Incorporating natural variation into IVF clinic league tables

Oscar Lemmers1, Jan A.M.Kremer2 and George F.Borm1,3

1Department of Epidemiology and Biostatistics, 133 Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands and

2Department of Obstetrics and Gynaecology, 791 Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

3To whom correspondence should be addressed at: Department of Epidemiology and Biostatistics, 133 Radboud University Nijmegen Medical Centre, PO Box 9101, 6500 HB Nijmegen, The Netherlands. Tel: +31 243617667; Fax: +31 243613505; E-mail: g.borm@epib.umcn.nl

BACKGROUND: More and more league tables are being published every day to rate the performance of health boards, hospitals and surgeons. However, they do not show the magnitude of uncertainty caused by natural variation. METHODS: We propose a new method to present league tables in which the ratings are easy to interpret. Instead of just giving one score, we suggest the addition of best-case scenario and worst-case scenario scores. The true performance of a clinic, accounting for natural variation, is most likely to be between the best-case scenario and the worst-case scenario for its rating. These ratings can be computed easily, without any special software. RESULTS: We illustrate our method based on data of Dutch IVF clinics from 2004. Six (out of 13) clinics shared a ‘top of the league’ position when considering the best-case scenario. CONCLUSION: There is great uncertainty about the ratings. To show the magnitude of uncertainty, league tables should include the best-case scenario and the worst-case scenario ratings of each clinic.

Key words: IVF/league tables/outcome assessment/quality of care

Introduction

Many countries throughout the world collect health figures. These data can vary from indicators that measure an aspect of the process of care (such as practitioner adherence to clinical guidelines), or outcomes of care at hospitals, to the mortality rates associated with individual surgeons. The idea behind collecting and publishing this information is that the performance of all participants in the public sector should be measured, as the stakeholders (government, patients, insurance companies, hospitals themselves, etc.) have a right to know what the services are achieving. Public trust is not as obvious as it used to be and it has to be earned and re-earned. One way to achieve this is to make the medical sector as transparent as possible, which avoids it being labelled as a ‘cosy culture of professional self-regulation’ (Power, 1997, p.44). Another reason for the continuing demand to broadcast health figures is that publication is thought to act as an incentive for low performers to adopt best practices from the top of the league in pursuit of improvement.

League tables are often published without any mention of the health statistics they are based on, so valuable information is lost. Patients will be reluctant to go to a heart surgeon who stands 10 places lower in the league than his colleague next door, as this suggests a large quality difference. However, they might change their mind if they can see that the difference in mortality rates is only marginal. Therefore, it would be preferable to reveal the real figures. However, even then, the results may be difficult to interpret. This is particularly the case when several outcome indicators are combined into a total score, such as in the US News list of best hospitals in the USA (US News, 2006). Then ratings in league tables have to be used anyway. Owing to the great interest in league tables, we should develop methods to analyse and communicate these figures in the best possible manner.

In this paper, we focus on IVF clinic quality ratings based on the pregnancy per cycle rate. As a relatively large proportion of IVF treatments take place in the private sector, the competition to improve methods, results and league table positions is even stronger than usual. The pregnancy rate is influenced not only by quality differences, but also by treatment policy, case mix and natural variation. These factors obscure the true performance of a clinic, which will also be reflected in the ratings.

Natural variation causes fluctuations in the results. There is a debate whether this unexplained variation is only chance variability or that it mainly consists of yet unknown factors. However, as this is impossible to decide, in the statistical analysis, this variation will be treated as chance variability. An example of natural variation is the number of babies born in a small town today. Without any apparent reason, it might be very different from the number born yesterday. This is because the numbers are small and a few extra births will make a big (relative) difference. Nevertheless, fluctuation in the annual figures will be far less, because the numbers are large. This phenomenon can also be observed in the results of IVF clinics. By chance alone, without any apparent change in treatment or patient mix, the results will vary from time to time. The effect of natural variation will be much greater at hospitals that treat
relatively small numbers of patients. Therefore, a small hospital is more likely to seem to be overperforming or underperforming than a large one, even if this is not true in reality. This natural variation is present in all data and causes difficulty with interpretation: real differences might be obscured, whereas apparent differences might be fictitious.

To determine the size of the influence of natural variation on quality ratings, we suggest that a best-case scenario rating and a worst-case scenario rating should be incorporated into the score of each clinic. The true rating of a clinic, accounting for natural variation, is most likely to be between the best-case scenario and the worst-case scenario for its rating. It can be as good as in the best-case scenario, which is reassuring for patients. On the other hand, it can be as bad as in the worst-case scenario. This is a warning for the clinic: without any action from their side, next years’ rating may drop to the level of the worst-case scenario rating of this year. It is possible to give 95% confidence intervals to the ratings, as is also the case with, for example, percentages, but in our opinion, best-case and worst-case scenarios are more natural and more appealing. They are also much easier to calculate and no special software is required. Our results showed that although there were substantial differences between the IVF clinics, there was great uncertainty about the positions in the league table.

Judging a clinic solely on the outcome, even when a measure for the natural variation is added, does not do justice to the clinics. It is necessary to correct for case mix (such as age of the woman or number of previous cycles of IVF) and treatment policy (such as the number of embryos transferred). The proposed method allows that all these factors can be incorporated. After adjusting, a single league table can be created which takes into account not only these factors, but also the natural variation. The table then gives the true performance of a clinic, which was previously obscured by case mix and natural variation.

We present our method using IVF data from 13 hospitals in the Netherlands.

Materials and methods

Data

Data were obtained from the Dutch Society for Obstetrics and Gynecology (NVOG, 2005).

Table I. Dutch IVF clinic league table for 2004, with observed, best-case scenario and worst-case scenario ratings

<table>
<thead>
<tr>
<th>Clinic</th>
<th>Treatment cycles</th>
<th>Pregnancies</th>
<th>Pregnancy rate (%)</th>
<th>League table position</th>
<th>Best-case—worst-case scenario</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1264</td>
<td>409</td>
<td>32.4</td>
<td>1</td>
<td>1-6</td>
</tr>
<tr>
<td>B</td>
<td>1027</td>
<td>321</td>
<td>31.3</td>
<td>2</td>
<td>1-7</td>
</tr>
<tr>
<td>C</td>
<td>1513</td>
<td>453</td>
<td>29.9</td>
<td>3</td>
<td>1-8</td>
</tr>
<tr>
<td>D</td>
<td>525</td>
<td>154</td>
<td>29.3</td>
<td>4</td>
<td>1-9</td>
</tr>
<tr>
<td>E</td>
<td>654</td>
<td>186</td>
<td>28.4</td>
<td>5</td>
<td>1-9</td>
</tr>
<tr>
<td>F</td>
<td>285</td>
<td>81</td>
<td>28.4</td>
<td>6</td>
<td>1-10</td>
</tr>
<tr>
<td>G</td>
<td>639</td>
<td>177</td>
<td>27.7</td>
<td>7</td>
<td>2-10</td>
</tr>
<tr>
<td>H</td>
<td>399</td>
<td>102</td>
<td>25.6</td>
<td>8</td>
<td>3-10</td>
</tr>
<tr>
<td>I</td>
<td>775</td>
<td>182</td>
<td>23.5</td>
<td>9</td>
<td>6-11</td>
</tr>
<tr>
<td>J</td>
<td>180</td>
<td>41</td>
<td>22.8</td>
<td>10</td>
<td>4-12</td>
</tr>
<tr>
<td>K</td>
<td>817</td>
<td>164</td>
<td>20.1</td>
<td>11</td>
<td>9-12</td>
</tr>
<tr>
<td>L</td>
<td>688</td>
<td>119</td>
<td>17.3</td>
<td>12</td>
<td>10-13</td>
</tr>
<tr>
<td>M</td>
<td>412</td>
<td>61</td>
<td>14.8</td>
<td>13</td>
<td>12-13</td>
</tr>
</tbody>
</table>

The NVOG monitors clinics in the Netherlands that are licensed to carry out IVF. Every year, records are kept at each clinic of the number of treatment cycles started, the number of pregnancies, singleton ongoing pregnancies, twin ongoing pregnancies and triplet ongoing pregnancies. A pregnancy is defined as a positive test in urine or serum (>50 IU l⁻¹), not earlier than 15 days after the ovum pickup. We used the results of treatment cycles that started in 2004.

Statistical methods

The idea is that every hospital operates within a certain bandwidth or margin: there is a best-case scenario and a worst-case scenario at either end of the rating scale. Table I summarizes the results recorded at Dutch IVF clinics in 2004. The clinics were rated according to their observed pregnancy rates. The clinic with the highest rate took top position, the clinic with the second highest rate took second position and so on.

Owing to natural variation, there is always a margin of uncertainty around this observed rate. For example, when we compared clinic D with clinic G, the former was found to have a pregnancy rate of 29.3% (Table I), whereas the latter had a pregnancy rate of 27.7%. The 95% confidence interval of the difference in rates between clinic D and clinic G was −3.6 to 6.8%. Possible interpretation of the right-hand margin of 6.8% is that this figure is the best that clinic D can reasonably achieve when compared with clinic G. In this case, clinic D has been given the benefit of the doubt: the margin of uncertainty has been explained completely to their advantage. Then the performance of clinic D is better than clinic G, thus clinic D performs better than clinic G in a best-case scenario for the rating of clinic D. Similarly, possible interpretation of the left-hand margin of −3.6% is that this figure is the worst that clinic D can reasonably achieve when compared with clinic G. Clinic D has then been given all the doubt, because the margin of uncertainty has been explained completely to their disadvantage. Now the performance of clinic D is poorer than clinic G, thus clinic D performs worse than clinic G in a worst-case scenario for the rating of clinic D.

The best-case scenario rating of IVF clinic D can be obtained as follows.

(i) Calculate the 95% confidence intervals of the difference between the pregnancy rates at clinic D and each of the other clinics.

(ii) Count the number of intervals with a positive right-hand margin: 12.

(iii) Thus, when clinic D is given the benefit of the doubt, its performance is better than that of 12 of the other clinics. As there
are in total 12 other clinics, the best-case scenario leads to clinic D taking position 1.

The worst-case scenario rating of IVF clinic D can be obtained as follows

(i) Calculate the 95% confidence intervals of the difference between the pregnancy rates at clinic D and each of the other clinics.
(ii) Count the number of intervals with a negative left hand margin: 8.
(iii) Thus, when clinic D is given all the doubt, its performance is worse than that of 8 of the other clinics. In this worst-case scenario, it takes position 9.

To calculate the best-case scenario and the worst-case scenario ratings of each clinic, simply repeat the procedures described above.

Results

On the basis of data from the NVOG recorded in 2004, the Dutch IVF clinics were rated according to their pregnancy per cycle rate. Then the best-case scenario and worst-case scenario ratings were calculated for each IVF clinic. The results are shown in Table I.

Figure 1 shows the league table position, the best-case scenario and the worst-case scenario ratings of each of the Dutch IVF clinics according to the pregnancy per cycle rates.

In almost every case, the best-case scenario and the worst-case scenario ratings lay far apart.

Discussion

The best-case scenario rating of a clinic is the position in the league table that it would take if it had everything going for it. In other words, the highest position that a clinic can reasonably achieve. The worst-case scenario rating shows the position that a clinic would take if it had everything going against it, i.e. the lowest position that it can reasonably achieve. Although 100% certainty cannot be given, the true ranking of a clinic is most likely to be between the best and the worst-case scenario ratings.

Figure 1. Graphic representation of the 13 Dutch IVF clinics rated according to their results in 2004. An equality sign indicates the league position of the clinic. Best-case scenario rating is indicated by an upwards triangle, whereas the worst-case scenario rating by a downwards triangle.

Natural variation

Incorporation of a best-case scenario rating and a worst-case scenario rating into the quality rating showed the size of the influence of natural variation on the position in the league table. This gave much more information than one rating alone. Our results showed that in most cases, there was a large difference in the position of a clinic between the best-case and the worst-case scenarios. A clinic performs much better in a ‘good’ year, which results in a high position, but much worse in a ‘poor’ year, which results in a low position. This means that due to natural variation, there is great uncertainty about the true position of a clinic in the league table.

The influence of natural variation will be larger or smaller if the number of treatment cycles is smaller or larger, respectively. This is illustrated in Table I, where clinics I and J have similar pregnancy rates (23.5 and 22.8%, respectively). The best-case scenario rating for clinic I (775 treatment cycles) is 6, whereas its observed rank is 9. For the much smaller clinic J (180 treatment cycles), with observed rank 10, it is 4.

Several authors have described the shortcomings of league tables, including Marshall and Spiegelhalter (1998). They discussed league tables of IVF treatment clinics in the UK and used simulation methods to place confidence intervals around the individual ratings to indicate the level of uncertainty about the position. One of their findings was that in 1996, only 1 out of 52 UK IVF clinics could confidently be placed in the bottom quarter of the league table. Green and Wintfeld (1995) evaluated the ratings of heart surgeons based on mortality rates. They noted that even after risk adjustment, 46% of the surgeons moved back and forth between the top half and bottom half of the league table over the course of 1 year. Once again, natural variation had a great deal of influence and made it difficult to draw firm conclusions based on league table positions alone.

Case mix and risk adjustment

Another well-known cause for the differences in performance (besides natural variation) is case mix, because many person-specific factors influence IVF pregnancy rates. Examples are the age of the woman, number of previous cycles of IVF, basal FSH concentrations, total ovarian response and the number of embryos transferred. Usually, adjusting for certain factors will change the ratings, which might cause substantial position switches between individual clinics; see, for example, Parry et al. (1998). They placed hospitals into a league table according to their annual neonatal mortality rate. The hospital with the lowest crude average mortality rate was relegated to sixth position (out of nine) after adjustment for clinical risk and the severity of illness. Therefore, to be informative for patients and fair to clinics, all data should be adjusted for case mix. In our example of Dutch IVF clinics, the patients in the clinic on the last place were relatively old (on average 36 years). Adjustment for age would improve the rating of this clinic. In this paper, we have not adjusted for case mix, because we wanted to focus on the ranking method only. Also, the data from the clinics were too incomplete to allow meaningful adjustment. This again...
shows the importance of collecting and providing all relevant information.

To account for case mix differences between IVF clinics, logistic regression can be used to calculate the adjusted results (Parry et al., 1998). Then our new method can be applied to the adjusted estimates. Another solution might be to divide the patients into groups (e.g. according to age, number of embryos transferred) and to report results per group. But the numbers in some groups would be too small to draw any further conclusions.

Confidence intervals
An alternative to publishing league tables is to present pregnancy rates with confidence intervals. These intervals indicate the margin of uncertainty about the true performance. Nevertheless, sometimes we are more interested in positions in a league table, to see who is ‘top’, or who is ‘bottom’. The disadvantage is that we no longer see the margin of uncertainty, or the magnitude of the underlying difference in performance. By incorporating the best-case and the worst-case scenarios into the quality rating of a clinic, the individual league table positions are put into better perspective. In this way, the large body of data has been reduced into three easily interpretable numbers that facilitate transparency. We consider that this is a worthwhile improvement for the patients and the clinics.

Summary
Our new method makes it possible to distinguish clinics that show significantly poorer (or better) performance and they can subsequently be monitored. The method also helps to avoid the opposite, namely the premature naming and shaming of a clinic. Another advantage is that it can be used to analyse a league table that is based on a total score, i.e. a combination of several outcome indicators. A difference in these scores is difficult to interpret, therefore it is more appropriate to look at the differences between the positions in the league table. Our method quantifies the influence of natural variation on the position in the league table in an easily interpretable way.

Appendix
Suppose we have two hospitals A and B, with number of pregnancies \( n_A \) and \( n_B \) and number of IVF treatment cycles \( N_A \) and \( N_B \), respectively. The margin of uncertainty in the observed difference between their pregnancy rates is

\[
1.96 \sqrt{\frac{n_A}{N_A} - \frac{n_B}{N_B}} 
\]

Note that the ratios \( n_A/N_A \) and \( n_B/N_B \) are the exact pregnancy rates at hospitals A and B, respectively.

In the present paper, we considered that the performance of hospital A was better (or poorer) than hospital B in the best-case scenario (or worst-case scenario) when the right-hand margin (or left-hand margin) of the 95% confidence interval was positive (or negative). In the formula above, we used the factor 1.96, because a two-sided 95% confidence interval corresponded with \( z_{0.025} = 1.96 \). Obviously, a different percentage can be selected, for example, 90%. Then the factor 1.96 has to be replaced by \( z_{0.05} = 1.65 \).

References

Submitted on August 3, 2006; resubmitted on December 19, 2006; accepted on January 11, 2007