How to Present Statistical Comparisons between Swedish Hospitals and Counties

Author: Binyan Xia
Supervisor: Adam Taube

Uppsala University | Statistics Department
2011. 5. 15
Abstract

Background The Swedish Association of local authorities and regions in collaboration with the national board of health and welfare produces the yearly “Swedish Health Care Report” in order to provide evaluations of the hospitals and counties in Sweden for both the politicians and the general public.

Method We describe several standard methods which have been used to present the performance of each hospital or county: Forest plot (FL), League Table (LT), League Plot (LP) and Funnel Plot (FP). Using simulation technique to produce the League Plot of rank is also presented in order to illustrate the unreliable of the ranking principle.

Results The league plot with confidence interval is easily understood by people, but it should provide the total number of operations (sample size) as well. The resulting multiple-indicators system gives a clear overview of the whole system, but the cut-off points used in the traffic light method is not the best choice. Several possible improved methodologies are: A league plot traffic light method and a standard funnel plot traffic light method is recommended when aiming at finding the outliers; A p=0.67 funnel plot traffic light method is suggested when wishing to divide the units into approximately equally large groups; A one-side traffic light method seems to be a wonderful choice when focusing on the bad performance units.

Keywords: Quantitative indicators, Forest Plot (FP), League Table (LT), League Plot (LP), Funnel Plot (FP), Traffic Light method (TL)
Contents

Introduction .................................................................................................................................. 1

1 Theoretical background ............................................................................................................. 1
  1.1 Quantitative indicator ........................................................................................................... 1
  1.2 Variation .............................................................................................................................. 3
  1.3 An indicator applied in this paper ........................................................................................ 4

2 Standard approaches .................................................................................................................. 4
  2.1 The forest plot ....................................................................................................................... 4
  2.2 The league table ................................................................................................................... 5
  2.3 The league plot ..................................................................................................................... 6
  2.4 The league plot of ranks ....................................................................................................... 7
  2.5 The funnel plot .................................................................................................................... 9

3 Over-dispersion .......................................................................................................................... 11
  3.1 The reason of Over-dispersion .............................................................................................. 12
  3.2 The modification of Over-dispersion .................................................................................... 12

4 Further discussions on the “Swedish Health Care Report” ..................................................... 13
  4.1 Comments on the main part of the report ............................................................................ 13
  4.2 Comments on the result part of the report .......................................................................... 13
  4.3 Some possible methodological improvements ...................................................................... 17

5 Conclusions ................................................................................................................................. 20

References ..................................................................................................................................... 21

Appendix I ..................................................................................................................................... 23

Appendix II .................................................................................................................................. 25
Introduction

At present it is a strong political ambition that people should be given “equal” health care no matter where you and therefore the comparisons between different hospitals and regions have become a hot topic. It is quite easy to access voluminous evaluations of hospitals’ performance through the internet and some other public media. These evaluations mostly come from ranking some values such as the mortality or the survival rate after being treated in the hospital. Then a natural question arises: How are these evaluations obtained and are they really reliable?

The aim of this essay is to describe and critically discuss various ways to present and analyse statistical data when conducting comparisons between hospitals or geographic regions. Since 2006, the Swedish Association of local authorities and regions (SALAR) in collaboration with the national board of health and welfare has published 10 open comparison reports which concern healthcare and medical care, care and elderly care, comprehensive school, senior high school, safety and security, public health, etc. The source of inspiration of this paper is one of these open comparison reports----The yearly Swedish Health Care Report\(^1\) which concerning the quality of various units in the Swedish health care system. Taking one indicator (the proportion of re-operations after the first hip replacement surgeries) as example, several standard methods which are widely used for comparisons of hospitals’ performance are presented first. Then we discuss the “over-dispersion” phenomenon, and give some critical comments on this Swedish Health Care Report. At last, the paper provides several possible improved methods for further study.

1 Theoretical background

Which kind of indicators do the statisticians apply when they are performing a hospital comparison? Why do not different hospitals perform exactly the same? Before this paper presents the standard approaches, we will firstly provide some background knowledge to the audiences about two key words, namely “quantitative indicator” and “variation” and the specific indicator we will present.

1.1 Quantitative indicator

“Quantitative indicators” are important and useful tools for monitoring and evaluating progress in the health care systems. Statisticians enjoy using routinely collected

\(^1\) Öppna jämförelser av hälso- och sjukvårdens kvalitet och effektivitet – Jämförelser mellan landsting 2010.
statistical data to construct “indicators” in order to measure the outcome of health care which is latent variable and cannot be described directly. Typically, “indicators” such as mortality after treatments, the time waiting for treatments and the proportion of re-operations after first operations are focusing on reviewing the hospitals’ performance.\textsuperscript{2} The identification and use of suitable indicators is also a central issue. A good indicator should be not only accurate but also presented in a proper way that does not result in any unfair criticism or unjustified praise. Different researchers have various view of what an “indicator” is, while there is one general agreement that a single “indicator” does not describe the whole situation. In this report, a multiple-indicators system is introduced to give an overview of the Swedish health care system, however, even applying multiple indicators, we still cannot make a description of the whole system. This paper will present the standard approaches using an indicator selected from the “Swedish Health Care report”: the proportion of hip re-operations after the hip replacement surgeries in two years at different Swedish hospitals and counties during 2006-2009.

### 1.2 Variation

It is crucial to study the actual variation in order to learn something about the various units such as hospitals or counties. The variation in the health care system can be explained by a number of major categories:\textsuperscript{3}: 

1. Different types of patients (“patient mix”)
   e.x: Different sex distributions  
   Different age distributions  
   Different health level of the patients  
   Different job distributions
2. Different quality of health care provided 
   e.x: Various traditions at the different hospitals  
   Various skill levels among the operating doctors  
   Various technical levels of the different hospitals
3. Different environment 
   e.x: Different air pollution level  
   Different radioactive elements level
4. Different data collection methods
5. Chance


Thus, there are a number of known, as well as unknown, possible explanatory background factors which might influence the variation of the final results. The part of the variations which cannot be explained by different background factors, are traditionally lumped together under the name of random variation. One thing needs to be clarified here is that this “random” variation does not always follow a well defined random distribution.

Even if a hospital is exactly the same during two or more years, the doctors are the same and the treatments are the same etc there will not be exactly the same outcome for different years in the given Swedish health system report. This means that there is variability over time and the results can be considered as a realisation from a certain process over time.

Different researcher can work more or less intensively with the eliminating the effects of various background factors. It is easy to take the sex and age distributions into account, but it is much more difficult to consider the various job distributions, depending upon the non-availability of statistical data.

1.3 An indicator applied in this paper

What is a hip replacement surgery? Hip replacement surgery is using a prosthetic implant to replace the damaged hip joint. Generally, there are two kinds of hip replacement: Total replacement (Figure 1) and Hemi (half) replacement. Hip replacement is one of the most successful orthopedic surgeries and it is quite common. Of course, the artificial implant can wear out over time, so health care providers put great effort on modifying the hip replacement implants in order to get a long-lasting
result. So how long will it last until another replacement surgery is highly concerned by the patients.4

The indicator in this paper is the proportion of re-operations within 2 years after the first total hip replacement surgeries which is under the Musculoskeletal disorder (MSDs) category. The denominator is the number of re-operations within 2 years after hip replacement surgery and the nominator is all operations of hip replacements surgeries which have been registered in the Swedish register system 2005-2008.5

In the data collection process, only complications leading to surgeries are included. And if one patient has had several surgeries, it will be counted as one. If a patient did his or her re-operation in a different hospital or county, then it will be counted into the original hospital or county where he or she did the first surgery. In this medical follow-up case, there are patients who might do a hip re-operation but not in this time period, and there can also be patients who died or moved away from Sweden during this time period. Exactly how the actual indicator and its confidence interval have been calculated, we have not been able to find out.

2 Standard approaches

Generally, there are four standard approaches which are widely used in the comparisons among health care systems: the forest plot of proportion, the league table of proportion and its confidence interval, the league table of rank and its confidence interval and the funnel plot of proportion. As the aims of the researches can be different, the researchers can also present these methods from different aspects. In this part, we will present the data from two different situations, taking hospitals as units and taking counties as units. The “Swedish Health Care Report” supplies the data: the proportions of re-operations within 2 years after hip replacement surgeries during 2006-2009.

2.1 The forest plot

Figure 2 shows a simplified version of a forest plot or a “caterpillar plot” which is quite similar to the ones used in meta-analysis. But we will not present meta-analysis in this paper. In this forest plot, each point presents the value of hip re-operation


5 Öppna jämförelser av hälso- och sjukvårdens kvalitet och effektivitet – Jämförelser mellan landsting 2010 Appendix II.
percent for each hospital, and each line shows the confidence interval for each hospital. From this forest plot, we can get some general statistical information of each hospital. Then, we find that the hospitals show great variations, and the confidence intervals of different hospitals cover the proportion region from 0.00 to 7.90. If some particular standard value can be obtained, the patients can compare the hospital they intend to go with this value and get a simple evaluation themselves but it is possibly wrong.

![Forest plot of hip re-operation ratio in Sweden, 2006-2009](image)

Figure 2: The forest plot of hip re-operation case in Sweden, 2006-2009

In this diagram, the random variation is described by means of 95% confidence intervals for each proportion of re-operations at different hospitals. We notice that the smaller the total number of operations is, the wider the confidence interval will be.

### 2.2 The league table

League tables are well known for comparing sport teams, schools or companies by ranking their achievements and abilities all over the world. Nowadays, it is also widely used by the health care system in the production of hospital comparison reports. The league table introduces an order between the units from the “best” to the “worst”. The readers get a direct intuitive feeling for the results and often the ranking obtains a great public interest. It is easier for the reporters to tell the patients and the governments: which one is the “best” and which one is the “worst”. This is why league tables are so popular in the reports for all kinds of results. Statisticians also provide 95% confidence intervals of the percents (possibly risk adjusted) because the random variation might distort the results.
The journalists must learn that as far as there is variability, differences between the units will exist even if there is no quality difference between the unit ranked as the “best” one and the unit ranked as the “worst” one.

Table 1: League table of hip re-operation percent (part) 2006-2009

<table>
<thead>
<tr>
<th>Hospital</th>
<th>Percent(CI)</th>
<th>Rank</th>
<th>No. Of re-operation</th>
<th>No. Of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vaxjo</td>
<td>0.00 (0.00, 0.00)</td>
<td>1</td>
<td>0</td>
<td>504</td>
</tr>
<tr>
<td>Ystad</td>
<td>0.00 (0.00, 0.00)</td>
<td>2</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Falkoping</td>
<td>0.20 (0.00, 0.49)</td>
<td>3</td>
<td>2</td>
<td>981</td>
</tr>
<tr>
<td>Sundsvall</td>
<td>4.05 (2.46, 5.63)</td>
<td>77</td>
<td>24</td>
<td>593</td>
</tr>
<tr>
<td>Gävle</td>
<td>4.20 (2.56, 5.85)</td>
<td>78</td>
<td>24</td>
<td>571</td>
</tr>
<tr>
<td>Sunderby</td>
<td>4.87 (2.06, 7.68)</td>
<td>79</td>
<td>11</td>
<td>226</td>
</tr>
</tbody>
</table>

Table 1 is a part of traditional league table which includes the institution’s names, the ranks, the ranking standard—hip re-operation percent and its confidence interval, the number of re-operations and the total number of hip replacement surgery. 79 hospitals are ranked according to their hip re-operation percent. As already mentioned, it is a considerable variability between the hospitals. Even if we realize that a confidence interval contains all possible hypothetical values of the actual parameter which can not be discarded on the basis of available data, we notice that there is no proportion which is covered by all the different intervals. Thus there must be significant differences between the hospitals. In general this can be expected since the patients at the different hospitals are indeed not random samples from the population. But does it mean that the high ranking hospitals are “bad hospitals”?

2.3 The league plot

A league plot of hip re-operation percent for hospitals generated from the league table shown above in Table 1 is presented in Figure 3.a and similarly Figure 3.b can be obtained. In this hip re-operation case, a weighted mean of the whole population which equals to 0.0179 is calculated first as a target line. Each bar represents the percent of re-operations operated patients for hospital or county per hundred in 2 years during 2006-2009 and it is visually quite obvious that there is a considerable variation between the hospitals. The intervals at the right head of the bars give a 95% confidence interval for each hospital. We can identify the “outliers” from the league plot by searching the ones whose CI do not cover the target line. The smaller the hospital’s patient size is, the larger the confidence interval is. Thus, the hospitals’
sizes play a key role in this type of hospital comparisons. Taking ‘Helsingborg’ as an example, from the league table, it might belong to the “bad” hospitals by the public due to the high ranking, but its CI covers the target line. Actually, people cannot say this hospital is worse than the other hospitals. The league plot shows a clearer pattern than the league table.

Figure 3.a: The league plot of hip re-operation percent for different hospitals in Sweden, 2006-2009

In Figure 3.a, Karlstad, Solna, Mölndal, Sundsvall, Gävle and Sunderby are the “outliers” with high rank. As the confidence level is 95%, there would be “4 outliers” (∼79 × 0.05) in a normal system for a population of 79. Nevertheless, there are 20 “outliers” in the whole system which highly exceeds the expect value. Actually, it shows “over-dispersion” which will be described more specifically in Part 3. On the other hand, five outliers will be found in Figure 3.b which is also more than the expect number of 1 (∼21 × 0.05). Thus, “over-dispersion” appears in both of the league plots.

2.4 The league plot of ranks
In the league plot of median rank (Figure 4.a and Figure 4.b), the units are also ordered according to their rank which means that each hospital or county has the same rank value as in the league plot of percent. The median is chosen in this plot for it is not affected by the “outliers” in the system. Obviously, these rank values are open for considerable random variation, therefore it is interesting to give a confidence interval for each rank value.

Bradley Efron (1979) “invented” bootstrap since variances, tests and asymptotic confidence interval can be misleading when the sample size is not large enough. We can use bootstrap simulation to estimate the median and rank the hospitals by ranking their bootstrap medians. We can also calculate the confidence interval of the median’s rank given the bootstrap ranks, and then plot the result against the original rank as in Figure 4.a and Figure 4.b. First, we generate 1000 replications for each hospital according to the approximate binomial distribution and resample them with replacement. Second, we rank the new samples and get 1000 bootstrap ranks for each hospital. Third, we calculate the bootstrap samples’ median for each hospital and rank the medians. At last, we calculate the 95% confidence interval for the rank of the medians based on the 1000 bootstrap ranks. Actually, the 1000 replications reflect possible observations under the same situation.

![Figure 4.a: League plot of median rank and its CI for hospitals in Sweden, 2006-2009](image-url)
In Figure 4.a, there are a lot of confidence intervals for the hospitals’ ranks which are quite wide which means that in the same situation these hospitals can rank quite differently. Figure 4.b shows the same result. Thus it is still a debatable question whether we can rely on the ranks. Concerning this point, some statisticians have extremely critical opinions that these ranking methods are presented without any consideration of random variability.

2.5 The funnel plot

When the sample size of each unit is different, a funnel plot is an alternative way of presentation, proposed by many statisticians such as Spiegelhalter. A funnel plot presents the observed indicators against its precision such as sample size. Thus when the units have exactly the same sample size, the funnel plots lose their meaning. Traditionally, it contains three parts:

1. A scatter plot of all the units (e.x: Swedish hospitals) with the sample size (e.x: patients numbers) as X-axis and the indicator (e.x: the percents of hip re-operation) as the Y-axis.

---

2. A horizontal target line in the middle equals to the target value mentioned in the league table part.

3. Two “control” curves and two “warning” curves. The “Special cause method” holds that there are two kinds of variability: ‘common-cause’ variability and ‘Special cause’ variability. A threshold of 3 standard deviations being commonly used as a demarcation between the “in-control” hospitals which are assumed to be subject to the former variability and “out of-control” hospitals on the contrary. So the control curves are the 3σ confidence interval of the target and the warning are the 2σ confidence interval.

From a purely statistical point of view, the funnel plot is a graphical representation which tests whether each value $Y_i$ belongs to the given distribution.

$$H_0: Y_i = \theta \quad H_1: Y_i \neq \theta$$

$$Z = \frac{Y_i - \theta}{\sigma/\sqrt{N}}$$

The empirical distribution of the observed hip re-operation data follows a Binomial distribution approximately. There are two possible methods to draw the control limits: a normal approximation and an exact formula in the cases of discrete distribution such as Binomial distribution and Poisson distribution. Practical research results show that if the sample size $N>100$, the approximated curves and the exact curves coincide. As most of the sample sizes are larger than 100, we can use a normal approximation method to obtain the control curve as follow:

$$y_p(N) = \theta \pm z_p \frac{\sigma}{\sqrt{N}}$$

, where $y$ is the indicator value; $\theta$ is the target value; $p$ is the P-value; $z_p$ is $\Phi(z_p) = p(Z \leq z_p) = p$ for a standard normal variate $Z$, e.x: $z_{0.025} = -1.96$; standard deviation $\sigma = \sqrt{\theta(1 - \theta)}$ and $N$ is precision parameter—number of patients.

---

Figure 5.a is a funnel plot of the hip re-operation percents for the hospitals and Figure 4.b is for the counties. The first funnel plot clearly identifies high-percent low-volume hospitals as well as low-percent low-volume hospitals. The control limits of funnel plot will be wider for the lower volume hospitals. If we do not take this into account, the lower volume hospitals are more likely to be ranked misleadingly and unfairly to the “top” or “bottom” group.10 Funnel plot and other statistical control process measures identify the “special cause” variations described in the “Special cause method”. The hospitals which fall outside the control limits are “outliers” and need more discussion by the researchers. In Figure 5.a, Mölndal, Sundsvall, Gävle and Sunderby are the “outliers” which are found above the upper control limit. The funnel plot also shows slightly “over-dispersion” of this hip re-operation data as there are six “outliers” instead of the expect value 1. Figure 5.b also shows some “over-dispersion” due to 2 counties falling outside the control limits.

3 Over-dispersion

As the league plot and funnel plot shows, over-dispersion phenomenon appears in this hip reoperation proportion system.

---

3.1 The reason of Over-dispersion

“Over-dispersion” is a consequence of greater variability in a data set than can be expected on the basis of a given simple statistical model. The variance function is defined as:

\[ \text{Var(response)} = \varnothing \cdot V(\mu) \]

It explains how much the response variable relies on its mean.

Generally, the error distributions are defined as:

1. Normal distribution: \( V(\mu) = 1, \varnothing = \sigma^2 \);
2. Binomial distribution: \( V(\mu) = \mu(1 - \mu), \varnothing = 1 \);
3. Poisson distribution: \( V(\mu) = \mu, \varnothing = 1 \);

As we see, both the response variance of Binomial distribution and the Poisson distribution depend totally on their means. In such case, it describes the reason for over-dispersion phenomenon in these distributions. For example, in this proportion case, the binary observations (the re-operation cases and the total operations) seem to be correlated with each other. Because of this non-independence, the variance of observations can be greater or less than the assumption.\(^{11}\)

When a funnel plot has a 95% limits, it is to be expected that five percents of the units fall outside these limits or being “outliers”. The same line of argument is true also for league plots. According to the league plot of percents and the funnel plot of percents for hospitals and counties concerning hip re-operation case, the systems both show “over-dispersion”.

3.2 The modification of Over-dispersion

There are lots of strategies to accommodate this phenomenon and plenty discussion concerning it also.\(^{12}\) Some researchers hold that instead of comparing the small units, we can cluster them into more homogenous groups. And other researchers even introduced a different target. They recommend the usage of an interval as a target rather than a weighted mean value. Some other researchers gradually remove the values of “special” reason in order to reduce the variation and make the “outlier” shift down into the normal range. For example, in the cancer mortality report of South-East England, the reporter stepwise exclude breast cancer for women and lung cancer for

---


men to investigate the over-dispersion reason.\textsuperscript{13} There is also another way that we can change the value of the dispersion parameter $\varnothing$ according to the degree of freedom by modelling the data as: $\log \frac{p_i}{1-p_i} = \alpha_i + \beta \log n_i$.

4 Further discussions on the “Swedish Health Care Report”

The “Swedish Health Care Report” which gives a detailed view of the Swedish health care system contains 134 indicators, produced by Swedish Association of Health and Welfare. The report is aimed at giving an overview of the indicators using a format readable for both the politicians, the administrators and the general public.

4.1 Comments on the main part of the report

In this report, each paragraph starts with a brief introduction of each indicator which contains both medical explanations and background knowledge. These texts help the readers to understand the meaning of the indicator. Then, the researchers provide a time series graph with curves for women, men and total respectively from 2000 to 2009 which gives an overall impression of the development for the indicator. Thus it is easy for the researchers to find out if there is a time trend for a specific indicator. In general, the trends for men and women are quite similar. Then, two league plots combined with confidence intervals are provided. In these two diagrams, women and men are plotted separately for each county both for the time period 2005-2008 (if available) and 2006-2009. It can be calculated that there are high correlations between women and men for some indicators, but lower correlation between them for some others. This is related to the indicator’s character which will not be discussed in this paper. Finally, a league plot for each hospital is provided for more detailed discussion.

In these league plots, nothing is mentioned about the sample size at all, although the indicator value and its confidence interval for each unit is provided. If there are some researchers who are interested in this study, they can hardly conduct in-depth studies based on the data in this report only. We have had great help from Mr. Thomas Fröjd in obtaining more detail information.

4.2 Comments on the result part of the report

After presentation and discussion about each indicator separately, the report ends with an effort to summarise the results concerning all the counties and indicators. A multiple-indicator panel makes it possible to overview the whole system instead of being limited by one single indicator.

In this part, a rather primitive method--the “Traffic light method” (TL) is applied to present the counties’ performance. (See Appendix 1) Each row presents an indicator and each column stands for a county. According to the ranking of the indicator, the counties are divided into three equally large groups. Thus, each group contains 7 counties for each specific indicator. Each grid is coloured with three different colour based on the following rule: The high ranking group is presented by red light which means that these counties might be high risk units and need to be improved; The low ranking group is presented by green light which indicates that these counties are low risk and might be recommended; And the middle group is given with yellow light.

The “Traffic light method” is widely used in the risk management. In a TL method, two sharp cut-off points are set in order to divide the units. Each unit is given with a single traffic light according to their location: red, yellow or green. It is important to choose appropriate red-yellow boundary line and yellow-green boundary line in the traffic light representation. Generally, there are two kinds of boundary line selection methods: the fixed-cut and the percentiles.

The fixed-cut might be “golden value” coming from some empirical results or some worldwide standards. Applying this fixed-cut traffic light method to the hip re-operation case, two cut-off points should be selected (e.x: R/Y=0.2 G/Y=0.1) in advance. Then, the counties will be presented with the different colour traffic lights based on the results of comparing them with these two fixed-cuts. In principle, it might even happen with this method that all the counties are coloured with red or green. So in this fixed-cut traffic light method, group size can be different from each other and a colour distribution shift can also be detected over time. We notice that in this approach, it is in principle also possible to make comparisons with similar data from other sources.

Instead of setting “fixed-cut”, a percentile boundary of time series is widely used in the fishery management of describing the stock conditions in order to achieve sustainable development. (Red: negative for fishing; Yellow: intermediate for fishing; Green: positive for fishing) 33rd percentile and 66th percentile of the time data series
are set to be the boundary lines of each indicator in the fishery system. So each group is supposed to have approximately the same size.\(^\text{14}\)

In the “Swedish Health Care Report”, the counties are divided into three equal large groups (7 counties for each group), so it equivalents to setting the 33\(^{\text{rd}}\) percentile and 66\(^{\text{th}}\) percentile of cross-section data as the boundary lines for the traffic light method. Therefore, this method can be named as “observed traffic light method”. This method might be suitable for performing comparisons inside the one system, but it is hard to make comparisons between different systems. The resulting multiple-indicators panel aimed at giving a summary of the whole report, however, has several shortcomings:

First, a “traffic light method” might lose some important information. There are only two sharp cuts but no smooth transition among the units in this system. So the meaning of the traffic lights close to these cuts can be very sensitive. It may give the same colour when the difference between units is significant and also give different colours even when the results are nearly identical, and the difference between the “best” and “worst” can be sometimes huge and sometimes tiny. For example, we know that this hip replacement re-operation proportion ranges from 0 to 0.79 which is quite wide. But if the same indicator happens to be range quite narrow from 0.10 to 0.20, it will show almost the same result to the public using this traffic light method. So the real meaning of the light can be confusion.

Second, it can be quite misleading to divide the hospitals or the counties completely due to their values of indicators without any consideration of the sample size. As it appears in the funnel plot, the units with a smaller sample size can easier happen to be on the top or the bottom of the ranking, so ignoring the importance of the sample size may cause an unfair evaluation of these smaller units. In the report, the approach is commented upon in the following humble way\(^\text{15}\):

1. The comparisons only aim to imply the demands of improvement for some hospitals or counties in specific field but not really mean that any of them is “best” or “worst”;
2. The ranking method is not absolute but rather a signal that the result should be further discussed;
3. And the approach with different colours is a part of the signal system which gives a simplified picture.

---


Third, this method does not take advantages of introducing any distribution assumptions. Some of the indicators approximately follow the Binomial distribution, while some others approximately follow the Poisson distribution and so on. Taking waiting time for operation as an example, it approximately follows a Poisson distribution which is different from this hip re-operation case presented in this paper. Having proper distribution assumptions is important when testing the significance of the deviation of separate units.

A new funnel plot (Figure 6) can be made in the hip re-operation case to illustrate the misleading part of this traffic light method as it is applied in the Swedish Health Care Report:

1. A scatter plot of all the counties is plotted which is the same as the one in Figure 4.
2. Two horizontal lines which are named as “yellow-green” and “yellow-red” cuts dividing the counties into three groups. One cut equals the mean of the 7th and the 8th value of the indicator. The other cut equals to the mean of the 14th and the 15th value of the indicator.
3. A target line which is the same as the one in Figure 4.
4. Two new “warning” curves. The new “warning” curve here have a confidence level of p=0.67. This means that there supposed to be 33% units fall above the higher “warning” curve and 33% below the lower “warning” curve.

Figure 6: Funnel plot of hip re-operation percents for counties in Sweden, 2006-2009
Of course, it is especially interesting to focus on the counties coloured with red which are usually defined as the “worse” ones and need to be improved and the counties above the new upper “warning” line in the funnel plot (Figure 6) which is also labelled as the “worse” ones. Obviously, there is a region between the upper “warning” line and the “traffic light method” division cut which can be a “misleading” region which is light blue in Figure 6. The countries falling into this region are the ones who have a smaller size and are unfairly classified by the “Swedish Health Care Report” into the “red” group as “bad” performance. In this example, ‘Vaster Norrland’ is a country which is wrongly labelled as a “bad” county in the “traffic light” presentation.

4.3 Some possible methodological improvements

According to the standard approaches presented in part 2, a funnel plot could perhaps be a good alternative presentation in the main part of the report. The league plot focuses on each unit, but the funnel plot emphasizes the whole system. Although the results of league plot and the funnel plot are quite similar to each other, it is still important to provide a clear pattern to the readers using a funnel plot.

On the other hand, there can be several other methods to choose the cut-off points in the traffic light presentation system of the result part which can be applied for different aims. In order to compare all these improved methods and the method used in the report, Table 2 is provided.

Table2: four different method of present the result

<table>
<thead>
<tr>
<th></th>
<th>Stockholm</th>
<th>Uppsala</th>
<th>Jönköping</th>
<th>Kalmar</th>
<th>Gothenburg</th>
<th>Halmstad</th>
<th>Västra Götaland</th>
<th>Östergötland</th>
<th>Dalarna</th>
<th>Gotland</th>
<th>Värmland</th>
<th>Jämtland</th>
<th>Norrbotten</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first row in Table 2 is the “Percentile Traffic Light method” presented in the result part of the report. Automatically, it contains 7 red lights, 7 yellow lights and 7 green lights. This will be so irrespective of the shape of distribution of the actual indicator.
The second row in Table 2 provides the result of what we will call the “League Plot Traffic Light method”. The counties can be divided based on location of the 95% confidence intervals and the target line in the league plot (Figure 3.b). Thus,

- **RED**: Counties with confidence intervals to the right of the target line;
- **YELLOW**: Counties with confidence intervals containing the target line;
- **GREEN**: Counties with confidence intervals to the left of the target line;

This method describes each unit separately and compares them with the overall mean. It contains 1 red light, 16 yellow lights and 4 green lights.

The third row in Table 2 shows the result of what we will call the “Standard Funnel Plot Traffic Light method”. In Figure 7.a, we divide the counties into three groups based on comparing the location of each unit and the funnel curve (95%). Thus,

- **RED**: Counties above the higher warning limit;
- **YELLOW**: Counties between the warning limits;
- **GREEN**: Counties below the lower warning limit;

This method is based on the “Special Cause theory”. It focuses on the whole system and gives a clear pattern of the whole system. Then, it compares each unit with the system warning curve. It contains 3 red lights, 14 yellow lights and 4 green lights.

The fourth row in Table 2 presents what we will call “p=0.67 funnel plot TL method”. Compared to Figure 7.a, Figure 7.b applies p=0.67 as its confidence level intend to obtain equally large groups. Based on the same rules as the method above, it divides the units as follows:

- **RED**: Counties above the higher p=0.67 warning limit;
- **YELLOW**: Counties between the p=0.67 warning limits;
- **GREEN**: Counties below the lower p=0.67 warning limit;
This modification approach of the standard funnel plot method aims at dividing the units into approximately equally large groups as they do in the Swedish Health Care Report. It contains 6 red lights, 7 yellow lights and 8 green lights.

The fifth row in Table 2 gives what we will call a “One-side funnel plot TL method”. As it is more interesting to focus on the “bad” performance units, it might be smarter to divide the hospitals or counties in a one-side fashion. This can be more meaningful for the improvement of the health care system. Figure 7.c is a combination of the standard funnel plot and the p=0.67 funnel plot. There is both p=0.95 and p=0.67 limits in this method. Thus,

RED: Counties above the higher 0.95 warning limit;
YELLOW: Counties between the higher 0.95 and higher 0.67 warning limits;
GREEN: Counties below the higher 0.67 warning limit;

This method focuses on the higher rank units. It makes more detailed division at the upper part of the funnel plot. It contains 3 red lights, 3 yellow lights and 15 green lights. So if the readers want to have a more detailed view of the good performance units, they should check the league plot again.

The last row in Table 2 is what we will call the “Fuzzy traffic light method”. A “fuzzy traffic light” method makes some improvements from the standard traffic light method. It is a combination of a pie chart and a common traffic light method. All the ranks that one unit can obtain under the same situation is given by Figure 4.b. Then the chance of this unit being present as red light (rank>66), yellow light (rank>33 and rank<66) and green light (rank<33) can be calculated. Thus, each county or hospital can be presented with a pie chart based on this ratio. So there are some counties given one single colour and also some other counties presented by multiple colours with two or three colours. This fuzzy traffic light method is newly introduced by some statisticians in order to avoid the shortcomings of traditional traffic light method. It provides the reader possible risk suggestion, but it seems to be too difficult for general public to understand.

All the improved methods take the sample size or the patient number into consideration which is really important. Although the researches of the report say the readers should make further discussion themselves, it is still important to give consideration of sample size. If not, it can be unfair for the counties which have a

smaller size when we compare the counties directly in the “Swedish Health Care report”.

5 Conclusions

The Swedish Health Care Report carries on an open comparison of 134 different indicators between hospitals and counties which detailed described the Swedish health care system. Some simple evaluations of the hospitals and the counties are made by the application of the League Plot method (the main part) and the Traffic Light approach (the result part). In spite of the fact that this report is clear and easy to understand, there are still some shortcomings in it which we should point out: the missing of sample size, the misleading of ranking technique and the traffic light method. In this paper, we present some standard methods widely used in the health care system, and also provide some possible improved methods against the traffic light approach used in the Swedish Health Care Report. In brief, a League plot traffic light method and a standard funnel plot traffic light method is recommended when aiming at finding out the outliers; a p=0.67 funnel plot traffic light method is suggested when wishing to divide the units into approximately equally large groups; a one-side traffic light method seems to be a better choice when focusing on the bad performance units.
Reference


Öppna jämförelser av hälso- och sjukvårdens kvalitet och effektivitet – Jämförelser mellan landsting 2010 Appendix Ⅱ .


Appendix I: Result of the Swedish Health Care Report

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Kostnad</th>
<th>Sweden</th>
<th>USA</th>
<th>Japan</th>
<th>England</th>
<th>Russia</th>
<th>Sweden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost per consumer DRG-pension</td>
<td>40,377</td>
<td>41,499</td>
<td>41,866</td>
<td>44,190</td>
<td>41,437</td>
<td>34,350</td>
<td></td>
</tr>
<tr>
<td>Cost per säljkontakt i primärvård</td>
<td>1,263</td>
<td>1,263</td>
<td>1,357</td>
<td>1,357</td>
<td>1,357</td>
<td>1,357</td>
<td></td>
</tr>
</tbody>
</table>

**Omräkningsindikatorer**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Graduation, försörjning och nyfödda</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bäst i område under graviditet, Kx</td>
<td>5,51</td>
</tr>
<tr>
<td>Dödsfall i barn</td>
<td>3,50</td>
</tr>
<tr>
<td>Neonatal dödsfall</td>
<td>1,17</td>
</tr>
<tr>
<td>Likartyp: patient-patient hos nyfödda</td>
<td>1,25</td>
</tr>
<tr>
<td>Brister i vaksamhet</td>
<td>3,79</td>
</tr>
<tr>
<td>Koxigdom vid omgängned graviditet</td>
<td>7,38</td>
</tr>
</tbody>
</table>

**Kvinnafinish**

| Indicator                      | 
|--------------------------------|--------------------------------------|
| Omsorgsdeltagande efter borttagande av ben | 2,25 | 3,26 | 1,34 | 2,85 | 1,09 | 1,09 |

**Röntgenorganens sjukdomar**

| Indicator                      | 
|--------------------------------|--------------------------------------|
| Implantattjänst, total köp | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 |
| Implantattjänst, total köp M. | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 |
| Implantattjänst, total köp M. | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 |
| Implantattjänst, total köp M. | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 |
| Implantattjänst, total köp M. | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 |
| Implantattjänst, total köp M. | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 | 196,0 |

**Patientrapportera resultat, total köp | 0,368 | 0,378 | 0,388 | 0,398 | 0,408 | 0,418 |

---

This table provides a summary of the results from the Swedish Health Care Report, including various indicators such as cost per consumer DRG-pension, cost per säljkontakt in primary care, and indicators related to graduation, pregnancy, neonatal mortality, and other medical procedures.
<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Chinese</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Native American</th>
<th>Native Hawaiian</th>
<th>Pacific Islander</th>
<th>Other</th>
<th>Asian</th>
<th>Combined Total</th>
<th>African American</th>
<th>Canada/Northeast</th>
<th>Canada/Southwest</th>
<th>Canada/Southeast</th>
<th>Canada/West</th>
<th>Combined Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.59</td>
<td>6.99</td>
<td>7.89</td>
<td>4.94</td>
<td>1.81</td>
<td>1.21</td>
<td>0.85</td>
<td>1.35</td>
<td>2.50</td>
<td>3.00</td>
<td>3.36</td>
<td>3.18</td>
<td>1.12</td>
<td>1.12</td>
<td>1.23</td>
<td>1.77</td>
<td>1.35</td>
</tr>
<tr>
<td></td>
<td>2.57</td>
<td>7.03</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>3.59</td>
<td>7.03</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>3.59</td>
<td>7.03</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>3.59</td>
<td>7.03</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>3.59</td>
<td>7.03</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
<td>5.36</td>
</tr>
</tbody>
</table>
Appendix Ⅱ: Part of R code

#Variable: hospital ratio l h case total county
#Target value:
ratio_star <- sum(case)/sum(total)
library(gplots)
##league plot
od1<-order(ratio, decreasing=TRUE)
ratio_od <- rbind(ratio)[od1]
l_od <- rbind(l)[od1]
h_od <- rbind(h)[od1]
hospital_od<-hospital[od1]

pdf('hip_hospital_league.plot.pdf', height = 21, width = 14)
par(las = 1, mar = c(4, 15, 5, 2))
barplot2(ratio_od, width=4, space=2, names.arg=hospital_od,
beside=TRUE, col="light blue", horiz=TRUE,
main="hip reoperation", xlab="ratio",
plot.ci=TRUE, ci.l=l_od, ci.u=h_od, ci.color="black",
ci.lty="solid", ci.lwd=1,plot.grid=TRUE,
grid.lty="dotted", grid.col="grey")
abline(v=ratio_star, col="red")
dev.off()

## funnel plot
ratio_rank <- rank(ratio)
plot(total, ratio, main="hip reoperation", sub="hospital scale",
    xlab="volume", ylab="reoperation rate", col="red")
points(total[ratio_rank<=52&ratio_rank>=27],
    ratio[ratio_rank<=52&ratio_rank>=27],col="yellow")
points(total[ratio_rank<=26], ratio[ratio_rank<=26],col="dark green")
abline(h=ratio_star)

#normal approximation
curve(ratio_star-qnorm(0.995)*sqrt(ratio_star*(1-ratio_star)/x),from=0,to=4000,add=TRUE,type="l",col="dark blue")
curve(ratio_star-qnorm(0.975)*sqrt(ratio_star*(1-ratio_star)/x),from=0,to=4000,add=TRUE,type="l",col="purple")
curve(ratio_star+qnorm(0.995)*sqrt(ratio_star*(1-ratio_star)/x),from=0,to=4000,add=TRUE,type="l",col="dark blue")
curve(ratio_star+qnorm(0.975)*sqrt(ratio_star*(1-ratio_star)/x),from=
0, to=4000, add=TRUE, type="l", col="purple")
# binomial
r1<-qbinom(0.975,1:3500,0.01809)
afa1<-(pbinom(r1,seq(1,3500),0.01809)-0.975)/(pbinom(r1,seq(1,3500),0.01809)-pbinom((r1-1),seq(1,3500),0.01809))
y1<-(r1-afa1)/(1:3500)
r2<-qbinom(0.995,1:3500,0.01809)
afa2<-(pbinom(r2,seq(1,3500),0.01809)-0.995)/(pbinom(r2,seq(1,3500),0.01809)-pbinom((r2-1),seq(1,3500),0.01809))
y2<-(r2-afa2)/(1:3500)
lines(y1[2:3500], col="red")
lines(y2[2:3500], col="blue")
## bootstrap median
median <- rep(NA, 79)
lboot <- rep(NA, 79)
hboot <- rep(NA, 79)
seboot <- rep(NA, 79)
median_rank <- rep(NA, 79)
r <- matrix(NA, 79, 1000)
rsample <- matrix(NA, 79, 1000)
rsam_rank <- matrix(NA, 79, 1000)
for(i in 1:79){
  r[i,] <- rbinom(1000, total[i], ratio[i])/total[i]
  index <- seq(1,1000)
  rsample[i,] <- r[i, sample(index, 1000, replace=TRUE)]
  for(j in 1:1000){
    rsam_rank[,j]<-rank(rsample[,j])
  }
  seboot[i] <- sd(rsam_rank[,])
  median[i] <- median(rsample[,])
  median_rank <- rank(median)
  lboot[i] <- median_rank[i] - qnorm(0.975)*seboot[i]
  hboot[i] <- median_rank[i] + qnorm(0.975)*seboot[i]
  if(lboot[i]<1) lboot[i]=1
  if(hboot[i]>79) hboot[i]=79
}
Z <- cbind(X, median_rank, lboot, hboot)
median_rank_od<-rbind(median_rank)[od1]
lboot_od<-rbind(lboot)[od1]
hboot_od<-rbind(hboot)[od1]
hospital_od<hospital[od1]
pdf('hip_hospital_league.plot_bootstrap.pdf', height = 18, width = 10)
par(las = 1, mar = c(4, 15, 5, 2)) # las=1 horizontal axis label mar marginal size (bottom, left, top, right)
barplot2(median_rank_od,width=4, space=2, names.arg=hospital_od, beside=TRUE, col="light blue", horiz=TRUE,
main="hip reoperation", xlab="median rank",
plot.ci=TRUE, ci.l=lboot_od, ci.u=hboot_od, ci.color="black",
ci.lty="solid", ci.lwd=1, plot.grid=TRUE,
grid.lty="dotted", grid.col="grey")
dev.off()