Author’s response to reviews

Title: Incorporating natural variation into IVF clinic league tables: The Expected Rank

Authors:

Hester F Lingsma (h.lingsma@erasmusmc.nl)
Marinus JC Eijkemans (m.eijkemans@erasmusmc.nl)
Ewout W Steyerberg (e.steyerberg@erasmusmc.nl)

Version: 2 Date: 6 April 2009

Author’s response to reviews: see over
Dear Dr Todd,

Thank you very much for considering our manuscript entitled: ‘Incorporating natural variation into IVF clinic league tables: The Expected Rank’ for publication in BMC Research Methodology.

We would like to thank the reviewers for the care and attention they put into the review of our paper. Three of the reviewers found the paper important to either those with related research interests or in its field. One reviewer did not find it of interest but also declared that he was less statically skilled (“I do not feel adequately qualified to assess the statistics”).

Most of the comments of the reviewers were of statistical nature and about the important question whether any ranking should be attempted at all. We have edited the method section to make the statistics and the formulas more clear and we have addressed the problems with ranking more extensively in the discussion. Another comment was that this example does not show all the features of our method, so we have also added an example to the discussion about what would happen in a dataset with more natural variation.

We are pleased to send you now the revised version of the paper, in which the changes are highlighted. Below we have addressed the reviewers’ comments separately in detail.

We hope you will find our revision appropriate and that will find it suitable for publication in BMC research methodology. We look forward to hearing further from you.

Sincerely yours,

Hester Lingsma Msc

Reviewer 1

The authors investigate a novel method to analyze and present rankings. I think the approach makes sense and is worth investigating. However, I feel that the conclusion of the paper should be exactly the opposite of what the authors propose: despite the fact that the method looks promising at first sight, its results on the example in the paper are disappointing and a closer (theoretical) look does not suggest that the method is not worth pursuing any further. Which is still an interesting finding.

I will try to explain this below.

We choose this example for our paper because it was previously used to demonstrate another alternative ranking method. In retrospect it might have not been the best example since the natural variation is quite small. We disagree that the method is not worth pursuing further, the results might be disappointing in this example but that is due to the data, not to the method. We have now mentioned also in the background section why we chose this dataset. Page 3, last paragraph.

Major compulsory revisions

The main problem with rankings is, that they are over interpreted. Readers think that the second ranked centre is definitely worse than the first one, while in fact the difference may small and may be the result of random fluctuations. Rankings are not as precise and accurate as they seem. However the method the authors does not solve this problem, as they propose expected rankings with 2 decimal places. This suggests an even higher (but false) precision and accuracy. But more importantly, there is no statistical basis for it. Taking into consideration the assumptions of the statistical model (which may not be –fully- correct) and the standard errors of the estimates, those two
last digits are purely random. So we should round off the expected rankings, but then we see that there is virtually no difference between the raw and the expected rankings. This is an interesting point with which we agree. We have now rounded the ERs to one decimal instead of two to prevent over interpretation. Some clinics obtain almost the same ER now which indeed better demonstrates that they are rather similar. We also rounded the PCERs. Table 1.

My conclusion would be: keep it simple and stick to the original rankings. Or, if you want to be more subtle, as the authors suggest in their discussion, publish the raw (pregnancy) data. An even better approach may be to publish the (shrunked) random effects estimates (BLUPs) for those data (and rank those). On this point we do not agree completely. As mentioned above, we feel that the ERs give more subtle information than the original rankings. The raw data might be indeed a good idea, but disregard that in a smaller center a high or low pregnancy rate might be just coincidence. Publishing the shrinken random effect estimates (BLUP) is also a good idea, but is essentially the same as we do with the ER (which shrinks the rankings). We don’t think ranking the BLUPs is a good idea since this disregard the uncertainty again, i.e BLUPs of 0.18 (clinic D) and 0.15 (clinic E) would be still ranked 4 and 5 and not 5.0 and 5.1. In this dataset there was actually no difference between the ranking of the fixed and random estimates since the natural variation was relatively small. This is more a problem of the data than of the method, and is now also addressed in the discussion. Page 10, third paragraph.

Detailed remarks

1. The authors write that the random effect model provides estimates (I assume they mean BLUPs, but it depends of course on how they interpret the random effects model) for the ‘true’ performance of the centers, and implicitly suggest that the fixed effect analysis does not. In fact both approaches estimate the ‘true’ performance, but the underlying model, i.e. the assumptions differ. I agree with the authors that the BLUPS may be more appropriate, but the wording is somewhat unfortunate.
We used ‘true’ to indicate the difference between ‘real’ variation and variation by chance. We think it is an appropriate term here. In the method section we now explain better what we mean by ‘true’. Page 4, last lines.

2. The authors mention that they have used R for their analyses. They should perhaps also mention that it is not always an easy task to calculate the BLUPs. In SPSS, for example, it is impossible, I think? I am not sure that SPSS can even do random effects logistic regression? In any case, they might explain (or warn) that BLUPS are not the same as least squares estimates and that it is not always straightforward to derive them and to select the correct information from the output of the various statistical programs.
We agree, as far as we no it is not possible to fit random effect logistic regression models with SPSS. It is however in some more advanced programs such as SAS, Stata and Splus. This is now also mentioned in the discussion. Page 8, last lines.

3. I would suggest to use N instead of F to denote the normal distribution, as F suggests the F distribution. It may also be worth mentioning that F (or N) refers to the standard normal distribution and not just any normal distribution.
We changed $F$ into $\Phi$ since we actually refer to the cumulative normal distribution. This was also suggested by one of the other reviewers. Page 5, first formula.

4. I am not sure that $p(#i<#j)$ is the probability that $i$ is worse than $j$. I would say that it is the probability that the result $#i$ is less than the result $#j$ when 'in fact' they are equal. This is definitely the case in the fixed effects model. It may be worth mentioning here that in the random effects approach, the variance of $#i$ is the variance of the BLUP? I am not sure that BLUPs are independent estimators, because they 'share' the estimate of $#$. Therefore: is it correct to simply add the variances?

This is a philosophical point which is intuitively correct. Theoretically it is indeed questionable whether the BLUPs are independent. Practically, in this example the shrinkage was small so we consider the BLUPs independent.

5. The authors introduce the rankability. They find a rankability of 0.9 and conclude this is high. But would 0.7 also have been high? Or 0.5? How to interpret this rankability? Why is it based on the median, and why not on some other percentile or on the mean? In fact there already exist quite a number of methods to estimate heterogeneity (because that is what we are talking about): ICC, $I^2$, $Q$, tau... Why introduce yet another one?

The rankability can be interpreted as the part of heterogeneity between the clinics that is due to true differences (as opposed to 'natural variation'). We observe a certain amount of variation, of which in this example 90% is due to 'true' differences. We considered this high, but we agree it is a value judgement, similar to the discussion about what can be considered as an acceptable $I^2$. Furthermore this method is a proposal; it also could have been different. Other methods are indeed similar in both statistical computation and interpretation, but none of them is exactly the same. We have addressed the issue above in the discussion. Page 10, second paragraph.

6. On the one hand, the authors claim that their method leads to a more simple presentation of the rankings than the methods proposed by other authors, because they list only one ranking ('number') per clinic. (As I mentioned, this single ranking may still lead to over interpretation.)

On the other hand, they propose a whole matrix of over a hundred probabilities $p(#i<#j)$. I am not convinced this is more simple. It also suggests a much greater precision than the data allow: a table of more than a hundred—estimated two-digit numbers has a huge total error rate. I do not think it offers any better guidance for couples to choose that raw pregnancy rates (or the BLUPs).

The calculations might be complex but the presentation is easy to interpret. There is a 36% probability that an average couple at clinic A has a lower chance of pregnancy than if they attend clinic B. The uncertainty is reflected in the probabilities. Also the table is less feasible if there are large numbers of clinics to compare. This issue is now also mentioned in the discussion. Page 9, first paragraph.

Reviewer 2

The authors have tackled an interesting and controversial issue. In my view “League table” is not a ‘politically’ correct concept to use for comparing institutions performance. The concept of ranking is too subjective and lead to more confusion. I would rather focus on performance level rather than rank. A
Discretionary:
1 - page 4 typos: First line “Data were ……originally from …”
The typo (form) is corrected. Page 4, first line Methods.

2 - Page 9 2nd line “We also see … perform in a very similar fashion, such as clinic D and E”.
This sentence is changed.

3 - page 5 “Where F (Phi is more appropriate for normal CDF) is the normal cumulative distribution…”
We have changed F into Phi.

4 - For Greek letters it’s better to use the equation editor in MSWord (object) rather than symbol as font (see the equations in page 5).
This is done now

Minor:
1 - Figure is interesting to display as long as the number of clinics is low. It would be very difficult to represent the graph if there are, say 100 clinics to compare.
The same applies for Table 2.
We agree, if there are large numbers of clinics to compare these graph and table less feasible and other types of graphs (e.g. a funnel plot) may be better. With regard to the table, if decision-making for patients is really the aim one could think of making separate tables for different regions for example.

2 - For Table 2, the upper part of the matrix is showing 1% isn’t 100% instead?
This should indeed be 100%, it is changed now

3 - I hope that, in the near future, the authors would carry on this work using case-mix adjustment for more credibility.
We agree that case-mix adjustment is absolutely crucial in comparing providers. Unfortunately in this dataset we don’t have the information to adjust.

Reviewer 3

This is an interesting paper on an important topic. The methodology is not new but is appropriate, although I think a more interesting example could have been found. Also at various points the methods are not sufficiently clearly described.

Minor Essential Revisions
1. Page 4. The term ‘natural variation’ is used for unavoidable random variability. This is not a bad term, but the first time it is introduced it should be explained that this may also be described in other ways.
In the introduction we have mentioned now that with natural variation we mean indeed variation by chance/random variation. Page 3, second paragraph.

2. Page 5. The formulae are not clear and make no distinction between parameters and observations. Are the theta_i’s the estimates? Then they should either have ‘hats’ on them or be represented by Roman letters, say t_i. In the
formula $p(\theta_i < \theta_k)$, are you using a Bayesian argument and referring to the true underlying rates? Or is the predictive probability of the two estimates being in a particular order (in which case this does not represent the rank probability). And are the variances based on the estimated rates after back-transforming from logistic (in which case normality will not hold), or in fact are they logistic transforms in which case it is untrue to call these the pregnancy rates. This section does need tightening up, made more precise, and maybe therefore put in an appendix.

We have tightened up the section Ranking and rankability -especially the formulas- taking into account the suggestions: the estimates have a 'hat' now and we made clearer that we indeed refer to the Empirical Bayes estimates. The variances are based on the logistic transforms and we have now called them logistic regression coefficients instead of pregnancy rates.

3. Page 5. why does $er_i$ have a 1 in its formula?
The ER has a 1 in the formula to avoid ranks below 1. If a center would be the 'best' with a high level of certainty, it would have only very low $p(\theta_i < \theta_k)$ and ER could be below one, which we find illogical for a 'rank'. This is added to the explanation of the formula. Page 5, last paragraph.

4. Page 6. your rho is just an intraclass correlation.
This would be true for a linear regression model. For logistic regression they are probably not exactly the same, also not because rho uses the median sigma2. This is now also mentioned in the method section. Page 6, last paragraph.

5. page 7. The problem with this example is that the clinics are fairly heterogenous, there is almost no shrinkage and the ER's are very close to the observed ranks. It would be interesting if they were all pulled towards the middle, which you could point out in the discussion that this would occur with rho smaller than 0.9.
We agree that in retrospect this is not the best example to make the features of the method clear. In the discussion we have added a short description of an example in stroke were rankability was only 0.55 and the ER indeed already shrunken to the median rank. Page 10, last 2 paragraphs.

6. page 10. line 3, NO - the heterogeneity is quite HIGH. rho =0.9, 90% of variance due to between-clinic variability, tau about 0.3, quite high on a logistic scale.
This is an interesting point. Tau is indeed 0.28, corresponding to 95% of the odds ratios for clinics between 0.57 to 1.74. Maybe we can indeed consider this substantial, but compared to previous work (e.g. in stroke, tau2=0.38) it is low. We have changed this in the discussion. Page 10, second paragraph, last line

7. page 17 Table 2: what does the phrase ‘A is expected to be worse than B with 36% probability’ mean? I think you mean something like: you assess a 36% probability that an average client at clinic A has a lower chance of pregnancy than if they attend clinic B.
We mean that there is a 36% probability that an average couple is more likely to get pregnant in clinic A than in clinic B. We have used your suggestion now in the table legend.

Reviewer 4
Major Comments

The authors start from the position that comparative performance of healthcare providers is often presented using league tables. They note that such league tables do not incorporate the natural variation in the rankings and propose a method for “improving” the ranking process to allow for chance variation.

The working premise in this paper is that ranking can be “improved”, when the arguments against ranking (empirical evidence that ranks are not reliable and are meaningless under chance causes and even more meaningless under assignable causes of variation) are overwhelming. Ranking is flawed! (Lancet 2001; 357: 463–67; Statist. Med. 2005; 24:1185–1202)

Fortunately, there are appropriate ways of presenting comparative data which avoid the pitfalls associated with ranking (Human Fertility 2006; 9(3): 145 – 151).

The reader would benefit from the being made aware of the above, and also about ways of presenting data which are statistically sound and avoid ranking altogether.

We agree that rankings can be flawed, and are not the ideal way to present performance data. Especially when natural variation is not taken into account. Nevertheless rankings are popular in the lay press, with the large risk of over interpretation. Therefore we argue that if one wants to make a ranking, the best way is to use the ER –which avoids the ranking pitfall of chance– instead of standard rankings. This is mentioned now more clearly in the background and in the discussion. Page 3, second paragraph, and page 10, last paragraph.

The rationale for the expected ranking procedure suggested by the authors is not well established in the introduction – in other words, why do we need another method of adjusting ranks for natural variation when at least two other methods already exists? Some justification is needed.

The main problem with ranking is that it leads to over interpretation. We think the ER and PCER prevent over interpretation better than previously proposed methods. This is now also mentioned in the aim in the introduction. Page 4, upper lines

The authors make use of random effects modelling to arrive at expected ranks and also make a probability statement of one clinics performance against another. Whilst a statistical review is necessary, the fundamental limitation appears to be that this approach does not distinguish between clinics that are varying purely by chance – i.e. even if we set up a test data set with variation in outcomes between clinics being purely random, then the proposed method will still give a rank and a probability – but these statistics would not be meaningful for a random variable. Clinic ranked 1 and clinic ranked 10 are all the same because they arise from the same random process and it is not clear how the proposed method overcome this issue?

In contrast to a fixed effect model, that is often used in estimating center differences, the random effect model we use in our study does does not distinguish between variation that is caused just by chance and variation that is caused by a true underlying difference in outcome. Theoretically, if there would be only random variation, all the clinics would get expected rank 7 (the middle rank), all the probabilities would be around 50%, and the rankability would be very close to 0. The conclusion would than be that all the clinics are essentially the same and that it makes no sense to compare or rank them. So actually the method would overcome the issue mentioned. However in this dataset the random variation is
small because of the large numbers, while there is quite some true underlying variation. That is why the fixed and random ranks do not differ. This is also visible in the high rankability. This is now mentioned in the discussion. Page 10 third and fourth paragraph.

The outputs of the proposed methods are not materially different from the ones arrived at by simple ranking. This could be due to the selected data set or may be a feature of the method. It seems to me that a more a rigorous and comprehensive evaluation (eg after adjusting for case-mix factors) is required to determine the “operating” characteristics of this method versus other approaches before this method can be routinely used. As also mentioned in the reply to the previous point, the fact that the output does not change substantially is more a feature of the data than of the method. We do feel however that giving the ERs of for example D and E of 5.0 and 5.1 is makes more clear that there is actually no difference between the two than giving them rank 4 and 5. So our method gives much more subtle results than the simple rankings, but we have chosen the best dataset to show the features of our method, since the natural variation is limited. This is now mentioned in the discussion. Page 10 third and fourth paragraph.