Eur Spine J (2009) 18 (Suppl 3):S305–S311
DOI 10.1007/s00586-009-0943-7

REVIEW ARTICLE

Benchmarking with Spine Tango: potentials and pitfalls

Christoph Röder · L. Staub · D. Dietrich · T. Zweig · M. Melloh · M. Aebi

Received: 22 December 2008 / Revised: 11 March 2009 / Accepted: 12 March 2009 / Published online: 1 April 2009

Abstract The newly released online statistics function of Spine Tango allows comparison of own data against the aggregated results of the data pool that all other participants generate. This comparison can be considered a very simple way of benchmarking, which means that the quality of what one organization does is compared with other similar organizations. The goal is to make changes towards better practice if benchmarking shows inferior results compared with the pool. There are, however, pitfalls in this simplified way of comparing data that can result in confounding. This means that important influential factors can make results appear better or worse than they are in reality and these factors can only be identified and neutralized in a multiple regression analysis performed by a statistical expert. Comparing input variables, confounding is less of a problem than comparing outcome variables. Therefore, the potentials and limitations of automated online comparisons need to be considered when interpreting the results of the benchmarking procedure.

Keywords Spine Tango · Online statistics · Benchmarking

C. Röder (✉) · L. Staub · T. Zweig · M. Aebi
MEM Research Center for Orthopaedic Surgery,
Institute for Evaluative Research in Orthopaedic Surgery,
University of Bern, Stauffacherstrasse 78,
3014 Bern, Switzerland
e-mail: christoph.roeder@MEMcenter.unibe.ch

C. Röder
Spine Service, Inselspital Bern, University of Bern,
Bern, Switzerland
e-mail: christoph.roeder@insel.ch

L. Staub
NHMRC Clinical Trials Centre, University of Sydney,
Sydney, Australia

D. Dietrich
Institute for Mathematical Statistics and Actuarial Science,
University of Bern, Bern, Switzerland

M. Melloh
Dunedin School of Medicine, University of Otago,
Otago, New Zealand

Introduction

Benchmarking is the process of comparing the cost, time or quality of what one organization does against what other similar organizations do. The result is often a business case for making changes in order to make improvements. Also referred to as "best practice benchmarking" or "process benchmarking", it is a process used in management in which organizations evaluate various aspects of their processes in relation to best practice, usually within their own sector. This then allows them to develop plans on how to make improvements or adopt best practice, usually with the aim of increasing some aspect of performance. Benchmarking may be a one-off event, but is often treated as a continuous process in which organizations continually seek to challenge their practices.

Translated to the medical field, a surgeon or a department would compare the quality of their own patients'outcomes with that of a peer group of surgeons in order to find out if their results are superior, equal or inferior to that benchmark. In the latter case, the desirable consequence would be an analysis and identification of problem areas and the implementation of new and improved practices.

Benchmarking is a powerful management tool because it overcomes "paradigm blindness." Paradigm blindness can be summed up as the mode of thinking, "The way we do it is the best because this is the way we've always done it." Benchmarking opens organizations to new methods, ideas and tools to improve their effectiveness. It helps overcome resistance to change by presenting successful methods of problem solving that are different to the ones currently employed.

Enabling benchmarking possibilities is one of the fundamental goals of the Spine Tango venture. Only such international projects can offer all participants the same language and set of variables in order to share their information in one and the same database. This data pool has the potential to represent the benchmark for state of the art spine surgery in Europe and in the future maybe even in other parts of the world.

As of December 2007, the online statistics tool was upgraded with a first version of benchmarking functionality. Although this represents a huge step forward in increasing the scientific value of the Tango, it also entails risks, i.e. a misinterpretation and misuse of the generated statistics by the methodologically less educated user. In the current article, we demonstrate the potentials and pitfalls of online benchmarking and explain when statistical modelling becomes indispensable.

Input variables versus outcome variables

Input variables are those variables that have an influence on the outcome. These can be patient characteristics like age, sex, main diagnosis, extension of lesion, spinal comorbidities or ASA status. Such variables are often referred to as "case mix" of a hospital. Other input variables include surgeon qualification, type of surgery (conventional, MISS, LISS, etc.), access or surgical measures. In contrast, outcome variables typically deal with the result of surgery. They can be found on the discharge subform of the surgery questionnaire (hospitalization times, complications) but foremost on the followup form, e.g. surgical goals achieved/partially achieved/not achieved, overall outcome rating by examiner, complications and also on the patient based followup forms (COMI neck and back questionnaires). Intraoperative complications could be considered a direct outcome variable of the circumstances of the case and the surgery, but they could also be considered an input variable for the final treatment result. Variables like "Goal of surgery" may be considered an input variable, but they are probably rather an independent type of information that can later indirectly be used to assess the outcome. Hence, not all variables can be clearly allocated to one of the two groups.

Potentials: online benchmarking of input variables

Because the online statistics function is not yet able to automatically link primary forms with their associated followups, most users perform online statistical queries and benchmarking based solely on the surgery questionnaires. As previously discussed, the form is mostly made up of input variables and consequently most statistical comparisons can directly be performed, e.g. age of patients, sex distribution, types of diagnoses, etc. However, these statistics only show descriptive analyses in the form of tables and figures, but do not include statistical tests of significance. As such tests have to meet certain assumptions of the distributions of underlying data, they should not be automatically generated. As we will show in the following, more profound analyses require a good knowledge of the dataset and expertise in statistics.

Pitfalls: online benchmarking of output variables

Looking at complications rates per se without reference to the actual surgical outcome makes them an output variable. Hospitalisation times are a similar case. In order to highlight and explain why this type of variable cannot be compared in a similarly direct way as the input variables, we describe the comparison of "raw" proportions of dura lesions in posterior spinal fusion of seven selected Spine tango hospitals and how corrections of these proportions are performed by multiple regression analysis and modeling in order to allow adjusted comparisons [4]. Adjustment is made for all those input variables that have a statistically significant influence on the dura lesion. These influential covariates are usually what we refer to as input variables in this article.

Step 1: display of raw proportions

Figure 1 shows the raw proportions of dura lesions in these hospitals which are 8.5, 2.3, 2.8, 0.6, 3.1, 4.0 and 2.9% (image format as provided by online statistics tool).

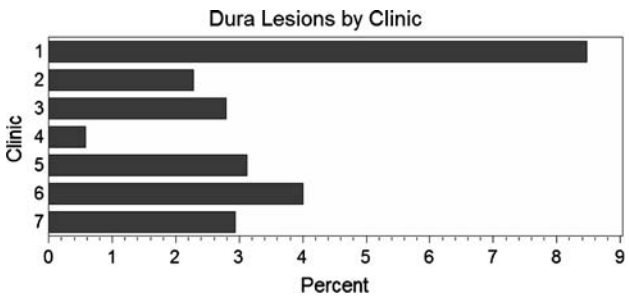

Fig. 1 Raw and unsorted proportions of dura lesions in seven selected Spine Tango hospitals

Step 2: sorting of raw proportions

When sorting these proportions (Fig. 2) it becomes obvious, that there probably is a true difference between hospital 1 and 4. However, what can be said about the other clinics? Surgeons could argue that their unfavorable case mix is the reason for high numbers of dura lesions, or that their surgical skills are responsible for low numbers. But how can we find out whether these covariates (input variables) really matter?

Step 3: calculating standard errors and average proportions

In a next, still descriptive step we have to calculate the unweighted average dura lesion rate of these seven hospitals which one could already consider a "raw" benchmark, even if hospital size is not yet accounted for. This average was 3.5%. More importantly, we have to provide error bars (i.e. standard errors for binomial proportions) which indicate how precise the estimated dura lesion rates are [1, 2]. This is essential because in real life none of the participants documents 100% of his operated cases and/or complications, be it intentionally or for the normal organisational problems like lost forms, forgotten forms or incomplete forms. Hence, the cases stored in the Spine Tango data base are only a sample of all the cases occurring in the participating hospitals. This depiction shows us a different picture already, namely that probably only hospital 4 is below and hospital 1 above the average dura lesion rate and that all other participants are comfortably within the benchmark. Note that large patient numbers in a clinic generally lead to smaller error bars due to a more precise estimation of the point estimates (Fig. 3).

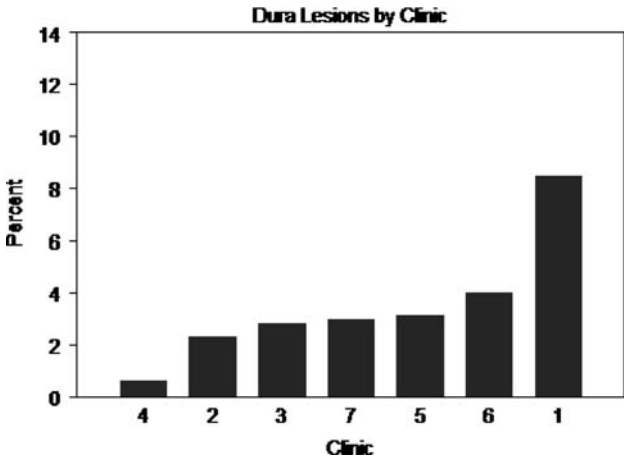

Fig. 2 Raw but sorted proportions of dura lesions in seven selected Spine Tango hospitals

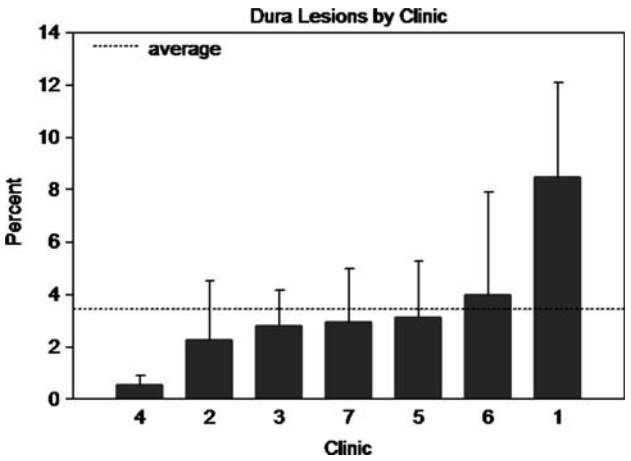

Fig. 3 Raw and sorted proportions of dura lesions with standard errors in seven selected Spine Tango hospitals. In addition, the average raw proportion of dura lesions as raw benchmark is displayed for the seven hospitals

Step 4: building a statistical model: calculation of probabilities by univariate logistic regression (instead of empirical proportions)

Now it is time for a statistician to move in and start with statistical modeling, a procedure that goes beyond what our standard online statistical routines can provide.

In a statistical model, the proportions become probabilities and these can be depicted as point estimates, for example as odds ratios [1, 2], with positive and negative error bars around them. By displaying the unadjusted dura lesion probabilities and the average dura lesion probability we can see that the initial "guess", namely that all but hospitals 1 and 4 are within the benchmark is still confirmed (Fig. 4). Note that errorbars are asymmetrical

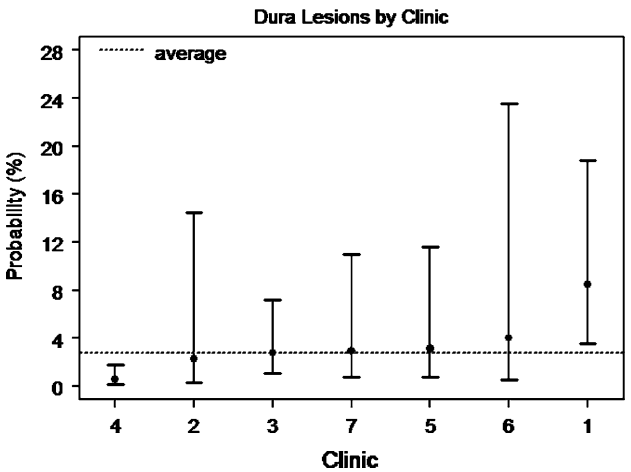

Fig. 4 Unadjusted and sorted probabilities of dura lesions with standard errors in seven selected Spine Tango hospitals. In addition, the average probability of dura lesions as raw benchmark is displayed for the seven hospitals

because they have been calculated on the logit scale and then transformed back to the probability scale. We used hospital 4 as reference hospital with an odds ratio = 1.0. The other odds ratios were 4.1, 5.0, 5.3, 5.6, 7.3 and 16.1 which means that the odds for a dura lesion in the other hospitals are between 4 and 16 times higher. The large confidence intervals show us, however, that these risk estimates can vary to a great extent.

Step 5: adding influential "input variables" (covariates) in a multiple logistic regression model

In order to assess if case mix, surgical skills or other covariates truly influence the dura lesion rate, we have to conduct a multiple regression analysis that includes all the parameters which we think could possibly affect this rate [1–3]. A multiple logistic regression corrects the probability estimates of dura lesions for imbalances in input variables between hospitals. A selection of the potential factors can be based on medical reasoning, but in case of a limited set of available information, as is the case in a basic registry data set like the Tango, we could also include all covariates. Sometimes, one can construct additional covariates by combining certain parameters into a new one. For example, we created a new covariate "type of fusion" by combining information about sole fusion, fusion with instrumentation and fusion with instrumentation and cage implantation.

Consequently, we included the following covariates from the Tango surgery form into the multiple regression model:

- age
- sex
- main pathology
- number of previous spine surgeries
- level of procedure
- number of fused segments
- operation time
- center of intervention
- surgeon credentials
- type of fusion

Step 6: interpretation of results of the regression analysis

Running a stepwise elimination procedure with a significance level of alpha = 0.05, non-significant covariates are sorted out of the model in a stepwise process. The following two covariates remain in the model as significantly influencing the dura lesion probability:

- center of intervention ($P = 0.020$)
- number of fused segments ($P = 0.018$)

Expressed in simple terms, with an error probability of 2% ($P = 0.020$) we can state that the center of intervention has a true influence on the dura lesion rates and with an error probability of 1.8% ($P = 0.018$) we can state that the number of fused segments has an influence in all posterior spinal fusion surgeries conducted in these hospitals. Remember, we are only looking at a sample of hospitals and at a sample of procedures and try to draw conclusions for the real world, i.e. for all hospitals and all procedures. Because we look at a representative sample of procedures from these hospitals but at a non-representative sample of hospitals as these seven participants do not necessarily represent the world of spinal surgery, we should limit our conclusion to these seven hospitals. With our significance level of alpha = 0.5, we reject all findings where the error probability is greater than 5%. This is done in the stepwise elimination procedure where covariates with the highest error probabilities, i.e. the highest P values are sorted out first. Our null hypothesis normally is that there is NO difference between the hospitals or that the number of fused segments has NO influence [1, 2]. This is why it is called NULL hypothesis. The alpha error or type-1 error is the probability to erroneously reject this hypothesis of no difference, i.e. to state that there is a difference though truly there is none. Mostly, the error probability of 5% ($P = 0.05$) is used. The opposite is a type-2 error or beta-error, which is erroneously accepting the null hypothesis of no difference though there truly is one. The type-2 error has to do with the so-called power of the study that directly depends on the sample size of a study. The power consideration is still overseen in many instances where researchers have not had any significant findings and conclude that there is no difference in the real world. Generally, a power of 0.8 is the target when sample size calculations are made for a study [1, 2]. Conducting an underpowered study is similarly unethical as conducting an overpowered study. In the first case, no conclusion can be drawn and time and resources were wasted for a worthless study, in the second case a statistically sound conclusion could have been drawn with less resources and patients.

In the case of a registry with an ongoing data collection and without clearly stated scientific hypotheses the power considerations are rather relevant when analyses are conducted that reveal no significant findings. It is less relevant for prospectively planning a sample size (though studies can be "nested" into this prospective data collection) but still helpful for calculating how many more observations would be needed in case an analysis was conducted which revealed no significant findings due to a small sample size.

Step 7: application of these results for the benchmarking procedure

The finding that center of intervention is a significant covariate already tells us that there is a significant difference in the probability for dura lesions in posterior spinal fusion surgery between at least two of the hospitals. We can guess that it is candidate number 4 and 1, but can we already stop here? No! The fact that "number of segments" also has a significant influence needs to be further pursued. What if hospital 1 predominantly operates cases with long fusions and hospital 4 mostly performs single level fusions? Then, the raw probabilities we are still looking at give us a skewed picture of clinical reality which is of disadvantage for hospital 1. There are two ways to tackle this problem and enable a correct comparison between the hospitals:

- stratification for "number of segments"
- adjustment to the same average number of segments for all hospitals

The variable "number of segments" has the following four outcomes: 1, 2–3, 4–5, >5. Stratification means that we will now separately compare the probability for dura lesions between the seven hospitals in cases with 1 level fusion, with 2–3 fusion levels, with 4–5 fusion levels, and finally with all fusions longer than 5 levels. That way, we neutralize the influence of the number of levels of fusion for each of the four comparisons and come closer to the true differences between the hospitals, but without being able to get an overall picture (Fig. 5).

For the mathematically interest readers:

Let π_{ij} be the probability for extension i and clinic j.

The logistic model is $\log(\pi_{ij})/\log(1-\pi_{ij}) = \mu + \alpha_i + \beta_j$

where μ is a constant, α_i the extension effect and β_j the clinic effect. With appropriate coding of effects (sum of effects equal to zero) μ becomes the overall mean and $\mu + \alpha_i$ the average within extension on the logit scale.

We can observe that the error bars around the point estimates get larger with an increasing number of levels of fusion. This is because the number of observations (the sample size) becomes smaller and hence the estimate of the true probability of dura lesions becomes less precise. Nevertheless, the four strata are the most precise comparison of the probability of a dura lesions between the seven hospitals for each situation. But what happens if a significant covariate has even more than four outcomes and if several covariates are significant? We would be faced with several dozens of possible stratifications. As this is impractical, we present a second possibility of such a correct comparison—adjustment to the same average number of fusions for all hospitals. This is an artificial mathematical procedure which becomes necessary because of the ordinal but non-numerical outcomes of the variable "segments of fusion". Figure 6 shows how the frequency of dura lesions increases with the number of fusion levels in the univariate model and why we cannot chose one

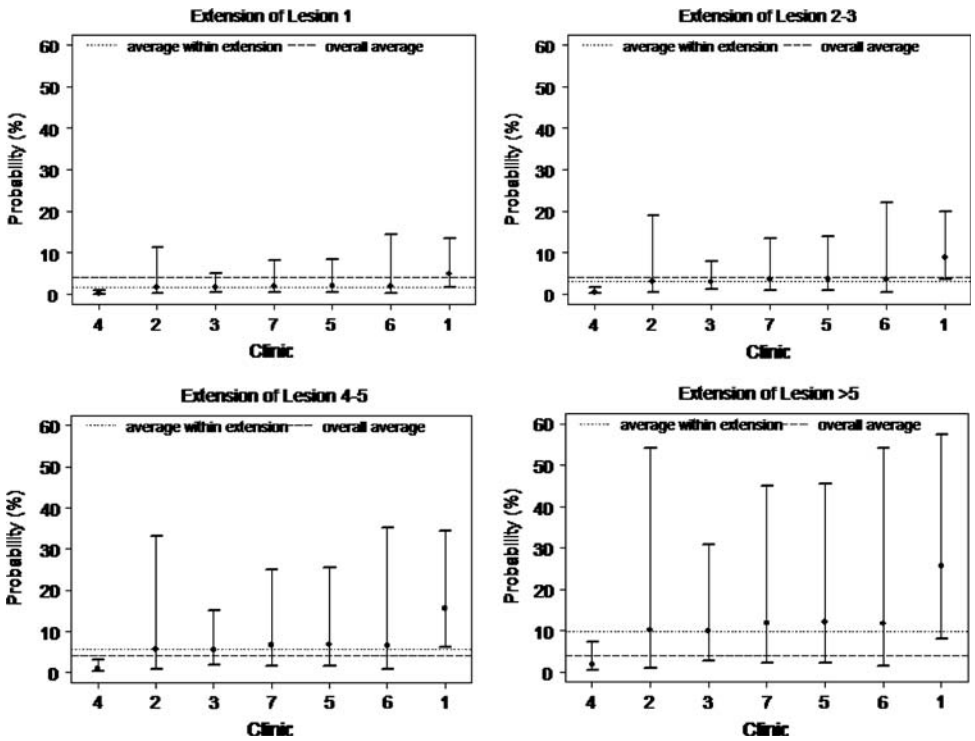

Fig. 5 Unadjusted but stratified (four outcomes) and sorted probabilities of dura lesions with standard errors in seven selected Spine Tango hospitals. In addition, the overall average probability of dura lesions (global benchmark) and the stratum specific average probability of dura lesions as stratum specific benchmark is displayed for the seven hospitals

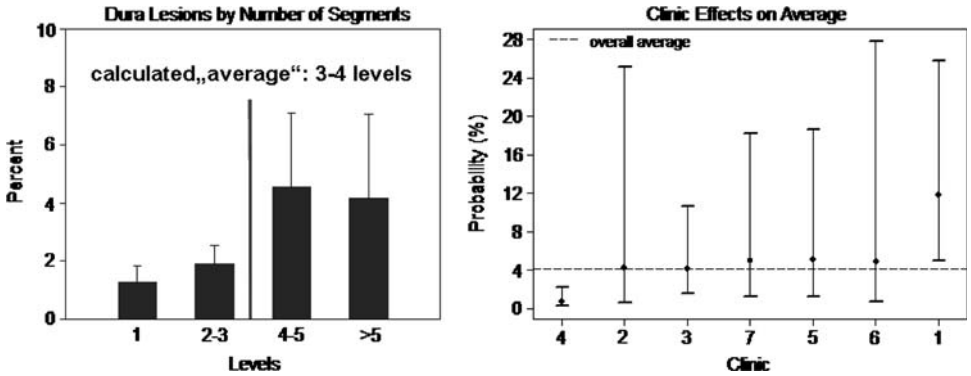

Fig. 6 Graphical visualization of the mathematical step for calculating a non-existing outcome level for dura lesions by number of segments (3–4 levels). Calculation and display of average (adjusted benchmark) and individual adjusted probability of dura lesions for the seven hospitals

outcome of the covariate "levels of fusion" as the average outcome and hence have to artificially create it.

The average number of dura lesions would be "3–4" and after mathematically creating the estimated probabilities with their error bars we have the best possible overall benchmark for comparing the seven hospitals. We do now not only see which hospitals are below, within and above the benchmark, we can also quantitatively describe the differences of probabilities. Hospital 4 still remains the reference hospital with OR = 1.0. The adjusted odds ratios for the other hospitals were now 6.0, 5.8, 7.1, 7.3, 7.0 and 18.2 which is different from those we had calculated previously (Fig. 4). Unlike the probabilities which are different for different levels of the covariates, odds ratios are the same for all levels of the covariates (as long as there are no interactions between the variables of interest and the covariates).

Conclusion

It is a long way to the true benchmark (Fig. 7)

Lessons learnt

The Spine Tango online statistical tool allows for direct comparisons of input variables and univariate, unadjusted comparisons of outcomes. A methodologically correct and true comparison can only be done after a multivariate analysis of other significantly influential covariates and adjustment for their influence by statistical modeling. Depending on number and type of these covariates it may be impossible to come up with one final benchmark. If for example, patient sex is revealed as significant covariate, we cannot mathematically create a benchmark for the average patient half man-half woman but would have to look at two types of comparisons: one for male and one for female patients. The Spine Tango data pool is constantly fed with new cases and the introduction of new surgical techniques or new types of implants may influence the interrelationships between outcomes and covariates. Consequently, the above-described analytical process has to be repeated periodically and adjustments have to be made for new covariates with a statistical influence. This highlights the fact that though the online benchmarking gives us a good but only rough idea about the truth, a methodologically correct analysis can never be automated but has to be manually conducted by an expert in the field. The current article and analysis has not considered the problems of data acquired in a voluntary observational registry like its reliability or biases that can be introduced. The acquired results shall serve as examples for the methodological limitations of the online benchmarking function and not as generalizable results regarding dura lesion rates in posterior spinal fusion surgery.

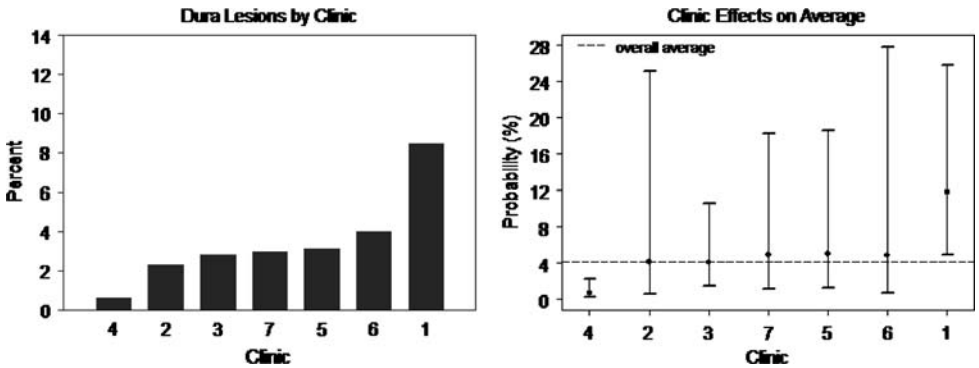

Fig. 7 Contrasting the raw display of proportions of dura lesions with the adjusted probabilities and adjusted benchmark of dura lesions

Conflict of interest statement None of the authors has any potential conflict of interest.

References

1. Hüsler J, Zimmermann H (2005) Statistical principles for medical research projects (Book in German). Verlag Hans Huber, Bern, Switzerland
2. Kirkwood BR, Sterne JAC (2003) Essential medical statistics, 2nd edn. Blackwell, Massachusetts
3. Kleinbaum DG (1992) Logistic regression. a self-learning text. Springer, New York
4. Melloh M, Staub L, Aghayev E, Zweig T, Theis J, Barz T, Chavanne A, Grob D, Aebi M, Roeder C (2008) The international spine registry SPINE TANGO: Status quo and first results. Sep;17(9):1201–1209. Epub 2008 Apr 30