. (2009), we systematically collected the published articles and searched the gray literature about the hospital efficiency studies of Australia and its peer countries and regions. The main requirement of a systematic review is to be comprehensive (Tranfield et al., 2003), while the fundamental idea is replicable. Following the procedure of data collection in Linnenluecke et al. (2020) and Choi and Oh (2019), we designed a collection and selection process in order to reach a confluence of possibly all related published articles and gray literature.

Firstly, we chose Scopus as the main database for its comprehensiveness in healthcare research and the adaptive format for the mainstream bibliometric analysis techniques, while the Web of Science or more specialized platforms such as Medline or PubMed could be good alternatives. The second crucial decision

[2] MEDLINE is a bibliographic database of life sciences managed by the U.S. National Library of Medicine.

[3] EconLit is an economics literature database provided by the American Economic Association.

[4] Gray literature is defined as the research produced by academics, government, industry, etc, which is not controlled by traditional publishers. Common gray literature publication types include theses and dissertations, working papers, conference papers and government documents (Paez, 2017).

for this theme-centric review[5]is the design of keywords for the Boolean search.

The idea of the Boolean search for a systematic review is searching the predetermined keywords combined by logical operators in the selected text fields to extract a paper pool that covers all the papers relative to the target topic. The keywords are determined by the review theme. The fields could be title, abstract, authors, etc of an article. As for the operators, "AND" and "OR" are the most commonly used so that most requirements could be satisfied by their combination.

Our review theme could be decomposed into three fields, which are "location", "topic" and "object". The "location" restricts the interested country or region as the research target during the data collection. Besides the country names, the sub-level district names were also included. For example in Australia, apart from "Australia" and "Australian" as the keywords, "Queensland", "Victoria" and names of all the other states and territories were used as the search keys in the "location" group as well. Moreover, the commonly used abbreviations, such as UK to United Kingdom, and synonyms, such as British to Britain, were also listed simultaneously. Since there are five interested countries and regions, we listed five keyword groups for Australia, UK, Canada, New Zealand and Hong Kong respectively.

The "topic" limits the research aim in the paper collection, and the terms used are "efficiency", "inefficiency", "productivity" and "performance analysis", which are the same for all targeted countries and regions. Finally, the researched "object" is constrained among "hospital", "healthcare" (or "health care") and "health services". Although "hospital" is the main target in our review, we chose to include similar phrases, which were a little more generalized, so that no crucial material would be missed for further analysis.

The logic operators for the Boolean search is clear and unified among all fields. The result is an intersection among the three keyword groups, obtained by the logical operator "AND". Meanwhile in each group, the union is combined with each equivalent keyword in each field (e.g. title or abstract) with the logical operator "OR". However, the fields for the Boolean search of each of the three groups were different. After trials and considerations for comprehensiveness and accuracy, "topic" and "object" were searched in the title and keyword (including author keys and index terms[6]) fields, while keywords in the abstract were not accepted. Some papers may mention "hospital" or "efficiency" in the abstract, but if such terms are not recorded in the keywords or title, they are mostly not related to our topic. On the contrary, some authors may not indicate a toponym in their title or keywords, not even in the abstract. Thus, we searched "location" terms in the text fields including title, keywords and affiliation of authors to filter out more related papers. Another configuration worth noting is that we added the wildcard "*" at both ends of the one-word keywords, such as "efficiency", to allow the similar spelled terms, such as "inefficiency". An illustration by of the Boolean

[5] Generally, reviews could be distinguished into author-centric or theme-centric by their orientation (Linnenluecke et al., 2020). More details could be found in Webster and Watson (2002).

[6] Index term is the controlled vocabulary assigned to the article to represent the main topic.

search logic discussed above is as presented in Figure 1.

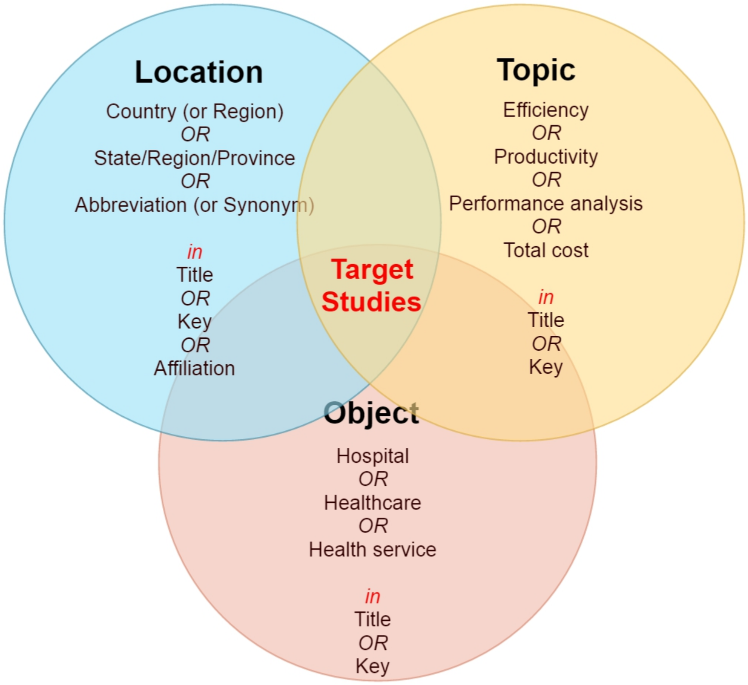


Figure 1: Illustration of the Boolean search logic

As for gray literature, we searched similar keywords in Google Scholar and RePEc. For example, regarding Australia, we found 6 papers which were not included in the previous search but were very close to our review topic. We would review these studies together with other selected papers from the published platform, but due to one limit of our study that we will discuss in later sections, the gray literature couldn't be included in further bibliometric analysis because of their lack of bibliometric information.

The collection at this stage is usually too wide for effective review or analysis. Similar to the procedure used in Linnenluecke et al. (2020), we dropped unrelated papers by manually reviewing them with the predetermined criteria. The selection criteria were determined by field experience and were adjusted by the condition of the existing paper pool, which is aimed at extracting every piece of paper that is strictly related to the theme of the review. The idea of manual selection is to leave unrelated papers out of the scope cautiously, rather than extracting the target papers from a search result by conditions. Another step worth noting is that we found that some research satisfying the Boolean search constraints were indeed focused on a certain disease, equipment or department. In research such as the evaluation of a new therapy, "hospital" is usually mentioned as the survey location and "efficiency" is used when describing the effects. These papers, however, are not useful efficiency studies of hospitals for our review. Therefore, we filtered out a portion of papers by detecting common disease terms in titles before the manual review.

3.2 Bibliometric analysis techniques

The academic knowledge is expanding dramatically in that it is more and more difficult for researchers to review, analyze and understand a field relying only on manual reading (Linnenluecke et al., 2020). In addition to the efficiency, the massive resources also force researchers to choose "high quality" materials but not necessarily consider a broader range of evidence (Tranfield et al., 2003). With the development of text mining and visualization techniques, bibliometric mapping approaches have been more helpful in systematic review. In fact, it is more than a beneficial tool, but to some degree a necessary route to conduct a confidential systematic review.

Analysis, such as co-citation analysis (Small, 1973) and co-word analysis (Callon et al., 1983), has been long applied in mapping research field structure (Nieminen et al., 2013). With the developing big data theories, scientometric researchers have programmed functional tools to solve the analysis demand on a huge data pool. Based on review and empirical research (e.g. Linnenluecke et al. (2020); Lee et al. (2014); Kohl et al. (2019); Ujum (2014); Nieminen et al. (2013)), we listed a group of commonly used tools and tested the mapping and visualizing functions. Some well-developed tools are based on the Java platform, such as VOSviewer (Van Eck and Waltman, 2010), Sci2 (Science to Science Tool) (Team, Sci, 2009) and SciMAT (Cobo et al., 2012), which are powerful in mapping visualization and network analysis. These Java-based tools are also usually equipped with text mining functions to handle data modification in advance of mapping. There are some other similar tools, such as CiteSpace II (Chen, 2006), whose main functions are similar to Sci2, and Network Workbench Tool (Workbench et al., 2006), which is also functional and has open access to data modification. Another series of tools are developed in the statistical programming language R (R Core Team, 2019). One advantage of these open-source packages is the flexible and extensible working environment, where researchers and practitioners could continuously provide updates for the functions (Linnenluecke et al., 2020). A representative package developed in R is Bibliometrix (Aria and Cuccurullo, 2017), which supports both descriptive analysis and network analysis. Besides, with the deployment of another package, Shiny (Chang et al., 2020), the functions of Bibliometrix could be applied in a user-friendly interaction interface through web applications. There are some other tools developed for mapping and network analysis, such as Pajek (Batagelj and Mrvar, 1998) and Gephi (Bastian et al., 2009), whose functions could be cooperated with other applications, such as Bibliometrix and VOSviewer. Another pioneer software Histcite (Garfield, 2009) is developed for networking analysis of key authors and articles, which however is no longer in development now.

Most of these popular tools are designed to allow importing bibliometric data from the Web of Science and

Scopus, which are usually stored as either Bibtex (bib), Plain Text (txt) or RIS (ris). However, the storage formats of different platforms are incompatible. Considering the function requirement and the condition of our bibliometric data set, we mainly use Bibliometrix and VOSviewer in further analysis.

3.3 Classification analysis by Random Forests

In Breiman et al. (1984), a model was introduced to solve the problem of classifying the new observations by their relationship with a set of interested predictors, which is the classification tree model (Beaulac and Rosenthal, 2019). In brief, the classification tree is a supervised learning model with an algorithm that divides the space of the response variable into regions by the specific values of the predictors. Following a level by level procedure, the original or the subgroup of the training set would be divided into secondary groups by comparing their values of a certain predictor with a critical value. The criteria of the critical value, or the criteria of establishing the tree are defined as a measure of impurity, which represents the degree of observations that are grouped in the same region, however, not with the same class. Hence, the algorithm of selection turns into computing the impurity of the possible combinations of the predictor conditions and choosing the condition with the minimal impurity. Then the predictors and their values in the selected condition can be regarded as the nodes in the tree that divide the input group into subgroups step by step, while the regions in the last level can be considered as the leaves, and therefore the name classification tree (Beaulac and Rosenthal, 2019).

A classification tree is an easy-to-implement model, which is also clear for interpreting. One way to improve the performance of the classifiers could be imagined as raising a forest of a set of decision trees, which is better at prediction than any individual classifier. Hence a critical procedure is the aggregation of the class predictions, where a voting method is always considered to be a solution where the final prediction of a new observation simply relies on a vote among all the individual classifiers. Besides, a prerequisite of the improvement on accuracy is the stability of individual classifiers over the training set, which, however, is proven not to be met by the classification tree model in another study by Breiman et al. (1996). Since then, a lot of interest has been attracted in aggregating the classifiers, especially methods that improve the stability of the classification trees (Beaulac and Rosenthal, 2019; Liaw et al., 2002).

One of the well-known methods is bagging, where it is proposed by Breiman (1996) that each tree is developed independently with a bootstrap sample of the training set (Liaw et al., 2002).[7] The sample for constructing each tree is of the same size and is randomly drawn with replacement from the original training set, which not only improves the classifier stability, but also greatly reduces the chances of overfitting (Beaulac and Rosenthal, 2019).

[7] Another popular method is boosting (e.g. Schapire et al. (1998)), in which the successive trees are more weighted in the voting than the earlier predictors with incorrect predictions (Liaw et al., 2002).

Based on the idea of bagging, Breiman (2001) introduced random forests, which add another layer of randomness in inputs. The algorithm is choosing the best predictor conditions in each node among a subset of randomly selected predictors, rather than among all the variables in the standard classification tree model. This modification of the random forests model greatly improves the accuracy, as well as the utility. In fact, there are only two parameters (the number of variables at each node and the number of trees) in the model and the results are usually not sensitive to the values of these parameters (Liaw et al., 2002). Following the discussion and inspired by the illustrations in Breiman et al. (1984) and Beaulac and Rosenthal (2019), we draw a brief illustration of this algorithm as we discussed above in Figure 2.

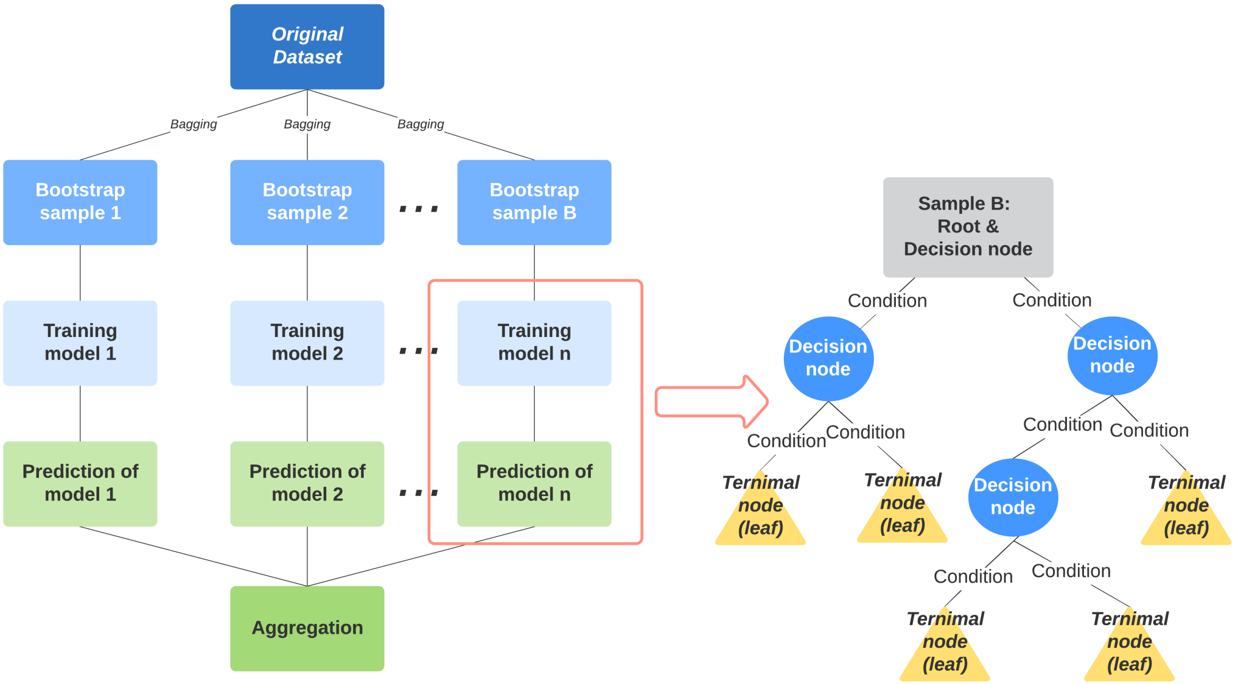


Figure 2: A brief illustration of the random forests model

In addition to the advantages discussed above (i.e. better accuracy than other algorithms, low risk of overfitting, user-friendly and easier in computation), another attractive property of random forests is that it allows for the computation of variable importance throughout the training process, which would be quite informative to our aim of evaluating the factors that impact the influence of research. Due to this meaningful result, random forests have been widely applied in predicting the response and interpreting the determinants in research. For example, Beaulac and Rosenthal (2019) ranked the factors impacting the student academic success with the variable importance obtained from a random forests model. As for bibliometric studies, Wang et al. (2019) performed random forests as one of the selected methods in comparing the capability of bibliometric indices in predicting the future success of articles. Nevertheless, with regard to the review of efficiency analysis, research about the factors impacting the scientific influence is very rare.

4 Results

4.1 Data collection

According to the paper collection procedure discussed in section 3.1, we obtained five paper pools for each interested country and region via the Scopus database, containing 1769 papers for Australia, 4719 papers for the UK, 5547 papers for Canada, 282 and 135 papers for New Zealand and Hong Kong from 1970 to 2020. Since searching using the same keywords, the number of results may indicate a degree of field prosperity of each selected country and region. As we present the process in Figure 3, after filtering out those papers containing disease or department names in their title, we manually reviewed the contents of each article and picked out the papers, which were strictly related to our review theme. As a result, we selected 9 papers for Australia, 15 for the UK, 8 for Canada, as well as 7 and 4 for New Zealand and Hong Kong, respectively. Intuitively, research in Australia regarding "hospital efficiency" is far less than in the UK or Canada, but the proportion of Australian research focusing on hospital-wise efficiency analysis is higher. On the other hand, the reason why more papers in the UK and Canada were excluded during the manual review stage is because of the higher rate of research focus on how new techniques and therapies improve the efficiency of a department or the treatment of a certain disease. These would meet the Boolean search constraints because "hospital" and "efficiency" were commonly mentioned in such studies, which does not match our selection criteria.

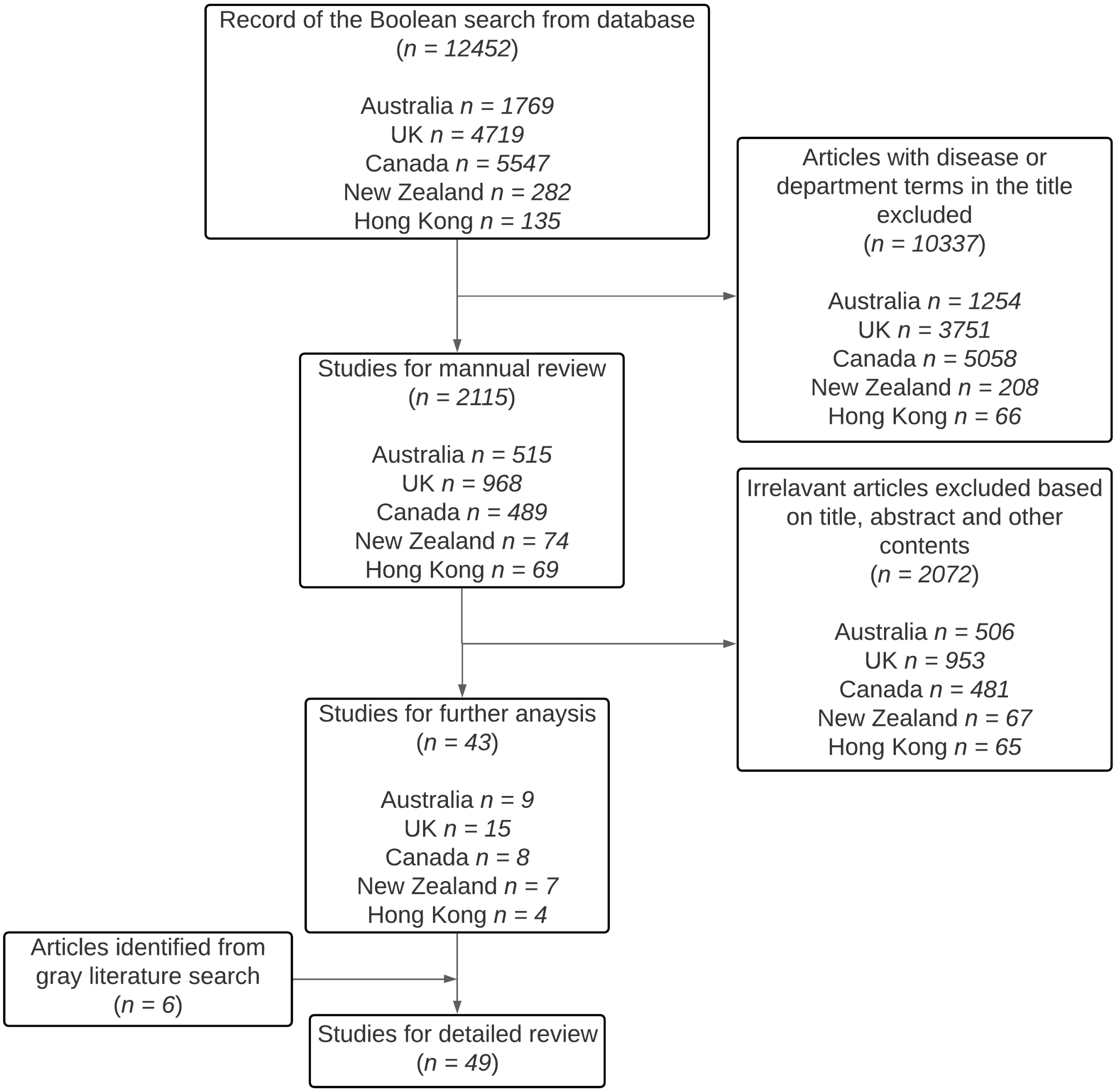


Figure 3: Literature collection flow

Another notable point is that the absolute quantity of papers about hospital-wise efficiency analysis is low among our interested countries and regions. Compared with the huge total amount of articles searched for by keywords and the wider review pool of the US in proceeding reviews, a lack of efficiency studies of hospitals is clear in Australia and its peers.

4.2 Bibliometric analysis

In recent years, visualization and mapping techniques, due to their powerful ability in helping to understand a huge amount of information in multiple dimensions, have attracted a lot of interest of authors of review studies. Linnenluecke et al. (2020) detailed the methodological steps of a systematic literature review. Moreover, they discussed the application of bibliometric mapping approaches, such as network analysis, Sankey diagram and topic extraction, to interpret and visualize the key findings of a comprehensive review. In another study by Choi and Oh (2019), the authors collected the papers published in the Journal of Productivity Analysis before early 2018 and analyzed the productive authors and their collaboration network. They also introduced a series of time dynamic word clouds of topics and collaboration networks to reflect the trend of research. Following the discussion in these papers, we designed a series of analyses based on the nature of our data.

At the beginning, we obtained a global sense of the relationship between the productive authors, popular keywords and the commonly published sources with a three-field plot (Sankey diagram)[8] for each country. The three fields used are keywords (both author keys and index terms), authors and published sources from left to right. The main field is "authors" in the middle, where the top productive authors were selected by each country, while the width of flow, whether between author and keyword or between author and source, represents the degree of relevance. Not all keywords or sources of each author are included in the analysis, but only those most commonly used by different authors.

As shown in Figure 4 for Australia, the top productive authors in the middle field show groups of productive collaborators. For example, Chua, Palangkaraya and Yong, whose works share the same keywords (author keys and index terms) on the left side, have articles published in the Economic Record and Health Economics as shown in the right side. As an active collaborator, Yong also has joint papers with another group of researchers, Cheng, Scott and Sundararajan. Moreover, if we focus on research methods and research objects in the keywords on the left side and ignore common terms such as "Australia", "article" and "humans", it indicates some trends in local research, such as "risk assessment" and "mortality" as the most frequently mentioned objects.

[8] Sankey diagrams are designed to visualize the flow of networks and processes. They illustrate the rate of flows, the relationships, and transformation with the arrow and the width (Froehlich, 2005).

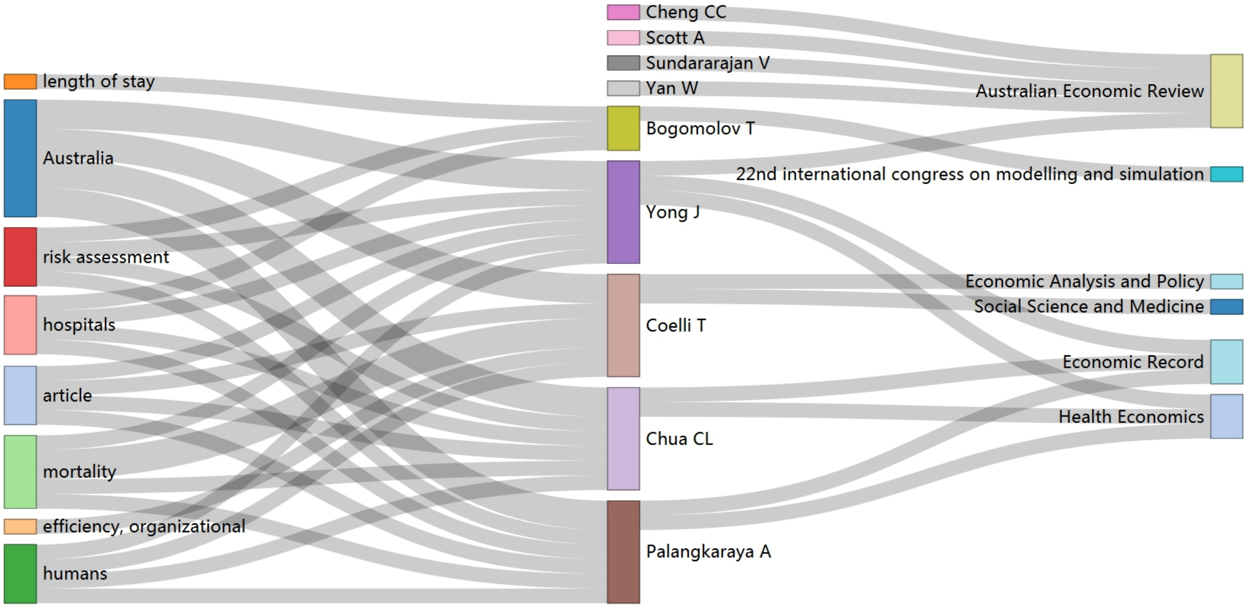


Figure 4: Sankey diagram of Australia

Similar plots were drawn for other peer countries for comparison. As shown in Figure 5 to Figure 7, it is similar to the case in Australia that productive researchers of this topic in the UK also show a trend of grouping, like the co-authorship among Bojke, Castelli, Street, Laudicella and Ward. Furthermore, researchers in Canada and New Zealand show more attention to the DEA method, such as the Chowdhury, Zelenyuk and Laporte group for Canada and Rouse, Harrison and Turner group in New Zealand. It is also shown by keywords on the left side that researchers in New Zealand are more interested in assessing private healthcare.

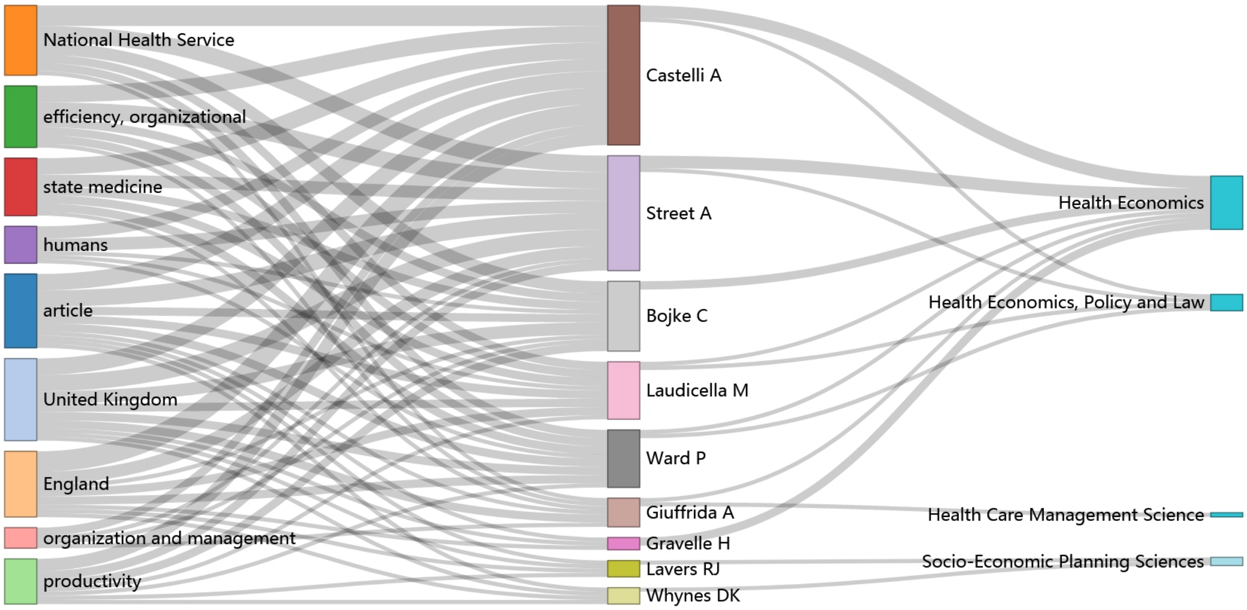


Figure 5: Sankey diagram of UK

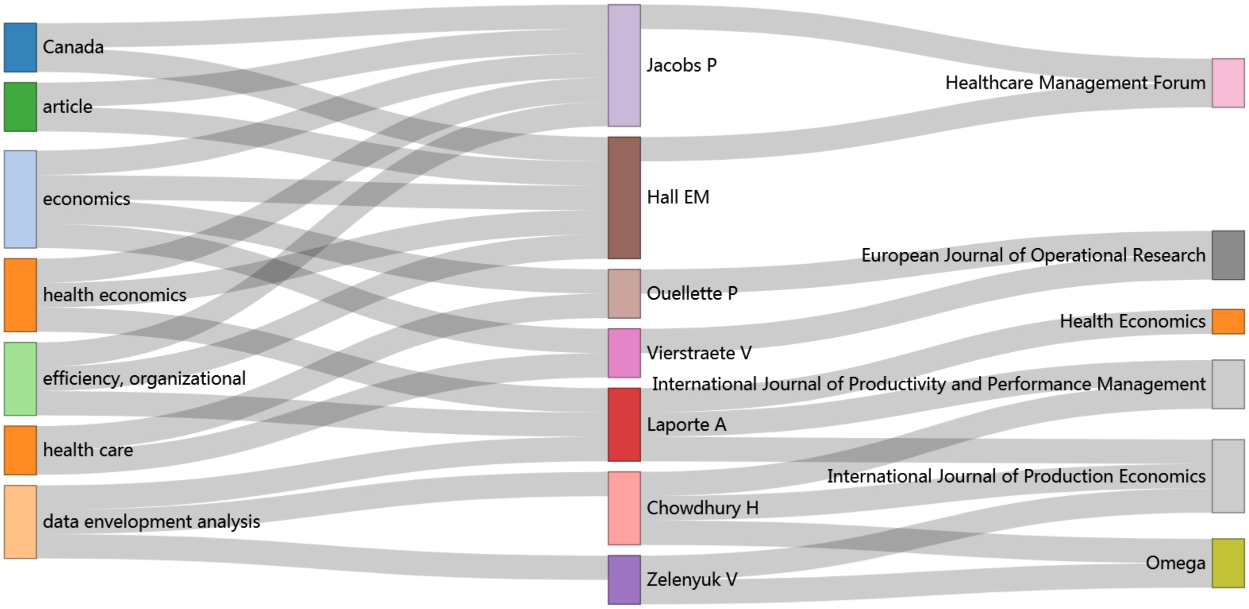

Canada
article
economics
health economics
efficiency, organizational
health care
data envelopment analysis
Jacobs P
Hall EM
Ouellette P
Vierstraete V
Laporte A
Chowdhury H
Zelenyuk V
Healthcare Management Forum
European Journal of Operational Research
Health Economics
International Journal of Productivity and Performance Management
International Journal of Production Economics
Omega

Figure 6: Sankey diagram of Canada

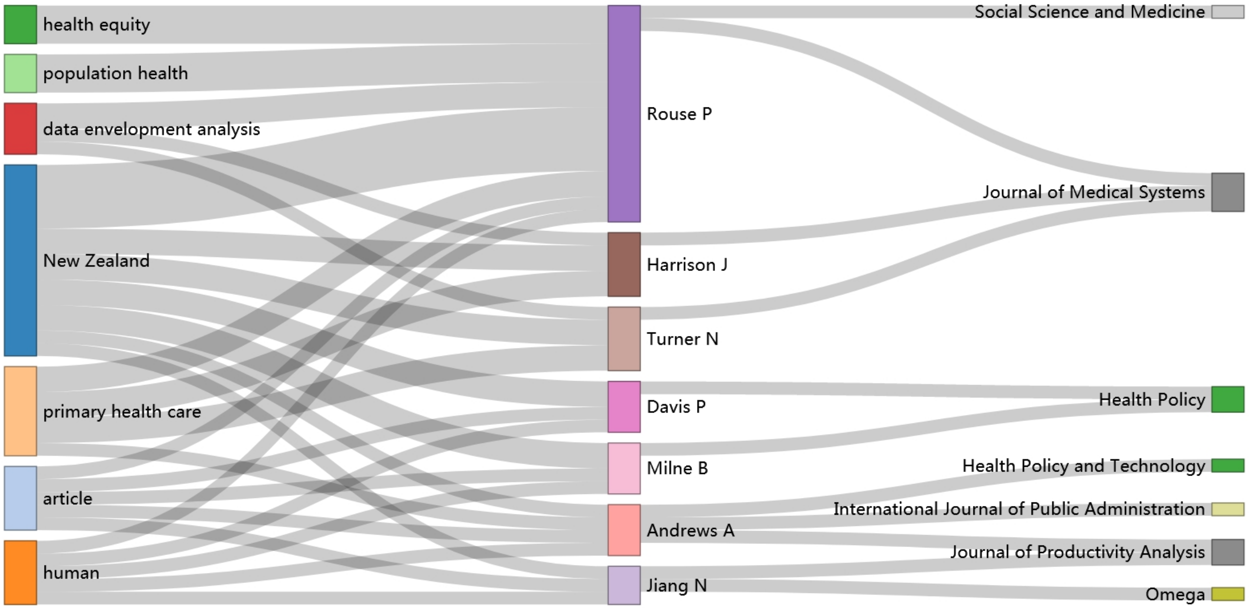
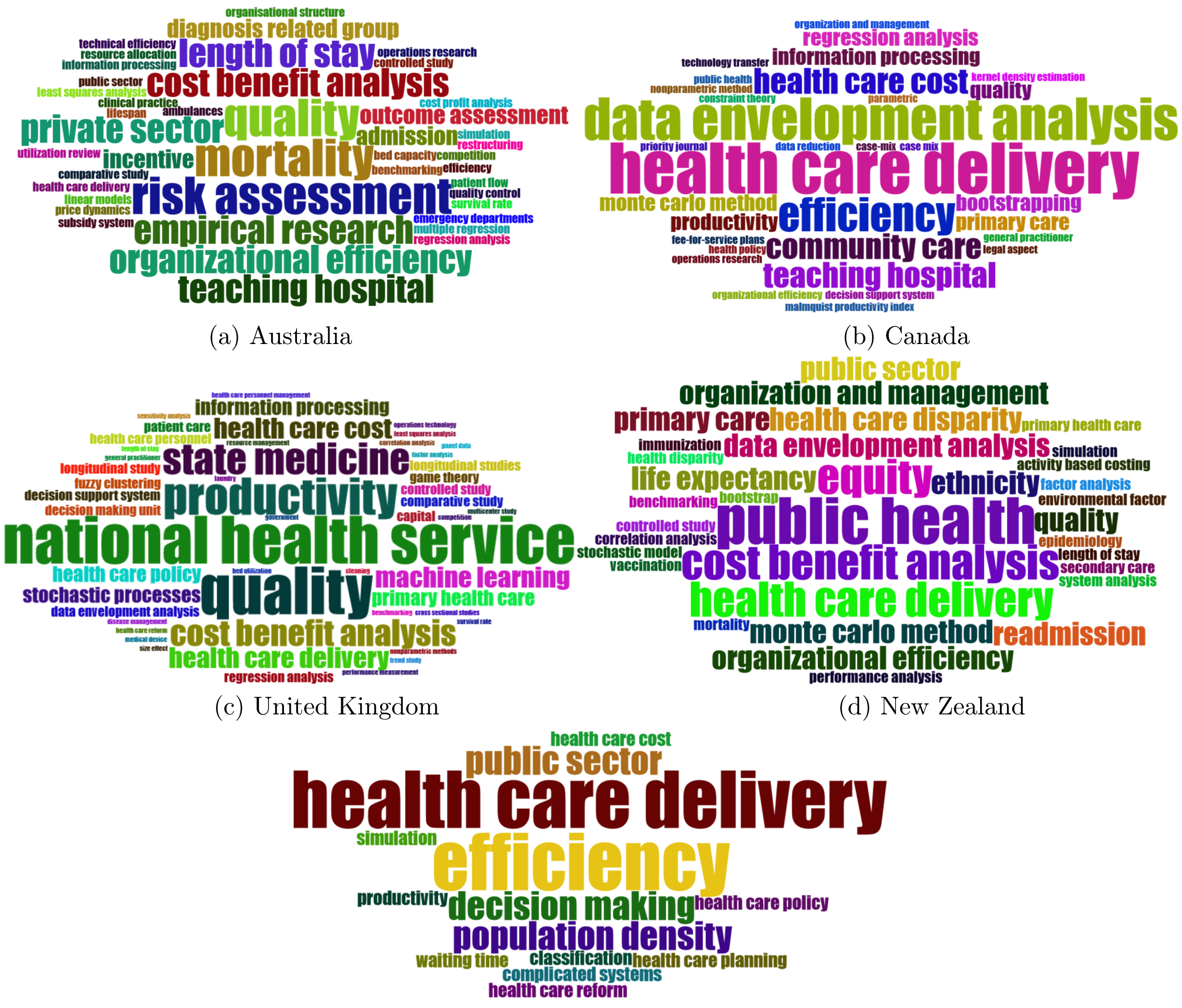

health equity
population health
data envelopment analysis
New Zealand
primary health care
article
human
Rouse P
Harrison J
Turner N
Davis P
Milne B
Andrews A
Jiang N
Social Science and Medicine
Journal of Medical Systems
Health Policy
Health Policy and Technology
International Journal of Public Administration
Journal of Productivity Analysis
Omega

Figure 7: Sankey diagram of New Zealand

A word cloud of keywords in each country and region was created to illustrate the research focus, where, as shown in Figure 8, the size of each term is determined by the frequency of appearance comparing with the other terms in a certain country or region. The keywords used were cleaned by dropping common terms (such as country names) to emphasize the methodologies and objects that researchers would pay more attention to. In addition to output assessment terms, such as “mortality” and “length of stay”, researchers in Australia preferred using “risk assessment” and “cost benefit analysis” in measuring “organizational efficiency”. Meanwhile, research in Canada focused more on methodology terms, such as DEA, “bootstrapping”, “Monte Carlo method” and “regression analysis”. Similarly as observed before, New Zealand researchers care more about “public health” and “primary health care”, while frequently using “cost benefit analysis”, DEA and “Monte Carlo method”. Researchers in the UK paid more attention to the National Health Service (NHS), “state medicine” and “quality”, as well as a new technique, “machine learning”. Hong Kong is most concerned with “health care delivery”, which is also a popular topic in Canada and New Zealand. Besides, “population density” is focused on in Hong Kong research, which is special among the other peers.



(a) Australia

(b) Canada

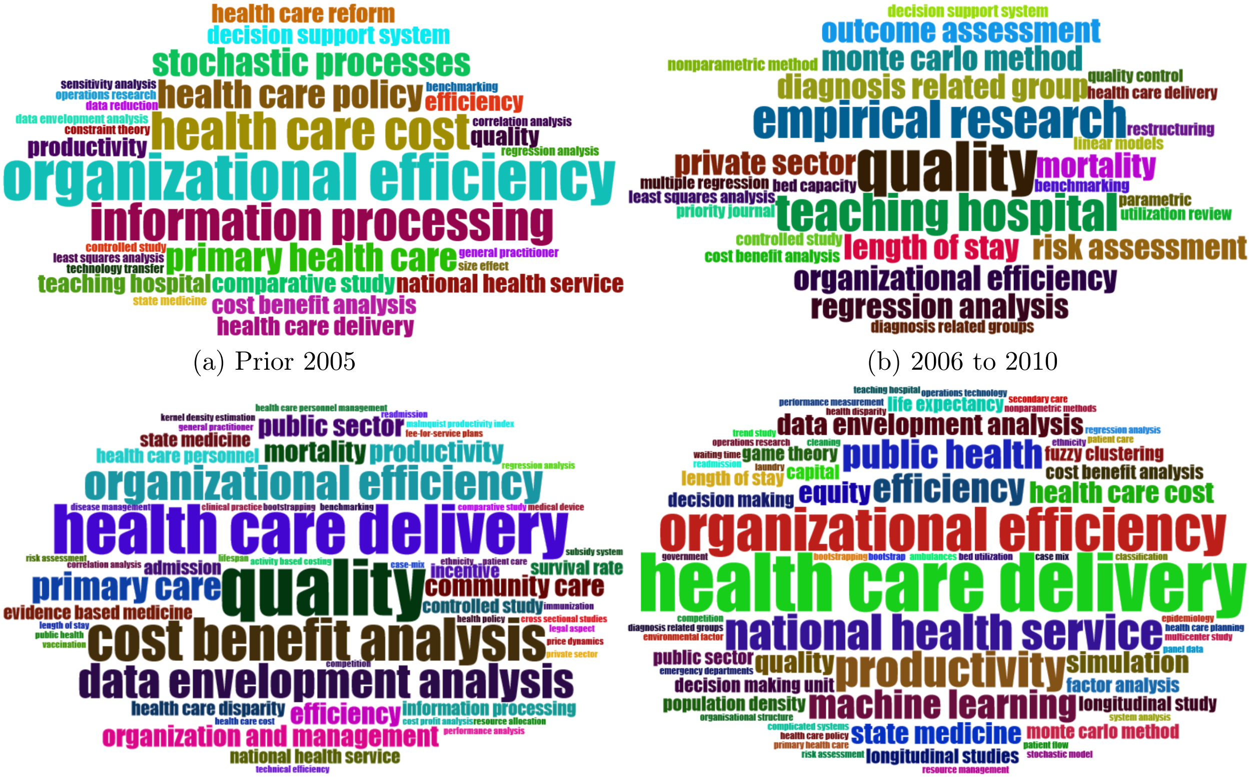
(c) United Kingdom

(d) New Zealand

(e) Hong Kong

Figure 8: Word clouds of keywords by countries and regions

In a time dynamic view of the key topics and methods as shown in Figure 9, the records of keywords from our interested countries and regions were combined and redivided by several periods of years, which are prior to 2005, 2006 to 2010, 2011 to 2015 and 2016 to the first half of 2020. With the same generating algorithm, the word clouds of later years is thicker and more informative, which is due to the fact that most research was conducted after 2010. The research in earlier years apparently focuses more on qualitative discussion about the effects of cost, policy and reform to the efficiency, while the focus later shifted to empirical methods where regression, Monte Carlo, DEA and machine learning gradually appeared more often. “Quality” is a special case, which is discussed a lot in the year from 2006 to 2015, but not in recent years.



(a) Prior 2005

(b) 2006 to 2010

(c) 2011 to 2015

(d) 2016 to 2020

Figure 9: Word clouds of keywords by period

We also checked the sources and authors which are most cited by local authors. For Australia, as summarized in Table 1, Health Economics is not only a journal where local research is frequently published, but also the most cited journal by local researchers. As for the most cited authors listed in Table 2, local researcher Braithwaite is the most cited author, while several global researchers with great reputations in efficiency analysis and the healthcare sector, such as Hollingsworth, Grosskopf, Valdmanis and Färe, are also commonly cited by Australian researchers on this topic.

Another phenomenon worth discussing is the similarity of the most cited authors by researchers in different countries and regions. Most of the top cited researchers in studies of Canadian hospital efficiency are located in the US, such as Grosskopf, Färe and Valdmanis, who are also most cited in the studies of Australia, which may indicate that Australian research closely follows the works from North America. On the contrary, the top cited researchers in the studies of the UK are local experts, such as Street, Castelli and Gravelle. The higher level of independence of research in the UK may be due to some popular and special topics, such as the performance of the NHS. The condition in New Zealand is similar to Australia that local researchers tend to follow the global top researchers in the field, as well as the regional leading authors.

Finally, we collected the global citations of local papers and the top six within each country and region are listed in Table 3. When comparing the total citations of these top cited papers, research in the UK and Canada are more influential in the field than those in Australia, New Zealand and Hong Kong.

For a more dynamic view of the paper production, we plotted the works of top productive authors (in descending order) among all the interested countries and regions in a time series chart. As shown in Figure 10, the label of different colors reflects the author's location, and the scope of time, from 1999 to 2020, covers the publication year of all the selected articles. The parallels connect every work of a certain researcher in our paper pool, where the size of the blue dots represents the number of papers published in each year. For example, Andrews published three papers in 2020 (Andrews, 2020b,a; Jiang and Andrews, 2020), while everyone else published one paper in the same year. Another dimension is the depth of the point, which is determined by the average citations per article per year. The range is from the smallest, which is around 0.6, to the largest, which is around 9.8.

Street and Castelli are the most productive authors who have been productive in the last one to two decades. For Australia, Yong is the author who has been the most productive during the past decade, who is also the most active collaborator locally. The research for Canada concentrates on the period 2008 to 2016. Research in New Zealand is mostly published by Andrews and Jiang in the last two years. However, research in Hong Kong is too few compared to other peer countries, and thus not shown in the plot. In general, most articles were published after 2011, which indicates the same conclusion as was discovered in the time series word clouds plot.

Table 1: Most local cited sources

AU		CA		UK		NZ		HK	
Sources	No. Cited	Sources	No. Cited	Sources	No. Cited	Sources	No. Cited	Sources	No. Cited
Health Economics	13	European Journal of Operational Research	19	Health Economics	17	European Journal of Operational Research	16	European Journal of Operational Research	11
Health Care Management Science	6	Health Care Management Science	16	British Medical Journal	7	Health Economics	12	Health Care Management Science	9
Administrative Science Quarterly	5	Health Economics	14	Health Economics	7	Ministry of Health	10	Omega	8
Medical Care	5	Socil-Economic Planning Sciences	10	Journal of Economics	5	Journal of Econometrics	8	Socil-Economic Planning Sciences	4
Health Services Management Research	4	Journal of Econometrics	8	Journal of Productivity Analysis	4	Health Policy	7	Annals of Operations Research	3
Review of Industrial Organization	4	Journal of Productivity Analysis	8	Public Service Output	4	Journal of Productivity Analysis	5	Central European Journal of Operations Research	3
Journal of Productivity Analysis	3	Health policy	7	Applied Economics	3	Benchmarking: An International Journal	4	European Journal of Health Economics	3
American Economic Review	2	Medical Care	7	Econometrica	3	Medical Care	4	Journal of Medical Systems	3

Table 2: Most local cited authors

AU		CA		UK		NZ		HK	
Authors	No. Cited	Authors	No. Cited	Authors	No. Cited	Authors	No. Cited	Authors	No. Cited
Braithwaite J	9	Grosskopf S	25	Street A	13	Färe R	14	Ozcan Y A	11
Hollingsworth B	6	Färe R	16	Castelli A	9	Cooper W W	11	Grosskopf S	4
Grosskopf S	4	Valdmanis V	14	Gravelle H	9	Charnes A	10	Joe Z	4
Hindle D	4	Ozcan Y A	12	Schmidt P	7	Grosskopf S	9	Seiford L M	4
Maniadakis N	4	Simar L	12	Dawson D	6	Simar L	6	Valdmanis V	4
Valdmanis V G	4	Wilson P W	10	Lovell C A K	6	Ashton T	5	Zhu J	4
Wilson P W	4	Banker R D	7	Newhouse J P	6	Gauld R	5	Cook W D	3
Färe R	3	Charnes A	7	Skinner J	5	Niakas D	5	Liu J S	3
Harris A	3	Cooper W W	7	Hollingsworth B	4	Wilson P W	5	Lu L Y Y	3
Linna M	3	Zelenyuk V	7	Laudicella M	4	Aletras V	4	Lu W M	3
Rosko M D	3	Lovell C A K	6	Parkin D	4	Coelli T J	4	Margaritis D	3

Table 3: Most global cited papers

AU		CA		UK		NZ		HK	
Sources	TC	Sources	TC	Sources	TC	Sources	TC	Sources	TC
Braithwaite et al. (2006)	36	Ouellette and Vierstraete (2004)	88	Jacobs (2001)	164	Davis et al. (2013)	28	Mcghee et al. (2001)	9
Nghiem et al. (2011)	9	Chowdhury and Zelenyuk (2016)	49	Street (2003)	48	Rouse et al. (2011)	6	Guo et al. (2017)	7
Chua et al. (2011)	8	Chowdhury et al. (2014)	35	Giuffrida (1999)	42			Li et al. (2019)	5
Eckermann and Coelli (2013)	8	Liu et al. (2008)	17	Mccallion et al. (2000)	37				
O'Donnell and Nguyen (2013)	7	Milliken et al. (2011)	15	Giuffrida et al. (2000)	16				
Chua et al. (2010)	7	Chowdhury et al. (2011)	15	Omrani et al. (2018)	15				

TC = Total Citations.

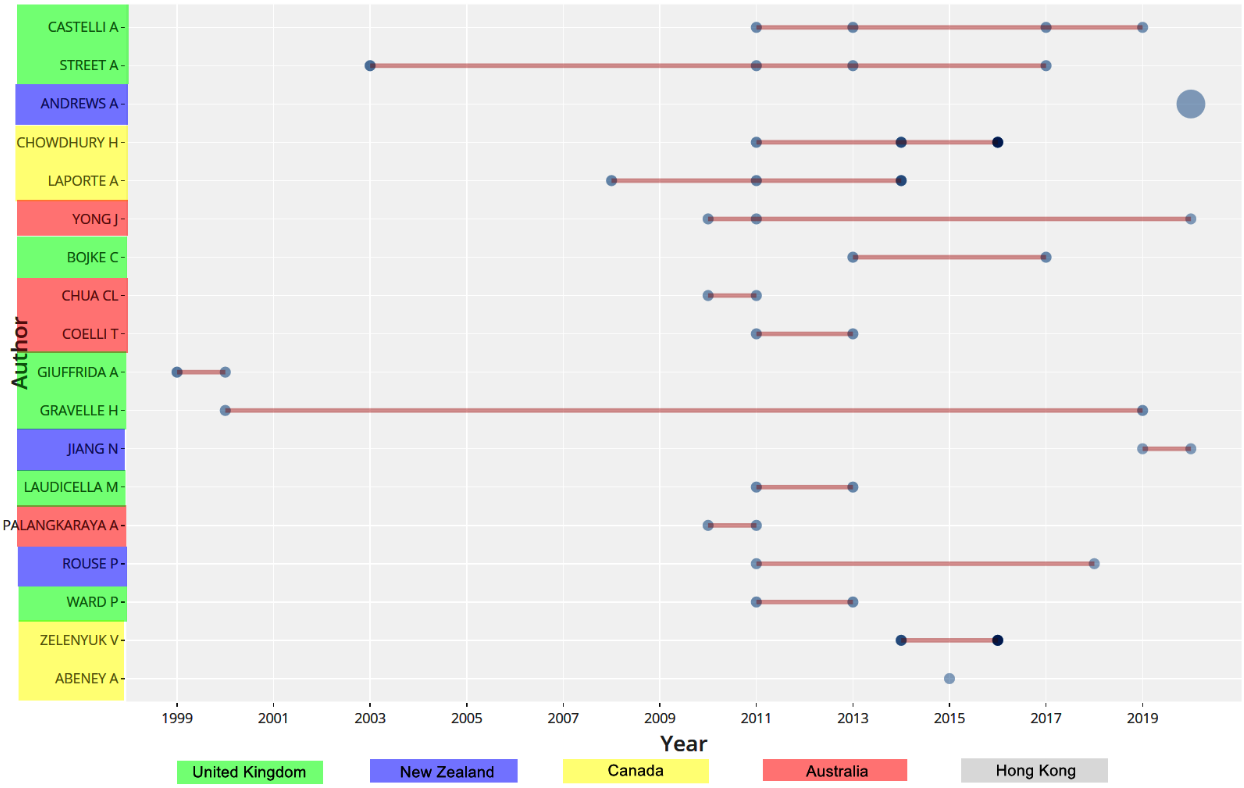


Figure 10: Production of top productive authors over time

Another dynamic trend worth exploring is the evolution of the keywords. Following a similar strategy of term cleaning for the word cloud, we dropped the commonly used terms and combined the records of all the countries and region. The results of co-occurence analysis of the modified keywords by VOSviewer (Van Eck and Waltman, 2010) are shown in Figure 11 and Figure 12. Firstly, the connections between the word frames represent the co-occurence existing between a pair of terms. While the size of each frame reflects the occurrence times of each keyword, the dimension presented by different colors are different in the two figures.

The frame in Figure 11 is colored by the average publication year. Firstly, each year in the publishing period of the whole sample is scored chronologically. For example, the year of the first publication is scored as $S = 1$ and the next year as $S = 2$ and so on. For a certain keyword, the score of publication year S_i is the average of the scores S_{ij} of all the papers which have used it, i.e.

$$S_i = m^{-1} \sum_{j=1}^{m} S_{ij}, \tag{1}$$

where i represents the i^{th} keyword in the sample, m is the number of papers that mentioned this keyword and j indicates the j^{th} paper in these m papers. Finally, the color gradient from a warm color to a cold one is assigned to each keyword based on the scores $S = \{S_i\}_{i=1}^{w}$, where w represents the number of keywords included in the analysis. As a result, the more times a keyword is mentioned in recently published papers, the warmer color the frame would be. Similarly, a more dark-colored frame indicates that the term is mentioned more in the earlier published works. On the other hand, the color of the frame in Figure 12 is an indicator of the average number of citations of the papers using a certain term as a keyword. The color gradient is determined by the citation scores $C = \{C_l\}_{l=1}^{q}$, where q is the number of keywords, l indicates the l^{th} keyword in the sample. The citation score for a certain keyword is

$$C_l = p^{-1} \sum_{k=1}^{p} C_{lk}, \tag{2}$$

where p is the number of relevant papers that used this keyword, k represents the k^{th} paper in these p papers and C_{lk} is the number of citations of the k^{th} paper. Therefore in Figure 12, the papers using warm-colored terms as keywords are cited more times on average than those using dark-colored keywords.

Consequently, the evolution of keywords reveals three main periods of research interests among our interested locations. In the early years around 2005, “cost-benefit analysis”, “information processing” and “teaching hospitals” are the most concerned methods or research objects. Whereas in the period around 2010 to 2015, more developed methods and indexes emerged in the topic. From “organization management”, “scoring system” to “length of stay”, “risk assessment” and “Malmquist productivity index” (MPI), “regression analysis” and DEA. Finally, new terms were introduced in recent years, such as “efficiency frontier estimation”, “bootstrap”, “factor analysis” and “longitudinal studies”. Moreover, the keyword shift over time is similar to that indicated in the previous time dynamic word cloud.

In fact, DEA is a special case that shows up in multiple periods, which further implies a situation that if a term is popular during all the time, the average score of the publication year may indicate it as mostly used in the middle time of the whole period. The keywords with a higher average score mean that they are for the most part recently introduced. Similarly, the keywords having a lower average score means that they are mostly only mentioned in the early years. However, the terms ranked in the middle by the average score of publication year could either be mostly mentioned in the middle period or popular through all the time. Therefore, more cautiousness is needed when interpreting the keywords colored in the middle of the color range.

The results in Figure 12 are clear that papers focusing on DEA, “bootstrap” and MPI are the most cited works, which indicates these methods are the most popular in this topic. When crosschecked with Figure 11, most of the terms which emerged in recent years have the least citations, while the terms “diagnosis related group” and “controlled study”, which are mainly used in the early stage, are frequently cited.

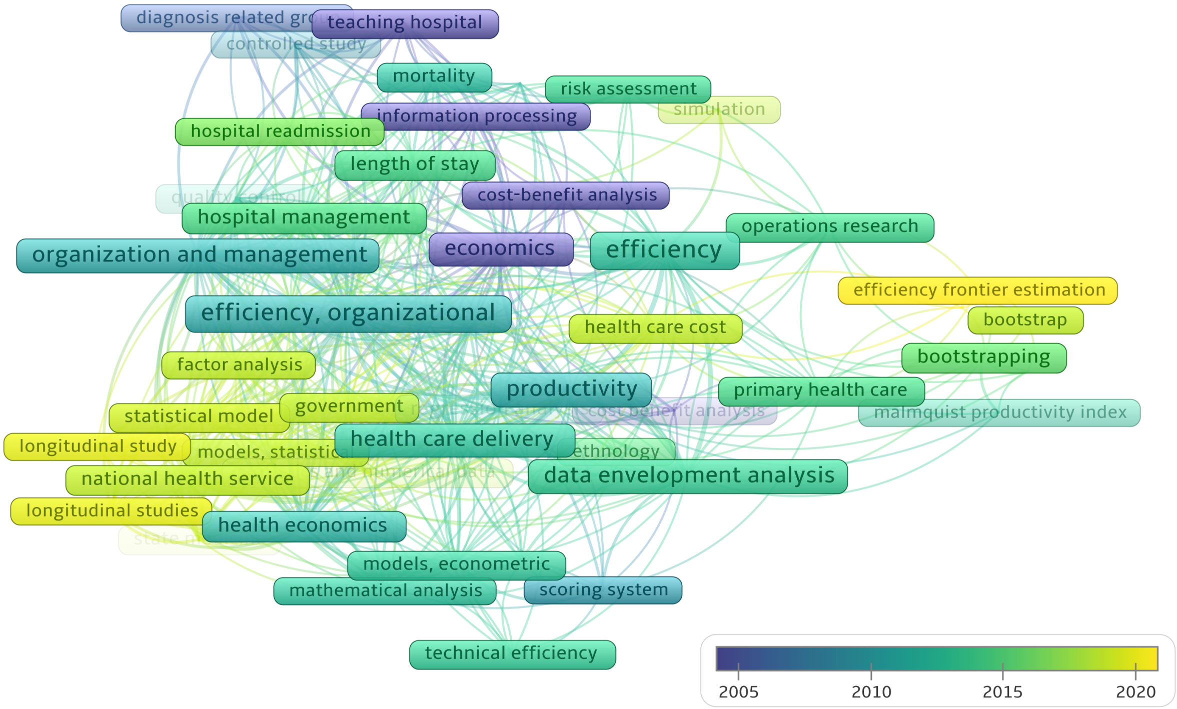


Figure 11: Co-occurrence network of keywords over time

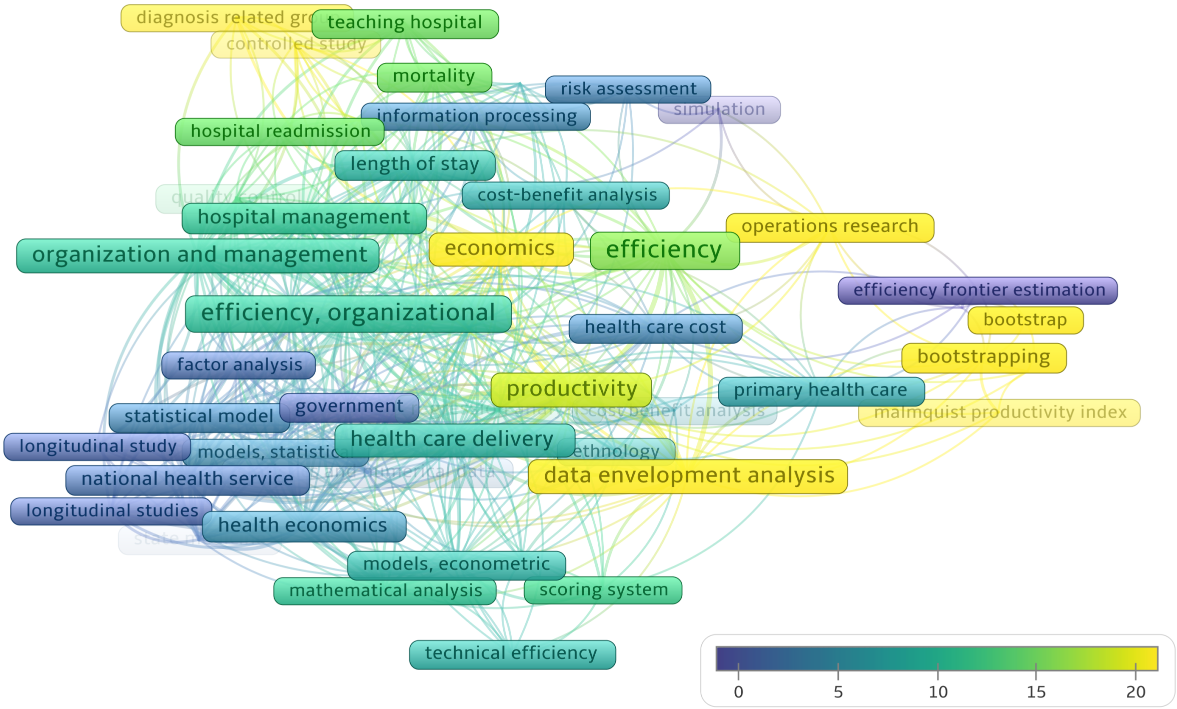


Figure 12: Co-occurrence network of keywords by number of citations

4.3 Influential factors on the impact of research

The influence of research is proof of the value or the contribution of the research by the peers and experts in the specific field. Thus, it is also a significant guide for the literature selection of studies, which is aimed at assembling a broad overview of a research field. The number of citations is frequently used as an indicator of the influence. However, the key challenges are identifying the determinants and interpreting how would these features affect the citations. Regression is one of the most commonly used methods. For example, Kossmeier and Heinze (2019) used regression with the least absolute shrinkage and selection operator (LASSO) to model the number of citations. Machine learning approach, such as random forests, is another popular choice, due to the accuracy in prediction and classification and the advantage in result interpretation.

As a novelty of a systematic review, we constructed a random forests model following Breiman (2001), which has been discussed in Section 3.3, to classify and predict the influence of the reviewed literature. Meanwhile, the impact of the features of a paper on the research influence was analyzed based on the variable importance measurement computed during the modeling process. The classes of the research influence are hierarchical rather than on a parallel level so that the interpretation of the important features is directional (e.g. how to be classified as higher influence research). Therefore, we further analyzed the marginal effect of the influential features on the probability of a paper being classified in the class of the highest influence.

The first question before training is the selection of variables, which could be considered in two aspects, the target variable, which indicates the scientific impact, and the predictors representing the characteristics of the research. Firstly, the citations are commonly used as performance or quality indicators of research for policy making, university ranking and academic hiring (Aksnes et al., 2019). Usually the count of the total citations is used as the expression of the citations. For example, in Kossmeier and Heinze (2019), they considered 21 predictor variables in predicting the citation count of manuscripts, which was used to represent the future scientific impact. Moreover, there are some other commonly used alternative measurements, including the averaged total citations per year (since publication), the total citations in the first three years after publication and the averaged citations by year of the first three years after publication (Wang et al., 2019). Taking an average is used for the consideration of reducing the variation caused by the length of time after publication. Besides, the statistics of the first three years after publication are used in some bibliometric studies to evaluate the research impact in more current circumstances of the research field (Glänzel, 2008). Since most of the papers in our sample were published in the last two decades, the research environment did not change significantly during the period. We chose the averaged total citations by year as the target variable, which is

$$AC = T^{-1} \sum_{t=1}^{T} NC_t, \tag{3}$$

where T is the number of years from the publication to the present and NC_t is the citations of the paper in a certain year. Besides, the averaged total citations over all the time and over the first three years of the sample papers were quite similar, which is also noted in the descriptive statistics in Table 4.

Based on the previous studies with regard to predicting the citations with bibliometric factors, we developed a set of predictor variables, which could be discussed in three aspects, which are "author", "article" and "journal" (Wang et al., 2019; Kossmeier and Heinze, 2019). For the authors, number of authors and institutions are considered, while two dummy variables were also created representing whether there were multiple authors and whether an international cooperation was involved. Besides, the h-index[9] of the first author, the corresponding author and the highest h-index of the co-authors were also included. Some basic characteristics of a paper were used, such as the length of the title, the number of pages and references. Since most articles in the review sample were empirical works, we also collected the sample size used in the research (if introduced) and the corresponding methods. The methods used among the analyzed papers appeared to have high homogeneity so we created a group of dummy variables to evaluate the influence of methodology. The group of dummy variables include DEA, SFA, MPI, TFP, regression, indices of input and output, Cobb-Douglas production function and simulation. For each method, it was used for at least two papers in the sample. Meanwhile, there were only two among the 43 analyzed papers that didn't use any of the listed mainstream methods. In the remaining 41 papers, one or more of these eight methods were applied. The citation count is also usually used for evaluating the influence of a journal. The impact factor (IF)[10] published by the Journal Citation Reports (JCR) and the CiteScore (CS)[11] by Elsevier of the published year were used to reflect the journal factors. Due to the fact that the IF or CS of some papers is not available in their published year, we used them together as a potential complement to each other.

Records of all the reviewed papers with these selected variables were collected from Scopus, while four of the total 43 papers were excluded for further analysis, because they were published in the year 2020 and haven't been cited yet, which may lead to potential bias. More detailed statistics are summarized in Table 4 that show there are about three authors of each paper in our sample, while the average number of institutions is around two. Interestingly, the h-index of the first author seems clearly lower than that of the corresponding author or the highest h-index of the co-authors. However, the interpretation should be

[9] The h-index is the maximum amount of papers that a scholar has published and each of these has been cited at least h times, which is usually used to indicate the influence of a researcher.

[10] The IF counts the citations in the selected year of all the publications in a journal that were published in the two preceding years and divides it by the number of all publications in the two preceding years.

[11] The CS counts the citations in the selected year and the previous three years of all the publications in a journal during the four years and divides it by the number of all publications during the same period.

treated cautiously, because the standard deviation of all the numerical predictors is relatively high. As for the dummy variables, it is clear that most research is conducted by multiple authors but not internationally cooperated. Besides, DEA is the most frequently used method, followed by regression. Another crucial situation of the whole data set is the condition of missing values. As shown in Table 4, the missing values mostly appeared in journal influential factors, which is due to the lack of information. Meanwhile, the other missing values in the h-index is because there is no co-author or corresponding author in some papers. The technique used to deal with the missing values will be discussed later.

Table 4: Descriptive statistics of the variables used in random forests

Predictor (Numerical)	Mean	Std. Dev	Max	Min	NAs
Author					
Number of authors	3.21	1.54	7.00	1.00	0.00
Number of institutions	2.33	1.24	6.00	1.00	0.00
H-inex of first author before publication	7.49	6.93	26.00	0.00	0.00
Highest h-inex of corresponding author before publication	12.79	9.52	34.00	1.00	10.00
Highest h-index of co-authors before publication	16.18	10.72	44.00	1.00	5.00
Article					
Number of pages	14.51	7.48	42.00	4.00	0.00
Number of references	33.46	19.66	88.00	5.00	0.00
Sample size (if)	283.43	249.62	1044.00	24.00	11.00
Number of title characters	85.82	27.61	140.00	23.00	0.00
Journal					
IF (publication year)	2.00	1.24	5.34	0.21	13.00
CiteScore (publication year)	2.75	1.71	5.50	0.30	26.00

Predictor (Categorical)	Positive (1)	Negative (0)	NAs
Author			
Multiple author(s)	34	5	0
International cooperation	9	30	0
Methods			
DEA	17	22	0
SFA	6	33	0
MPI	6	33	0
TFP	3	36	0
Regression	10	29	0
Input/Output indices	7	32	0
Cobb-Douglas production function	2	37	0
Simulation	3	36	0

Citations	Mean	Std. Dev	Max	Min	NAs
Total citation	18.95	30.69	166.00	0.00	0.00
Average total citation by year	2.04	2.49	10.80	0.00	0.00
Total citation in 3 years after publication	4.62	6.40	35.00	0.00	0.00
Average total citation in 3 years after publication by year	1.77	2.33	11.67	0.00	0.00

The random forests is capable of both classification and regression analysis. For the case of the citation count, the more commonly used method is by classifying the papers into hierarchies. The sample could be divided into subgroups, i.e. highly-cited papers (HCPs), medium-cited papers (MCPs) and low-cited papers

(LCPs) (Wang et al., 2019). While the criteria of hierarchy are relative, for example, in Plomp (1990), the papers cited 25 or more times were defined as HCPs. While, in Wang et al. (2019), the papers whose accumulated citation ratio reached the first 20% of the total citation counts of a journal are labeled as HCPs, the papers whose accumulated ratio located in the last 20% were LCPs, and the remaining papers accumulated for the other 60% of total citations were labeled as MCPs (Wang et al., 2019). According to the features of our sample, we sorted the papers by the average total citations per year and labeled them as HCPs, MCPs and LCPs by a ratio of 2 to 3 to 1, which is decided by the distribution in the histogram of the average citations per year.

For the missing values in the sample, one of the commonly used methods is dropping the observations with a missing value in any variable. While another method is filling up the missing cells with 0, or the median or mean of the variable that the missing value belongs to. A method better at prediction performance for random forests is imputing the missing predictor data using the average of non-missing observations weighted by the proximity. The proximity is a ratio of the number of trees that a pair of observations reached at the same terminal node to the total amount of trees, which is a good measurement of "nearness" between two observations in random forests. We used the functional R package "randomForest" by Liaw et al. (2002) for the whole modeling and analysis process and the function "rfImpute" was used for the imputing method with proximity. Finally, the sample was divided into the training and testing set by a ratio of 80% to 20% for independent validation.

Prior to the training procedure, the optimal number of variables of each node is chosen by the out-of-bag (OOB) error rate[12], with the function "tuneRF". The other parameter, the optimal number of trees, is also selected from a wide range of alternatives by the prediction error and the OOB error rate. Consequently, we built the model with the function "randomForest", which came out with a reasonable prediction accuracy both in the training and testing sample.

Rather than predicting the level of citations of a paper, the main purpose is to explore the most influential features with the variable importance obtained through the training process. The result of the ten most influential predictors is as shown in Figure 13. The left plot measures the accuracy decrease of each predictor, which is averaged by trees and normalized by the standard deviation of all the trees. The decreased accuracy of each predictor is reflected by the increase of the OOB error rate of each tree when the certain predictor was moving into the testing set (the OOB portion) compared to the OOB error of the original model. As a result, the more increase of the OOB error in the training step, the more irreplaceable the predictor is. In this case, the h-index of the first author is the most influential that the average OOB error of each tree

[12] The OOB error rate is the mean prediction error of the training sample, which is usually used as a validation of the model. The term OOB represents the remaining part of the training sample that is not chosen during the bootstrap procedure when building the forest.

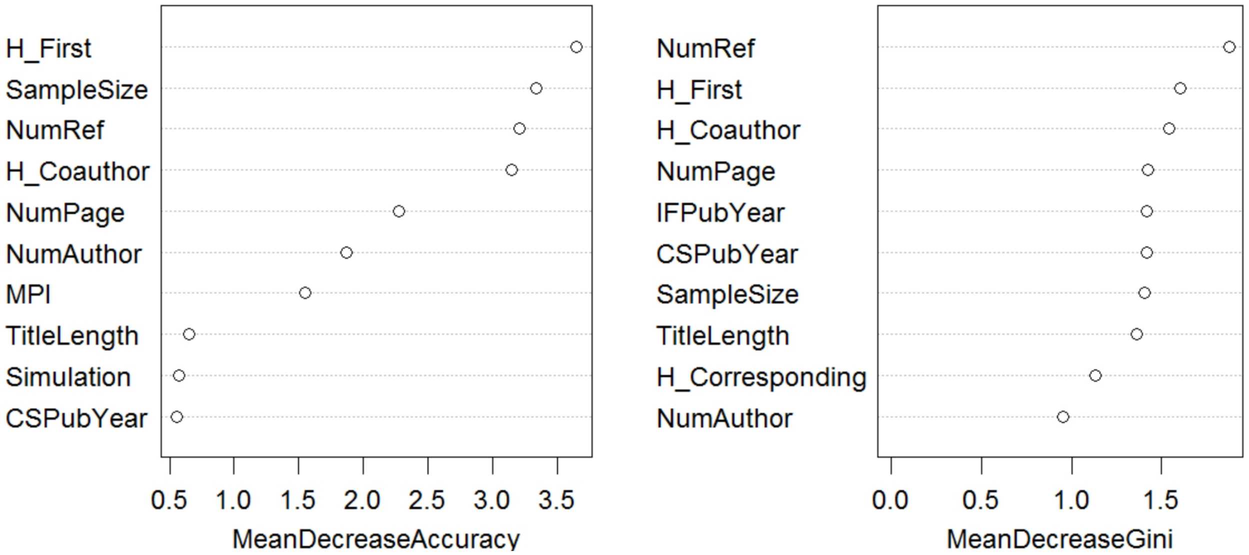
would increase by about 3.6 percent if it was not included in the training set but used in the OOB group for testing. The following important variables are the sample size, the number of references, the highest h-index of the co-authors, etc. On the right side, the decreased Gini index of each predictor was permuted after being averaged by trees. The Gini index is another evaluation of the predictor's contribution to the model accuracy, which actually represents the impurity used to evaluate the power of a classification tree model as discussed in Section 3.3. The Gini index is the probability of a classification tree that a random observation would be incorrectly labeled if the label is randomly chosen based on the distribution of the terminal nodes. Hence,

$$G = \sum_{r=1}^{L} \left(P_r \sum_{u \neq r} P_u \right) = \sum_{r=1}^{L} P_r \left(1 - P_r\right) = 1 - \sum_{r=1}^{L} P_r^2, \tag{4}$$

where P_r is the probability that the random observation is correctly labeled in the region r and L is the number of the final regions[13]. Hence, the number of references, for example, denotes an average drop of around 2 percent points of the mean Gini index over all the trees, which contributes the most among all the predictors. The following most important variables in this sense are the first author h-index and the highest h-index of co-authors, the number of pages and the IF in the publication year, etc.

Though the results of the two measurements are not in accordance with each other, the choice of the top ten predictors are similar. There is no rule of thumb about which measurement is more reliable, but the two results could indicate the power of each predictor in improving the prediction performance or the purity of each tree. Besides, the result of the variable importance is not identical, because of the randomness of the bootstrap procedure during the training. However, the variation of the importance ranking would be relatively small among different trials.

[13] See Breiman et al. (1984) for more details and discussion.



*H_Fist: h-index of the first author, NumRef: the number of references, H_Coauthor: the highest h-index of the coauthors, NumPage: the number of pages, NumAuthor: the number of authors, CSPubYear: the CS of the published year of the journal, IFPubYear: the IF of the published year of the journal, H_Corresponding: the h-index of the corresponding coauthor.

Figure 13: Dot chart of variable importance

Moreover, the more attractive conclusion would be the direction of the influence. Rather than identifying the h-index of the first author to be influential, we are more interested in how the change of h-index affects the level of citations. Thus the marginal effect of the top six important predictors (in mean decreasing accuracy) on the class probability were computed and plotted in Figure 14. The target class is HCPs so that the curve of each predictor reflects the marginal effect of it on the probability (estimated via logit model) of being classified as a highly-cited paper corresponding to its value on the X-axis. Though it might not be reasonable to compare the influence across different variables, it is still a meaningful indicator within each predictor.

For example, the impact of the h-index of the first author to the probability of being a highly-cited paper is negative when the h-index is around 0 to 2. However, it is an increasingly positive promotion when the h-index is approximately between 5 and 15, while the effect of the even higher h-index is positive and somewhat stable. The curve of the number of references shares a similar shape so that the papers with few references would be less possible to be classed into HCPs, while those with more than 20 references would be more likely to be highly-cited in our sample. As for the highest h-index of co-authors, the effect turns out to be positive if it is higher than 14, while the low h-index co-authors would indicate a lower probability of a highly-cited paper. It is interesting that with the increasing number of pages, the effect on the probability of being highly-cited moves from negative to positive and keeps constant during the middle range. However, a total page number of more than the majority drags the marginal effect back to 0. For the number of authors, the paper with a single author is more likely to be highly-cited. And for the sample size, it seems the papers with a relatively smaller sample are more likely to be highly-cited, thus increasing the impact to the literature and the state of knowledge in the field.

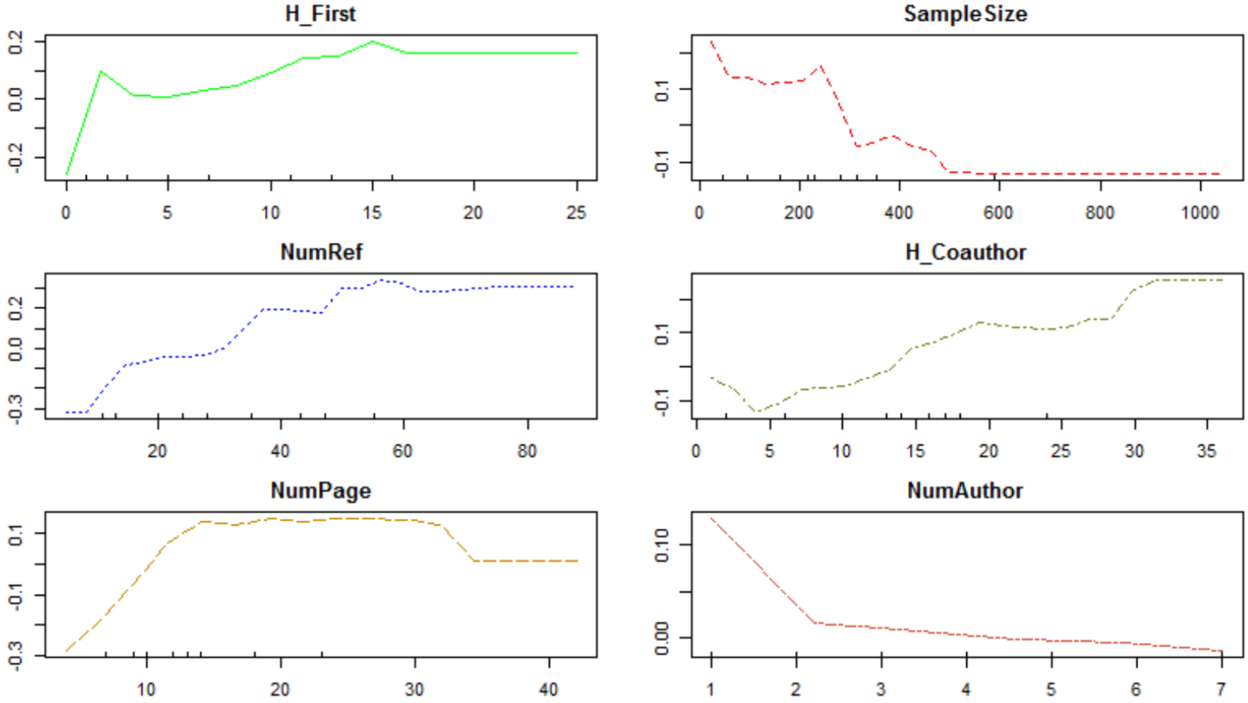


Figure 14: Marginal effect of predictors on the class probability

A limitation of this random forests application is that the number of observations is relatively low compared to the number of variables, which may enlarge the impact of noise during the training. However, the relatively lower OOB error and higher prediction accuracy indicate that the model is useful, fitting the data well, and helpful with the evaluation of the features associated with the impact of papers, as measured by citations.

4.4 Summary of findings of selected papers

Conclusions for the efficiency analysis across our reviewed countries and regions show consistency with a number of aspects as well as contrasting views in some cases. Table 5 briefly summaries some papers with representative findings in our review.

Table 5: Key findings of selected papers

Paper	Data	Method(s)	Key finding(s)
Chua et al. (2010)	Mortality rate in Victoria from 00/01 to 04/05	Two-stage regression	Teaching hospitals and larger local hospitals appear to perform better than others in the sample.
Nghiem, Coelli, Barber (2011)	35 Queensland public hospitals 1996-2004	DEA & SFA	Pure technical change dominates the growth; The number of nurses turns out to be the most influential factor.
Nguyen and Zelenyuk (2020)	15 Hospital and Health Services(HHSs) in Queensland in 16/17	DEA, FDH, and order-alpha-quantile frontier estimators	Some regional hospitals, most of which are small and located remotely, tend to perform at relatively low efficiency levels.
Giuffrida (1999)	English Family Health Service Authorities from 90/91 to 94/95	DEA & MPI	The small improvement in the productivity is mostly explained by the improvement of pure technical efficiency and scale efficiency, but not by the technology change.
Chowdhury et al. (2011)	Inputs and outputs data of hospitals in Ontario over the period 2003-2006	MPI index	The overall productivity and efficiency of hospitals in Ontario was concluded a decrease over the period of 2003 to 2006, while the technological progress achieved an increase of 5.95 percent on average.
Chowdhury and Zelenyuk (2014)	Hospital data of Ontario, 2003-2006	Two-stage DEA and truncated regression	Factors such as occupancy rate, outpatient/inpatient ratio, location, size and teaching status are significantly influential to efficiency.
Andrews (2020)	Quarterly data on DHBs from 2011 to 2017	DEA and truncated regression	A higher proportion of surgical, elderly, and acute inpatients has positive effects on the level of technical efficiency, while the impact of a longer LOS is negative.
Guo et al. (2017)	Public hospitals and institutes in Hong Kong from 2000 to 2013	DEA and Tobit regression	The public hospitals located in a richer district tend to perform at lower levels of efficiency.

The level of efficiency and/or productivity is one of the fundamental aspects of analysis for this topic. The result is usually reported in an aggregated measurement of the studied area. For example, Gabbitas et al. (2009) suggested that improvements of 10 percent in productivity are possible in aggregate for Australian public hospitals. In the report of Productivity Commission (2009), the technical efficiencies of Australian hospitals were about 20% below the hypothesized best practice. With regard to other interested countries in our review, the productivity of Strategic Health Authorities in England was found to be at 5% to 6% above the average of nationwide (Bojke et al., 2013). Meanwhile, the national technical efficiency of New Zealand public hospitals was evaluated at 86 percent on average from 2011 to 2017 (Jiang and Andrews, 2020). In a time dynamic view, multiple studies in the UK and Canada indicated an improvement in productivity and/or efficiency in the research area during the studied period (e.g. Castelli et al. (2011); Giuffrida (1999); McCallion et al. (2000); Valdmanis et al. (2017); Chowdhury et al. (2011)).

A meaningful question for further analysis is to identify the factors of a hospital that determine or impact or are associated with the efficiency scores and how they work. The size of a hospital is one of the most frequently considered factors in the studies for Australian hospitals, where the larger hospitals were mostly found to be more efficient than the others (e.g. Bogomolov et al. (2017); Cheng et al. (2020); Chua et al. (2010); Nguyen et al. (2020); Paul (2002)). While according to another view, the smaller hospitals may be more labor-intensive and perform better in scale economies (Wang et al., 2006). Meanwhile, in the research about Northern Ireland hospitals, the smaller hospitals tended to achieve more progress in productivity during the research period (McCallion et al., 2000). In Chua et al. (2010), teaching hospitals were found to be more efficient than the others, which may also be related to the larger hospital size. However, a higher level of education function in a hospital usually leads to higher costs (Yong and Harris, 1999) and in some analysis, the higher level of education had a negative relationship with the hospital efficiency (Paul, 2002).

The location of a hospital is another significant factor associated with the efficiency (Lavers and Whynes (1978); Chowdhury and Zelenyuk (2016)). The relatively low efficiency of some hospitals in Queensland, Australia was found to be partially explained by the remoteness (Nguyen et al., 2020), while no clear relationship was found between the location of hospitals in New South Wales (Australia) and their efficiency level (Paul, 2002). In the context of Hong Kong, the public hospitals in more affluent districts showed a lower level of efficiency, which may reflect the reality that people in better economic conditions prefer receiving services from private hospitals (Guo et al., 2017).

Some other factors or indicators also appeared to have a significant relationship with efficiency. For example, the higher length of stay (LOS) may negatively impact the efficiency of a hospital (Ali et al. (2019); Andrews (2020b)). The lower level of occupation indicated a positive relationship with efficiency in Paul (2002), while the occupancy rate was inversely related to inefficiency in Yong and Harris (1999). For the effect of the case-mix, in the comparison study by Chowdhury et al. (2014), the productivity and efficiency results of the model with or without case-mix weighted output were significantly different. Meanwhile, the results of the model using case-mix as a separate output were not significantly different from the results of the model without case-mix.

The type of hospital was also considered influential in some cases. A representative study is the comparative research of the operation and performance of public and private hospitals in Australia by Productivity Commission (2009) and its supplement (Productivity Commission, 2010). *Inter alia*, they concluded that the private hospitals showed higher partial productivity[14] of admitted-patient care than the public hospitals. Moreover, they also concluded that the public hospitals provided more non-admitted patient care in the

[14] Due to the lack of data to measure TFP of hospitals, the Commission examined the partial productivity measures by quantifying the output per unit of a single input (Productivity Commission, 2009).

sample while the private hospitals preferred to treat the least morbid patients (Productivity Commission, 2009). As for the technical efficiency, the public contract hospitals[15] performed the most efficiently in both the output and input oriented approaches to efficiency measurements (Productivity Commission, 2010).

5 Conclusion

In this paper, we systematically reviewed the research about hospital-wise efficiency in Australia and its peers, the UK, Canada, New Zealand and Hong Kong. We used the Boolean search in Scopus and manually reviewed the collected papers to construct a paper pool for each country and region. Before a detailed review of the selected papers, bibliometric analysis techniques were applied to explore the field, such as the preference of citation, the productive and influential authors over time and the shift of popular topics. Meanwhile, random forests was conducted to classify the averaged citations per year to reveal the most influential features of an article on its scientific influence.

For each country, there are some clear groups of productive authors as indicated by the Sankey plots. In the perspective obtained by a word cloud of each country, we found a clear difference among the most concerned keywords, for instance, "cost-benefit analysis" and "length of stay" in Australian research, "DEA" in Canadian works, and "public health" in New Zealand. According to analysis of local and global citations, the European Journal of Operational Research and Health Economics are the most cited sources by researchers in all our interested locations. Regarding the most cited authors, researchers in Australia, New Zealand and Canada show similar preference in global top researchers (mostly in US) and local pioneer authors of efficiency analysis in the healthcare sector. As a contrast, UK researchers focus more on domestic works rather than global ones. In another meaningful perspective of the keyword evolution, we used both co-occurence network analysis and time dynamic word cloud and obtained a shift of popular methods and research topics. For example, the most applied methods changed from "cost-benefit analysis" in the early years to "regression analysis", "MPI" in the middle period and to "frontier analysis" and "bootstrapping" in current time. Moreover, DEA has been applied in a wide period, which is also one of the keywords indicating highly cited papers. During the classification of citations by random forests, predictors, such as first author h-index, number of reference and number of pages, were found as the most influential to the citations of a paper. Meanwhile, the influential patterns were revealed through plotting the marginal effect of the predictors on the probability of a paper to be ranked as highly-cited. A summary of these key findings during this review is noted in Table 6.

[15] In this study, some public hospitals were re-classified as public contract hospitals, which are managed or owned by a non-government entity, but are established under legislation or are contracted by the government to provide public hospital services (Productivity Commission, 2010; Forbes et al., 2010).

Table 6: Summary of most important findings from this Review

Key findings	Method	Locations
The productive authors with their interested topics and the commonly published sources in each country: 1. Top productive authors in each country show a clear trend of collaboration. 2. Diversity in popular topics among the countries, i.e. risk assessment and mortility for Australia, NHS and state medicine for the UK, DEA for Canada, DEA and primary health care for New Zealand.	Sankey diagram	Section4.2, Figure 4 to Figure 7
The word cloud of key topics of all the literature in a time dynamic view: 1. The research in earlier years focuses more on qualitative discussion about the effects of cost and policy, which later has shifted to empirical methods such as, regression, DEA and machine learning. 2. "Quality" is discussed a lot in the year from 2006 to 2015, but not in recent years.	Word cloud	Section4.2, Figure 9
The journals that are most cited by the local researchers in each country and region: 1. The most cited journals show a high degree of similarity among the studied countries and regions. I.e. The Health Economics, European Journal of Operational Research, Health Care Management Science, Journal of Productivity Analysis, etc.		Section4.2, Table 1
The authors that are most cited by local researchers in each country and region: 1. Most of the top cited researchers in selected countries and region, except for the UK, are the global top researchers in the field and most of whom are located in the US. Several researchers are frequently cited in multiple countries, such as, Grosskopf, Färe and Valdmanis. 2. As for the UK, the top cited researchers are local experts, such as Street, Castelli and Gravelle.		Section4.2, Table 2
Time dynamic view of productive authors: 1. The most productive authors are identical to the results indicated in the Sankey diagrams. Besides, most of the studies were published in the last decade.	Time-series plot	Section4.2, Figure 10
The shift of interested topics by time by co-occurrence analysis: 1. In the publications around 2005, "cost-benefit analysis", "information processing", "teaching hospitals" are the most concerned keywords. 2. In 2010 to 2015, the focus shifted to more developed methods, such as from "scoring system", "risk assessment" to MPI, regression and DEA. 3. Recently, new terms, such as "bootstrap", "factor analysis" and "longitudinal studies" were the popular keywords.	Network analysis	Section4.2, Figure 11
Most cited topics by co-occurrence analysis: 1. Papers focusing on DEA, "bootstrap" and MPI are the most cited works, which indicates these methods are the most popular in this topic.	Network analysis	Section4.2, Figure 12
The crucial features of a paper that impact the research influence and the patterns of the impact: 1. Based on the importance measurement of the predictors on the mean prediction accuracy and mean impurity of all the trees, the h-index of the first author, the highest h-index of the co-authors, the number of references, the number of authors, etc are the most important features that impact the average citations per year of a paper. 2. According to the analysis of marginal effect, in general, the h-index of authors and the number of references have a positive effect on the number of citations per year, while the sample size and the number of authors have a negative impact.	Random forests	Section4.3, Figure 13, 14

One of the limitations as we mentioned before is that the quantity of the target research in the paper pool is small for each country and region. In contrast to the rapidly increasing and huge amount of papers regarding efficiency analysis of medical techniques and therapies, the efficiency analysis of hospitals is relatively sparse. Consequently, the data set used for bibliometric analysis and random forests is not big enough for their best effects. Another constraint is that the gray literature, though presenting some fruitful conclusions, is not included in the deeper analysis due to a lack of information in the systematic data collection. Moreover, a possible improvement in future study is evaluating the citations when using co-occurence analysis of keywords or as the response variable in random forests. With the text mining approach, for example, the total citations could be weighted by the impact factor of each cited paper, or the cited location within each paper.

Acknowledgments

The authors acknowledge the financial support from the Australian Research Council (from the ARC Future Fellowship grant FT170100401) and the support from their institution. We thank Jongsay Yong, Kok Fong See, Xinju He, David Du, Bao Hoang Nguyen, Hong Ngoc Nguyen and Evelyn Smart for their fruitful comments. We are also thankful for the feedback from other colleagues and participants of workshops and seminars where this paper was presented. These individuals and organizations are not responsible for the views expressed in this paper.

References

Abeney, A. and Yu, K. (2015). Measuring the efficiency of the Canadian health care system. *Canadian Public Policy*, 41(4):320–331.

Aksnes, D. W., Langfeldt, L., and Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *Sage Open*, 9(1):2158244019829575.

Ali, M., Salehnejad, R., and Mansur, M. (2019). Hospital productivity: The role of efficiency drivers. *The International journal of health planning and management*, 34(2):806–823.

Andrews, A. (2020a). Investigating Technical Efficiency and its Determinants: Case of New Zealand District Health Boards. *Health Policy and Technology*.

Andrews, A. (2020b). The Efficiency of New Zealand District Health Boards in Administrating Public Funds: An Application of Bootstrap DEA and Beta Regression. *International Journal of Public Administration*, pages 1–12.

Aragón, A. M. J., Castelli, A., Chalkley, M., and Gaughan, J. (2019). Can productivity growth measures identify best performing hospitals? Evidence from the English National Health Service. *Health Economics*, 28(3):364–372.

Aragón, Y., Daouia, A., and Thomas-Agnan, C. (2005). Nonparametric frontier estimation: a conditional quantile-based approach. *Econometric Theory*, pages 358–389.

Aria, M. and Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, 11(4):959–975.

Australian Institute of Health and Welfare (2020). Health expenditure Australia 2018-19. Technical report, AIHW.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Third international AAAI conference on weblogs and social media.*

Batagelj, V. and Mrvar, A. (1998). Pajek-program for large network analysis. *Connections*, 21(2):47–57.

Beaulac, C. and Rosenthal, J. S. (2019). Predicting university students' academic success and major using random forests. *Research in Higher Education*, 60(7):1048–1064.

Bogomolov, T., Filar, J., Luscombe, R., Nazarathy, Y., Qin, S., Swierkowski, P., and Wood, I. (2017). Size does matter: a simulation study of hospital size and operational efficiency. In *Proceedings-22nd International Congress on Modelling and Simulation, MODSIM 2017*, pages 1274–1280. Modelling and Simulation Society of Australia and New Zealand Inc.(MSSANZ).

Bojke, C., Castelli, A., Grašič, K., and Street, A. (2017). Productivity growth in the English National Health Service from 1998/1999 to 2013/2014. *Health Economics*, 26(5):547–565.

Bojke, C., Castelli, A., Street, A., Ward, P., and Laudicella, M. (2013). Regional variation in the productivity of the English National Health Service. *Health Economics*, 22(2):194–211.

Braithwaite, J., Westbrook, M. T., Hindle, D., Iedema, R. A., and Black, D. A. (2006). Does restructuring hospitals result in greater efficiency-an empirical test using diachronic data. *Health Services Management Research*, 19(1):1–12.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Callon, M., Courtial, J.-P., Turner, W. A., and Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Information (International Social Science Council)*, 22(2):191–235.

Castelli, A., Laudicella, M., Street, A., and Ward, P. (2011). Getting out what we put in: productivity of the English National Health Service. *Health Econ. Pol'y & L.*, 6:313.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). *shiny: Web Application Framework for R*. R package version 1.5.0.

Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377.

Cheng, C. C., Scott, A., Sundararajan, V., Yan, W., and Yong, J. (2020). An Examination of Public Hospital Productivity and its Persistence: An Index Number Approach. *Australian Economic Review*.

Chowdhury, H., Wodchis, W., and Laporte, A. (2011). Efficiency and technological change in health care services in Ontario. *International Journal of Productivity and Performance Management*.

Chowdhury, H. and Zelenyuk, V. (2016). Performance of hospital services in Ontario: DEA with truncated regression approach. *Omega*, 63:111–122.

Chowdhury, H., Zelenyuk, V., Laporte, A., and Wodchis, W. P. (2014). Analysis of productivity, efficiency and technological changes in hospital services in Ontario: How does case-mix matter? *International Journal of Production Economics*, 150:74–82.

Chua, C. L., Palangkaraya, A., and Yong, J. (2010). A two-stage estimation of hospital quality using mortality outcome measures: an application using hospital administrative data. *Health Economics*, 19(12):1404–1424.

Chua, C. L., Palangkaraya, A., and Yong, J. (2011). Hospital competition, technical efficiency and quality. *Economic Record*, 87(277):252–268.

Chua, C. L., Palangkaraya, A., Yong, J., et al. (2008). A two-stage estimation of hospital performance using mortality outcome measures: An application using Victorian hospital Data. Technical report, Melbourne Institute of Applied Economic and Social Research, The University of Melbourne.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., and Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, 63(8):1609–1630.

Commonwealth of Australia 2020 (2020). Budget strategy and outlook, budget paper no 1, 2020-21. Technical report, Commonwealth of Australia.

Davis, P., Milne, B., Parker, K., Hider, P., Lay-Yee, R., Cumming, J., and Graham, P. (2013). Efficiency, effectiveness, equity (E3). Evaluating hospital performance in three dimensions. *Health Policy*, 112(1-2):19–27.

Deng, Z., Jiang, N., and Pang, R. (2019). Factor-analysis-based directional distance function: The case of New Zealand hospitals. *Omega*, page 102111.

Eckermann, S. and Coelli, T. (2013). Including quality attributes in efficiency measures consistent with net benefit: creating incentives for evidence based medicine in practice. *Social Science & Medicine*, 76:159–168.

Färe, R., He, X., Li, S., and Zelenyuk, V. (2019). A unifying framework for farrell profit efficiency measurement. *Operations Research*, 67(1):183–197.

Forbes, M., Harslett, P., Mastoris, I., and Risse, L. (2010). Quality of care in Australian public and private hospitals. In *Australian Conference of Health Economists, Sydney*, volume 14.

Froehlich, P. (2005). Interactive Sankey Diagrams. In *IEEE Symp. on Information Visualization*, page 233.

Gabbitas, O. et al. (2009). Assessing Productivity in the Delivery of Public Hospital Services in Australia: Some experimental estimates-Productivity Commission Conference Paper.

Garfield, E. (2009). From the science of science to Scientometrics visualizing the history of science with HistCite software. *Journal of Informetrics*, 3(3):173–179.

Giuffrida, A. (1999). Productivity and efficiency changes in primary care: a Malmquist index approach. *Health care management science*, 2(1):11–26.

Giuffrida, A., Gravelle, H., and Sutton, M. (2000). Efficiency and administrative costs in primary care. *Journal of Health Economics*, 19(6):983–1006.

Glänzel, W. (2008). Seven myths in bibliometrics about facts and fiction in quantitative science studies. *Collnet Journal of Scientometrics and Information Management*, 2(1):9–17.

Guo, H., Zhao, Y., Niu, T., and Tsui, K.-L. (2017). Hong Kong Hospital Authority resource efficiency evaluation: Via a novel DEA-Malmquist model and Tobit regression model. *PloS one*, 12(9):e0184211.

Hanning, B. W. (2007). Length of stay benchmarking in the Australian private hospital sector. *Australian Health Review*, 31(1):150–158.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health care management science*, 6(4):203–218.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, 17(10):1107–1128.

Hollingsworth, B., Dawson, P., and Maniadakis, N. (1999). Efficiency measurement of health care: a review of non-parametric methods and applications. *Health care management science*, 2(3):161–172.

Hussey, P. S., De Vries, H., Romley, J., Wang, M. C., Chen, S. S., Shekelle, P. G., and McGlynn, E. A. (2009). A systematic review of health care efficiency measures. *Health services research*, 44(3):784–805.

Hyun-do Choi, D.-h. O. (2019). The importance of research teams with diverse backgrounds: Research collaboration in the Journal of Productivity Analysis. *Journal of Productivity Analysis*, 53(1):5–19.

Jacobs, P. and Hall, E. (1994). Key Operating and Financial Ratios for Alberta Hospitals. In *Healthcare Management Forum*, volume 7, pages 19–23. SAGE Publications Sage CA: Los Angeles, CA.

Jacobs, R. (2001). Alternative methods to examine hospital efficiency: data envelopment analysis and stochastic frontier analysis. *Health care management science*, 4(2):103–115.

Jiang, N. and Andrews, A. (2020). Efficiency of New Zealand's District Health Boards at Providing Hospital Services: A stochastic frontier analysis. *Journal of Productivity Analysis*, 53(1):53–68.

Kohl, S., Schoenfelder, J., Fügener, A., and Brunner, J. O. (2019). The use of Data Envelopment Analysis (DEA) in healthcare with a focus on hospitals. *Health care management science*, 22(2):245–286.

Kossmeier, M. and Heinze, G. (2019). Predicting future citation counts of scientific manuscripts submitted for publication: a cohort study in transplantology. *Transplant International*, 32(1):6–15.

Lavers, R. J. and Whynes, D. K. (1978). A production function analysis of English maternity hospitals. *Socio-economic planning sciences*, 12(2):85–93.

Lee, J.-D., Baek, C., Kim, H.-S., and Lee, J.-S. (2014). Development pattern of the DEA research field: a social network analysis approach. *Journal of Productivity Analysis*, 41(2):175–186.

Li, Y., Lei, X., and Morton, A. (2019). Performance evaluation of nonhomogeneous hospitals: the case of Hong Kong hospitals. *Health care management science*, 22(2):215–228.

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22.

Liaw, A., Wiener, M., et al. (2002). Classification and regression by randomForest. *R news*, 2(3):18–22.

Linnenluecke, M. K., Marrone, M., and Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management*, 45(2):175–194.

Liu, C., Laporte, A., and Ferguson, B. S. (2008). The quantile regression approach to efficiency measurement: insights from Monte Carlo simulations. *Health Economics*, 17(9):1073–1087.

Longo, F., Siciliani, L., Moscelli, G., and Gravelle, H. (2019). Does hospital competition improve efficiency? The effect of the patient choice reform in England. *Health Economics*, 28(5):618–640.

McCallion, G., Colin Glass, J., Jackson, R., Kerr, C. A., and McKillop, D. G. (2000). Investigating productivity change and hospital size: a nonparametric frontier approach. *Applied Economics*, 32(2):161–174.

McGhee, S., Leung, G., and Hedley, A. (2001). Efficiency is dependent on the control of supply. *Hong Kong Medical Journal*, 7(2):169–173.

McGlynn, E. A., Shekelle, P. G., Chen, S., Goldman, D. P., Romley, J. A., Hussey, P. S., de Vries, H., Wang, M. C., Timmer, M. J., Carter, J., et al. (2008). Identifying, categorizing, and evaluating health care efficiency measures.

Milliken, O., Devlin, R. A., Barham, V., Hogg, W., Dahrouge, S., and Russell, G. (2011). Comparative efficiency assessment of primary care service delivery models using data envelopment analysis. *Canadian Public Policy*, 37(1):85–109.

Nghiem, S., Coelli, T., and Barber, S. (2011). Sources of productivity growth in health services: a case study of Queensland public hospitals. *Economic Analysis and Policy*, 41(1):37–48.

Nguyen, B. H., Zelenyuk, V., et al. (2020). Robust efficiency analysis of public hospitals in Queensland, Australia. Technical report, School of Economics, University of Queensland, Australia.

Nieminen, P., Pölönen, I., and Sipola, T. (2013). Research literature clustering using diffusion maps. *Journal of Informetrics*, 7(4):874–886.

O'Donnell, C. and Nguyen, K. (2013). An econometric approach to estimating support prices and measures of productivity change in public hospitals. *Journal of Productivity Analysis*, 40(3):323–335.

Omrani, H., Shafaat, K., and Emrouznejad, A. (2018). An integrated fuzzy clustering cooperative game data envelopment analysis model with application in hospital efficiency. *Expert Systems with Applications*, 114:615–628.

O'Neill, L., Rauner, M., Heidenberger, K., and Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences*, 42(3):158–189.

Organisation for Economic Co-operation and Development (OECD) (2019). *Education at a glance 2019: OECD indicators*. OECD Paris.

Organization for Economic Co-operation and Development (2020). Oecd statistics.

Ouellette, P. and Vierstraete, V. (2004). Technological change and efficiency in the presence of quasi-fixed inputs: A DEA application to the hospital sector. *European Journal of Operational Research*, 154(3):755–763.

Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, 10(3):233–240.

Paul, C. J. M. (2002). Productive structure and efficiency of public hospitals. In *Efficiency in the Public Sector*, pages 219–248. Springer.

Penaloza, C. (2010). Healthcare productivity. *Economic & Labour Market Review*, 4(6):63–67.

Plomp, R. (1990). The significance of the number of highly cited papers as an indicator of scientific prolificacy. *Scientometrics*, 19(3-4):185–197.

Productivity Commission (2009). Public and private hospitals. *Research Report, Canberra, Australia.*

Productivity Commission (2010). Public and private hospitals: multivariate analysis. *Suppl. to Research Report, Canberra, Australia.*

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rouse, P., Harrison, J., and Turner, N. (2011). Cost and performance: Complements for improvement. *Journal of medical systems*, 35(5):1063–1074.

Sandiford, P., Consuelo, D. V., Rouse, P., and Bramley, D. (2018). The trade-off between equity and efficiency in population health gain: Making it real. *Social Science & Medicine*, 212:136–144.

Schapire, R. E., Freund, Y., Bartlett, P., Lee, W. S., et al. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269.

Street, A. (2003). How much confidence should we place in efficiency estimates? *Health Economics*, 12(11):895–907.

Team, RStudio (2019). RStudio: Integrated Development Environment for R. Boston, MA: RStudio, Inc; 2016.

Team, Sci (2009). Science of science (Sci2) tool. *Indiana University and SciTech Strategies*, 379.

Tranfield, D., Denyer, D., and Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, 14(3):207–222.

Ujum, E. A. (2014). *Identifying remarkable researchers using citation network analysis.* PhD thesis, University of Malaya.

Valdmanis, V., Rosko, M., Mancuso, P., Tavakoli, M., and Farrar, S. (2017). Measuring performance change in Scottish hospitals: a Malmquist and times-series approach. *Health Services and Outcomes Research Methodology*, 17(2):113–126.

Van Eck, N. J. and Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *scientometrics*, 84(2):523–538.

Wang, J., Zhao, Z., and Mahmood, A. (2006). Relative efficiency, scale effect, and scope effect of public hospitals: evidence from Australia.

Wang, M., Wang, Z., and Chen, G. (2019). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. *Scientometrics*, 119(3):1575–1595.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS quarterly*, pages xiii–xxiii.

Workbench, N., Team, N., et al. (2006). Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan.

Worthington, A. C. (2004). Frontier efficiency measurement in health care: a review of empirical techniques and selected applications. *Medical care research and review*, 61(2):135–170.

Yong, K. and Harris, A. H. (1999). *Efficiency of hospitals in Victoria under casemix funding: a stochastic frontier approach.* Centre for Health Program Evaluation.

Zou, C. and Wang, J. (2018). A Novel Method For Modeling The Large-scale Hospital. In *2018 5th International Conference on Systems and Informatics (ICSAI)*, pages 1231–1234. IEEE.

A Conclusions of reviewed papers

Table 7: Review papers with conclusions

Paper	Topic	Data	Method(s)	Conclusion(s)
		Australia		
Bogomolova et al. (2017)	Hospital size and efficiency	Simulated patient flow by different hypothesized hospital size, mix of arrivals, and diagnosis-related groups (DRG)	Compare indices such as length of stay (LOS) by DRG and occupied rate	1. Smaller hospitals are more likely to have a problem of overcrowding; 2. The function of the hospital, instead of the population that the hospital is serving, should be considered in priority when deciding on the the hospital size.
Braithwaite et al. (2006)	Does changing structure of hospitals results in better efficiency?	20 major teaching hospitals in New South Wales (91/92 to 96/97) and Victoria (91/92 to 95/96)	Cost-efficiency analysis	1. Whether changing the structure or not is not significantly influential to the efficiency of hospitals over the study period.
Cheng et al. (2020)	TFP of public hospitals	Hospital administrative data from Victoria from 07/08 to 11/12	Level and growth of TFP	1. When comparing the level and growth of TFP of hospitals by different size, the larger hospitals perform significantly better than the smaller ones; 2. Variation in TFP is substantial across hospitals, meanwhile the productivity level is highly persistent over time.
Chua et al. (2010)	Deriving a quality indicator for hospitals	Mortality rate from Victoria from 00/01 to 04/05	Two-stage regression	1. Teaching hospitals and larger local hospitals appear to perform better.
Chua et al. (2011)	Relationship between hospital efficiency, quality and competition degree	Admission data of Victoria in 96/97 & 04/05	DEA & Truncated regression	1. The relationship between efficiency and competition is positive when the Hirschman-Herfindahl Index (HHI) is used ,while it becomes negative when the number of private hospitals in competition is used instead; 2. The negative relationship indicates undesirable resource allocation when public hospitals were facing competition with a large number of private hospitals.
Eckermann, Coelli (2013)	Including quality in efficiency measures	45 acute care hospitals in New South Wales	DEA	1. With frontiers, shadow price could be estimated for individual hospital.
Gabbitas and Jeffs (2009)	Productivity measurement within the Australian public hospital system	Public acute care hospitals over the period 96/97 to 05/06.	SFA	1. Appreciable variance exists among the productivity estimations of State and Territory; 2. 10 percent of improvement in productivity may be achievable in aggregate for Australian public hospitals.

Table 7 (Continued)

Paper	Topic	Data	Method(s)	Conclusion(s)
Hanning (2007)	LOS benchmarking in private hospital sector	Data of Australian private facilities in 98/99 to 03/04	Compare LOS of hospitals	1. Data shows a steady increase in private sector same-day (SD) cases and a decrease in overnight average LOS; 2. Overall, the data shows significant variation in LOS parameters between private hospitals.
Nghiem, Coelli, Barber (2011)	Sources of productivity growth in health services	35 Queensland public hospitals 1996-2004	DEA & SFA	1. Pure technical change dominates the growth; 2. The number of nurses turns out to be the most influential factor.
Nguyen and Zelenyuk (2020)	The efficiency of public hospitals in Queensland	15 geographically based Hospital and Health Services(HHSs) in Queensland in 16/17	DEA, FDH, and order-alpha-quantile frontier estimators	1. Some regional hospitals, which are mostly small and located remotely, tend to perform at relatively low efficiency levels.
O'Donnell and Nguyen (2013)	Method to estimate support prices and productivity changes in public hospitals	116 Queensland public hospitals from 1999 to 2004	SFA	1. The levels of productivity and efficiency among Queensland hospitals vary appreciably over the study period.
Paul (2002)	Measurement of efficiency patterns for public hospitals	223 New South Wales public hospitals in 95/96	SFA	1. Hospital efficiency indicates positive relationships with larger scale, lower level of education and occupation and higher capital base; 2. Meanwhile, no clear pattern in terms of rural as compared to urban facilities.
Productivity Commission (2009)	Performance of public and private hospitals	508 Australian hospitals (368 public, 122 private, 18 public contract) in 06/07	SFA	1. On average, the sample technical efficiencies were about 20% below hypothesised best practice.
Wang, Zhao and Mahmood (2006)	Hospital-level inefficiency by hospital size	114 acute public hospitals of New South Wales in 97/98	Stochastic-frontier multiproduct cost function	1. Inefficiency accounts for 9.3% of total costs in large hospitals and 11.3% in small hospitals ; 2. Scale economies appear in very small hospitals, while very large hospitals are affected by diseconomies of scale; 3. Scope effects exist in both large and small hospitals; 4.Small hospitals are more labor-intensive.
Yong, Harris (1999)	Estimate the cost frontier for large public hospitals	35 large Victorian public hospitals in 94/95	SFA	1. Sample shows an average cost inefficiency of 3% of operating expenditure; 2.Teaching hospitals have a significantly higher level of costs than other hospitals; 3.Occupancy rate is inversely related to inefficiency.

Table 7 (Continued)

Paper	Topic	Data	Method(s)	Conclusion(s)
		United Kingdom		
Ali, Salehnejad and Mansur (2019)	Factors explaining variations in hospital productivity	Longitudinal data on English NHS hospital trusts	Unbiased panel regression tree and panel regression	1. Some approaches are significantly effective in improving hospital performance, such as reducing the LOS and increasing the outpatient surgery rate.
Aragón et al. (2019)	Does measure of industry suitable for individual hospital efficiency?	Data for 151 hospitals of 5 years	Compare TFP	1. The commonly accepted approaches for estimating productivity growth of the whole healthcare system are not suitable for analysis at hospital level.
Bojke et al. (2013)	Productivity of the NHS across England	Treatment data in the ten Strategic Health Authorities (SHA) in England in 07/08	Compare the ratio of total 'output' to 'input'	1. Levels of productivity of SHAs vary from 5% above to 6% below the national average.
Castelli et al. (2011)	Indices for NHS productivity	NHS in 03/04 to 07/08	Construct output and input indices to estimate productivity growth	1. During the research period, more treatments with better quality were provided.
Giuffrida (1999)	Efficiency of primary care provision	English Family Health Service Authorities (FHSAs) from 90/91 to 94/95	DEA & MPI	1. A small improvement in the productivity is indicated over the period of study; 2. The increase is mostly explained by the improvement of pure technical efficiency and scale efficiency, but not by the technology change.
Jacobs (2001)	Compare the efficiency rankings from different methods	Data of the UK health cost indices	DEA & SFA	1. It is concluded that there are not large differences in efficiency between NHS hospitals and savings from improving the performance of less productive hospitals would be quite modest.
Lavers and Whynes (1978)	Productivity of English maternity hospitals	193 English maternity hospitals	Cobb-Douglas and log-quadratic functions	1. The numbers of beds and nurses are the most influential factors to the efficiency level of a hospital; 2. The effects of hospital characteristics, such as location and type, indicate significant differences in efficiency levels among hospitals.
Longo et al. (2019)	Doe competition improved hospital efficiency?	Performance data of hospitals from 02/03 to 10/11	Unconditional quantile regression	1. When an additional equivalent rival emerged, there would be a 1.1% increase for the admissions per bed, as well as 0.9% for the admissions per doctor and 0.38% for the proportion of day cases. However, the number of cancelled elective operations would also increase by 2.5%.

Table 7 (Continued)

Paper	Topic	Data	Method(s)	Conclusion(s)
Mccallion et al.(2000)	Productive efficiency of 'larger' and 'smaller' hospitals	Northern Ireland hospitals during 1986 to 1992	Non-parametric frontier approach	1. The smaller hospitals tended to achieve more progress in productivity during the research period; 2. The improvement was attributed to progressive shifts in the best practice frontier outweighing a substantial decline in efficiency.
Street (2003)	Confidence ratio of efficiency estimate for hospitals	Cross-sectional data for English public hospitals	OLS & SFA	1. The estimation of hospital efficiency is sensitive to the modelling decisions so that the point estimates of individual hospitals should be treated cautiously.
Valdmanis et al. (2017)	Performance change in Scottish hospitals	Scottish hospitals from 2003 to 2007	MPI index and regression	1. The regression analysis showed a statistically significant trend of improvement in performance.
Canada				
Abeney and Yu (2015)	Efficiency of the Canadian health care system	Annual data of Canadian hospitals from 2001 to 2010	DEA	1. There were no great provincial differences in efficiency.
Chowdhury et al. (2011)	Productivity measure for hospital services in Ontario	Inputs and outputs data of hospitals in Ontario over the period 2003-2006	MPI index	1. The overall productivity and efficiency of hospitals in Ontario was concluded an decrease over the period of 2003 to 2006, while the technological progress achieved an increase of 5.95 percent on average.
Chowdhury, Zelenyuk, Laporte, and Wodchis (2016)	Does case-mix impact the analysis of productivity, efficiency and technological changes in hospital?	Panel data on Ontario hospitals for the period 2002-2006	MPI, DEA, and non-parametric density estimation	1. Results of productivity and efficiency analysis were significantly different between models with or without case-mix; 2. The results of the model using case-mix as a separate output were not significantly different from the results of the model without a case-mix variable.
Chowdhury and Zelenyuk (2014)	Production performance of hospital services	Hospital data of Ontario, 2003-2006	Two-stage approach: DEA and the truncated regression with double-bootstrap	1. Factors such as occupancy rate, outpatient/inpatient ratio, location, size and teaching status are significantly influential to efficiency.
Milliken et al. (2011)	Productive efficiencies of primary care service delivery	130 primary care practices in Ontario	DEA & regression	1. When evaluated by relative efficiency scores, community health centres performed the worst.
New Zealand				

Table 7 (Continued)

Paper	Topic	Data	Method(s)	Conclusion(s)
Andrews (2020)	Technical efficiency of New Zealand District Health Boards (DHBs)	Quarterly data on DHBs from 2011 to 2017	DEA and truncated regression	1. DHBs in the areas with high socioeconomic deprivation were estimated at a lower level of technical efficiency; 2. The efficiencies of DHBs, which provide secondary hospital services are lower than those of tertiary DHBs; 3. A higher proportion of surgical, elderly, and acute inpatients has positive effects on the level of technical efficiency, while the impact of a longer LOS is negative.
Deng et al. (2019)	Directional distance function (DDF) analysis for New Zealand hospitals	Public hospitals in New Zealand observed during 2011 to 2017	Factor-analysis-based (FAB) approach	1. The average efficiency score is around 90 percent, when the hospital readmission rate was used as an undesirable output.
Jiang and Andrews (2020)	Efficiency of DHBs	Multifaceted administrative hospital dataset of New Zealand public hospitals from 2011 to 2017	SFA	1. During the research period, the average national technical efficiency is 86 percent , while the cost efficiency is around 85 percent.
Sandiford et al. (2018)	Is there a trade-off between equity and efficiency in population health gain?	Life expectancy (LE) changes for 20 DHBs from 2006 to 2013	Stochastic data envelopment analysis and Monte Carlo simulation	1. The opportunity cost of achieving extra gains in equity beyond the point of maximum productive and allocative efficiency is relatively high .
		Hong Kong		
Guo et al. (2017)	Measure the Hong Kong Hospital Authority (HKHA) efficiency changes and the impact of exogenous factors	Public hospitals and institutes in Hong Kong from 2000 to 2013	DEA and Tobit regression	1. The public hospitals located in a richer district tend to perform at lower levels of efficiency; 2. In a sense, this phenomenon reflects the social-economic reality that people with better economic conditions prefer persuing for higher quality services from the private hospitals.
Li, Lei and Morton (2019)	Hospital efficiency of Hong Kong	37 hospitals in Hong Kong in 12/13	DEA	1. The efficiencies of hospitals in the study show a clear difference between the efficient and inefficient hospitals .