Author’s response to reviews

Title: Should policy-makers and managers trust PSI? An empirical validation study of five Patient Safety Indicators in a National Health Service

Authors:

Enrique Bernal-Delgado (ebernal.iacs@aragon.es)
Sandra García-Armesto (sgarciaarm.iacs@aragon.es)
Natalia Martínez-Lizaga (nmartinez.iacs@aragon.es)
Begoña Abadía-Taira (mbabadia.iacs@aragon.es)
Joaquín Beltrán-Peribañez (jbeltranp@aragon.es)
Salvador Peiró-Moreno (peiro_bor@gva.es)

Version: 2 Date: 23 December 2011

Author’s response to reviews: see over
Dear editor,

Hereinafter you will find enclosed a detailed response to each one of the queries posed on by the reviewers. The work has benefited from their comments, suggestions and criticisms.

As required, a new version of the manuscript, copy-edited by a native English speaker has been uploaded as well as, new versions of figures and additional files.

We are looking forward to hearing from you.

Sincerely,

Enrique Bernal-Delgado MD PhD
**Query #1:** P3, para 2: “c) Are measurements precise enough...?” Do you mean in terms of statistical power or amount of noise or something else?

- Response: Understood as a relative lack of random error.

**Query #2:** P5: the term “adjusted cumulative incidences” is used – why are they “cumulative”? The rates seem to be adjusted incidences.

- Response: The reviewer is correct; it is a mistake dragged along the text by the fact that MLM incidence is not adjusted. It has been corrected.

**Query #3:** P8: OR≥2 is used as a cut-off. This seems a bit large to me, as several comorbidities in the Charlson and Elixhauser indices, for example, have smaller ORs. Are the results still the same if you simply retain those variables with e.g. p<0.05?

- Response: We agree with the reviewer’s comment on the largeness of the cut-off. It is a discretionary boundary based on the need of reducing spurious findings on massive datasets along with increasing the clinical relevance of those findings. Nevertheless, we studied the potential influence of a different threshold (OR>1.5) both on the “area under the curve” and on the rho statistic, and differences were negligible.

For the purpose of responding specifically to the reviewer’s query, we replicated the analyses considering –as suggested- all the Elixhauser comorbidities with p value less than 0.05. The differences between the model included in the article and that one obtained by this new strategy showed meaningless differences in terms of ROC curve and negligible differences in terms of rho statistic; figures are shown in the table below.

<table>
<thead>
<tr>
<th>PSI modelled</th>
<th>ROC curve</th>
<th>Rho statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR &gt; 2 1</td>
<td>OR &gt; 2 2</td>
</tr>
<tr>
<td></td>
<td>P &lt; 0.05 2</td>
<td>P &lt; 0.05 5</td>
</tr>
<tr>
<td>Decubitus Ulcer</td>
<td>0.79</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>0.80</td>
<td>0.10</td>
</tr>
<tr>
<td>Catheter- related infection</td>
<td>0.68</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td>0.24</td>
</tr>
<tr>
<td>Postoperative PE or DVT</td>
<td>0.68</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.71</td>
<td>0.05</td>
</tr>
<tr>
<td>Postoperative sepsis</td>
<td>0.64</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.72</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Note: 1. Original scenario; 2. Suggested by the reviewer.

**Query #4:** P8: the c statistic is not a measure of goodness of fit, as stated.

- Response: Although there is some controversy on this point, we have preferred to follow the denomination used in Hosmer DW, Lemeshow S. Applied Logistic
Regression. John Wiley & Sons, Inc. 2000: New York. 2nd edition. In this book, both the proper Hosmer-Lemeshow c-statistic and the “area under the curve”, are included within the tools worth testing goodness of fit (chapter 5 section 2). In our work, both were used to judge the best models, although the main driver was ROC curve given the excessive “sensitivity” of the Hosmer-Lemeshow test when analysing big samples.

Query #5: P10, data sources: were ALL publicly funded hospitals in Spain included, even specialist ones whose casemix will be very different from the average? I didn’t know there was a central database for hospital admissions in Spain – how does this overcome the very regionalised administration?

• Response: Yes, it is correct. All the hospitals included were publicly funded. The Central Government maintains a central repository with all the discharges produced by hospital providers across the country.

Nevertheless, as expressed in the acknowledgments of the article, for the purpose of this work we have used a different dataset; that hold by the Atlas VPM project, a national-wide research initiative involving the universe of public hospital admissions in the country since 2002 (a more detailed description can be found on www.atlasvpm.org).

Query #6: P10, data sources: up to 30 secondary diagnoses can be recorded. It would be useful to know the typical (e.g. mean) coding depth, in order to compare internationally, as this will clearly influence the PSI rates.

• Response: The table below shows the distribution (percentiles) of the number of secondary diagnoses recorded, by PSI. As observed, although it is possible to record up to 30 conditions, 95 percent of the discharges have 11 or less secondary diagnoses.

<table>
<thead>
<tr>
<th></th>
<th>MLM</th>
<th>DU</th>
<th>CRI</th>
<th>PE-DVT</th>
<th>PS</th>
</tr>
</thead>
<tbody>
<tr>
<td>P5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>P25</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>P50</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>P75</td>
<td>4</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>P90</td>
<td>6</td>
<td>10</td>
<td>8</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>P95</td>
<td>7</td>
<td>11</td>
<td>10</td>
<td>10</td>
<td>11</td>
</tr>
</tbody>
</table>

Query #7: P12, first line: do you mean the rho rather than c statistic for the % of variance explained by patient factors? It would in fact be interesting to know the c statistics for the risk adjustment models.
• Response: The first values, those referred to the variance explained by patient characteristic are ROC values—it is written “Although most of the variance was explained by patient-related factors ranging from 64% in PS to 79% in DU (…)”. Unfortunately, we missed a specific reference to the rho statistic in the second line, misleading the reader. We have eased the reading by including a specific reference to the “area under the curve” and the rho statistic values, as follows:

“(…) Although most of the variance was explained by patient-related factors ranging from 64% in PS to 79% in DU in accordance to the area under the curve, still a significant proportion of the variance was explained by the hospital: from a small rho value of 6% in the case of MLM (CI95%: 3% to 11%) to a high rho value of 24% (CI95%: 20% to 30%) in CRI”.

Query #8: P12, bottom: the “underperformers” are mentioned, but it is important to say what % are “good performers”. For instance, poor coding will lead to a hospital having a low PSI rate. Asymmetry would give us another clue to the reliability of the recording.

• Response: Usually, PSI analyses are certainly more concerned on high rates of adverse events; this is certainly our case. But, the reviewer is correct, good performers could be used as gold-reference if no systematic poor-coding practices exist.

I do have some doubts with regard to the idea of symmetry or asymmetry providing additional insight. In fact, hospitals are ranked following shrunken residuals (random effect in the multilevel analysis), which follow a Gaussian distribution (mean equals 0).

On the other, asymmetry could be found for statistical reasons, just because the proportion of hospitals below the line (hypothetical good performers) are actually hospitals with larger standard errors (i.e. more difficult to find differences if they exist). As observed in the table below, in our dataset, except in DU, the remaining PSI showed this phenomenon, being particularly evident in MLM-DRG.

<table>
<thead>
<tr>
<th>PSI modelled</th>
<th>Number of hospitals statistically different to the expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>below the expected</td>
</tr>
<tr>
<td>Mortality in low-mortality DRGs</td>
<td>1</td>
</tr>
<tr>
<td>Decubitus Ulcer</td>
<td>54</td>
</tr>
<tr>
<td>Catheter-related infection</td>
<td>24</td>
</tr>
<tr>
<td>Postoperative PE or DVT</td>
<td>27</td>
</tr>
<tr>
<td>Postoperative sepsis</td>
<td>27</td>
</tr>
</tbody>
</table>
In any case, and on a different point, by using our data we are not able to confirm whether these hospitals statistically below 0 are there because they are poor coders. However, as long as we have learnt from a criterion-validity work carried out in Spain (quoted in the article), the risk of underreporting is pretty much related to decubitus ulcer.

**Query #9:** P13, near the bottom: “this distribution is expected to be centred on zero”. If you are referring to the PSI rate, then this will of course not be true, as rates lower than zero are impossible. Do you mean the log of the rates or something else?

- **Response:** The expression is a mistake. As the reviewer comments, this paragraph is referred to incidence of adverse events, not to the Empirical Bayes statistic in the next paragraph. The proper expression is close to zero, and it has been changed in the text.

**Query #10:** P14: the comment on whether variation is attributed to the patient or the hospital in the case where risk adjustment is limited is a good one.

- **Response:** Thanks for the comment

**Query #11:** P16, subheading: “detect hospital performing above the expected”. Do you mean “better performing” (i.e. with low rates) or with higher PSI rates? You need to be clear here and elsewhere what “above expected” means.

- **Response:** Thanks for the suggestion. The subheading has been changed and is clearer. “Are PSI sensitive enough to detect hospitals with rates above the expected?”

**Query #12:** P17, line 2 (also lower down on p17): “proven sensitive enough to detect hospitals above the expected”. I don’t think one can make this conclusion. Just because a graph has some statistical outliers does not help us estimate the sensitivity of the indicator. For that, we would need to know which hospitals are actual outliers (from some ideal data source) and we don’t. I wouldn’t use the word “sensitivity” for this reason – power might be better (esp in the Conclusion).

- **Response:** The term “sensitivity” has not been used in these paragraphs as part of the analytical strategy devoted to evaluate the robustness of the measures; specifically, the influence of outlier values on the estimates. As detailed in the main endpoints section Sensitivity was defined as the “statistically significant difference between the observed and the expected, as provided by the residual analyses in a multilevel approach”. Thus, it is more on the ability for the multilevel residual analysis to flag true cases –hospitals with higher PSI rates
than the expected. It is true that the term is used in a heterodox manner—there is not a gold standard to compare with—so we are basically relying on the precision of the estimates.

For the sake of reducing the risks of misunderstanding, some of the uses of the term “sensitivity” have been substituted by “ability” or “precision.”

Query #13: P18: if hospitals do not “properly report secondary diagnoses” they will more likely be flagged as good, not bad performers.

• Response: The reviewer is correct: incidence of adverse events is affected by the comprehensiveness in coding the concerned diagnoses—pressure ulcer, sepsis, etc.; thus, poorer coding practices affecting adverse events will flag underperforming hospitals as good performers. In addition, as mentioned in the discussion section, this is probably true in the case of decubitus ulcer, where false negative cases are more frequent.

    However, along the paragraph under discussion, we wanted to stress a different point. How a poor risk adjustment would affect the most complex hospitals. From our perspective, if secondary diagnoses affecting Elixhauser comorbidities are not properly recorded, when adjusting risks, we will observe a smaller smoothing effect than that expected by their complexity. Thus, the difference between the observed (estimated residual for each complex hospital) as compared to the expected (the average estimated incidence for the whole sample of hospitals) will increase, increasing the probability for a complex hospital to be flag as a bad performer.

Query #14: Table 1: for the EB figures, it would be helpful to say what a “high” value looks like as most people are unfamiliar with this statistic. The same would be useful for rho (could give a footnote).

• Response: Thanks for the comment we have added the following notes to both tables 1, and 2:

    Empirical Bayes statistic is a measure of systematic variation—variation beyond chance. A value different to 0 would represent systematic variation; as for its magnitude, the higher the value the more the systematic variation.

    A Rho statistic value different to 0 represents the existence of cluster effect—the propensity of having an outcome is more similar among the patients within a hospital, that among patients from different hospitals; as for the magnitude of rho, the more the value, the greater the clustering.

Query #15: Fig 1, Y axis label: I think this should be log of adjusted risk, as it’s not a log scale
Response: Thanks for the comment. This is a mistake. We have changed the label.

Query #16: Fig 2: this needs axis labels. Also, I wonder if a funnel plot would be more useful as it would show the outliers better. Alternatively, you could superimpose horizontal lines to show the threshold values for “high” and “low” residuals.

Response: Thanks for the comment. We have added axis labels. With regard to the comment on funnel plot, we agree on its potential virtues, although in this work we wanted to use caterpillar graphs representing the residuals distribution, as part of the multilevel approach (sort of an orthodox methodological exercise). On the other hand, using shrunken residuals caterpillar graphs, it is easy to identify outliers looking at those dots at the two ends, breaking the line-trend – since they are supposed to follow a Gaussian distribution.

Nevertheless, the suggestion on funnel plot is a very important one; just a simple comment on that. In our experience, a previous judgement on the existence of over-dispersion is needed to estimate the confidence intervals – estimates can follow either a Poisson or Negative-Binomial distribution.
Query #1: The authors must document whether the PSIs they studied identified only those adverse events that were acquired during hospitalization or whether they included adverse events that were present on admission. (From the comment in discussion section about decubitus ulcers, it appears that the adverse events that were present on admission were included). If the PSIs include adverse events that were present on admission, then these events would have contributed to systematic differences in rates between hospitals. The authors should evaluate the impact of the events that were present on admission on these differences. I don’t think the tests of the influence of patient comorbid conditions and hospital case mix index would have uncovered the impact.

- Response: “Present on admission” (POA) conditions are not available in Spain at the moment. Our work is just referred to conditions acquired along the hospital stay. The comment on the decubitus ulcer in page #19 is specifically referred to a different work led by ourselves, where POA conditions were assessed by using a specific nursing-care registry taken as gold-standard.

Query #2: In the discussion section, the authors were careful to point out the impact of incomplete reporting of secondary diagnoses and the need to compare "like" hospitals. They also commented on the validity of the PSIs, which has been tested to some degree by the Spanish NHS. However, in their closing remarks, the authors recommend that the PSI can be used as a screening tool to "identify those centres from which best practice lessons can be drawn out and those where intervention is clearly needed." Given the study results, I recommend that the authors be specific about how results that show some centres with low O:E ratios and others with high O:E ratios should be interpreted and also provide more specific guidance to centres that have these different results. It is possible that centres with high O:E ratios may not be adequately reporting secondary diagnoses and so they should improve the documentation and coding of patient comorbidities. In this scenario, the intervention is about coding, not improvements in patient safety. Yet, the authors use of the term "intervention" seems misleading because it suggests that centres with high O:E ratios have an unfavorable patient safety performance.

- Response. The discussion on the number of secondary diagnoses sought to assess in what way under/over-reporting might influence risk-adjustment; specifically, how the number of secondary diagnoses affects beta coefficients in Elixhauser comorbidities, and how it affects cluster estimates -rho values and random effects. Its influence on the latter one, the random effect, is critical since it represents the difference between the residual attributed to each hospital compared to the estimated average; ultimately, the difference between the observed and the expected, following the article terminology, for each hospital.
As detailed in additional file #1, after modelling ndx, a reduction in Elixhauser comorbidities b coefficients was observed, suggesting that the number of secondary diagnoses absorbed part of the variance. Thus, as suggested by the reviewer, change in the residual effect for each hospital is also expected.

For the sake of answering this query, a two-step analysis was carried out: 1) the first step, looking at the correlation between the two random effects estimation (those referred to the model without and the model with the number of secondary diagnoses); hypothetically, the larger the correlation the lesser the influence of coding-practices; 2) the second one, complementary to the last one, looking at to what extent hospitals observed as statistically different to the expected, turn into non-statistically different and the other way round; hypothetically, the lesser the number of hospitals changing, the lesser the influence of coding intensity.

An excellent correlation (Pearson coefficient values) between the two estimations was observed for the whole sample of hospitals: 0.83 in post-operative sepsis, 0.86 in post-operative PE-DVT, 0.94 in decubitus ulcer and 0.96 in Catheter related infection.

When it came to determine whether the estimation yielded changes in the statistical significance of the random effect, (i.e. a hospital found as statistically different that average turned into statistically similar, and the other way round) null or negligible changes were observed, except in the case of decubitus ulcer where, a noticeable number of hospitals changed their status. As observed in the table below, this change affected more to those hospitals below the average, in both ways: 9 hospitals turn to non-statistically different than the average and 13, became statistically different.

<table>
<thead>
<tr>
<th>PSI modelled</th>
<th>Number of hospitals changing to an “equals to average status”</th>
<th>Number of hospitals changing to a “statistical significance”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>below 0</td>
<td>Above 0</td>
</tr>
<tr>
<td>Decubitus Ulcer</td>
<td>9  (54)</td>
<td>3  (47)</td>
</tr>
<tr>
<td>Catheter- related infection</td>
<td>0  (24)</td>
<td>0  (55)</td>
</tr>
<tr>
<td>Postoperative PE or DVT</td>
<td>0  (29)</td>
<td>0  (39)</td>
</tr>
<tr>
<td>Postoperative sepsis</td>
<td>0  (31)</td>
<td>0  (50)</td>
</tr>
</tbody>
</table>

Some of these reflections have been included into the text:

a) As for the subheading of the section, a less misleading one would be: Are results dependant on the coding practices affecting Elixhauser comorbidities?
b) A paragraph has been added at the end of this section:

Given that the number of secondary diagnoses absorbed part of the variance in the new model and beta coefficients changed, variation is also expected in the random effects estimation for each hospital. However, an excellent correlation (Pearson coefficient values) between the original random effects and the new ones was found: 0.83 in post-operative sepsis, 0.86 in post-operative PE-DVT, 0.94 in decubitus ulcer and 0.96 in Catheter-related infection. On the other hand, except in the case of decubitus ulcer the changes in the statistical nature of the random effect (i.e. hospitals found statistically different than the expected turned into statistically similar, and the other way round) were null or negligible.

c) With regard to the word “intervention”, except for the case of decubitus ulcer, in our opinion, these findings –along with those from the other aforementioned work on criterion validity- support the need of a patient safety intervention. Nevertheless, in order to reduce misinterpretation, although we have continued to use the term intervention, we have added the following clause in the closing remarks specifying what happens with decubitus ulcer:

(...) and risk of hospitals misclassification in decubitus ulcer remained.

Query #3: In the methods section, the authors should explain why they chose to analyze the 5 selected PSIs and why they chose not to analyze other PSIs or even all PSIs.

- Response: This work, and the decision about the 5 PSI chosen, is based on a previous work [quote 23] where all PSI were empirically assessed for the Spanish case. We have included a short reference in the methods section.

Query #4: In the discussion section, the authors apply the lessons learned from this analysis of 5 selected PSIs to all PSIs. It would be helpful if they described how and why the analysis of 5 selected PSIs can be generalized to all PSIs.

- Response: There was no intention –it is completely inappropriate- to suggest that what is likely true for these five PSI, is also true for the remaining. After reading the discussion section and closing paragraphs, and taking into account the suggestion, we have reviewed and added terms intended to reduce the risk of misunderstanding; thus, we have included terms as “these five PSI”, “the studied PSI”. We have kept, though, the original subheading “Should policymakers and managers trust PSI?” because we wanted to be consistent with the question posed on in the paper’s title. We think that although used as a general term, it won’t mislead readers.
Query #5: In the results and discussion sections, the authors should provide a specific description of why the analysis of the PSI for death in low mortality DRGs produced results that were different from those of the other 4 PSIs and the implications of these differences.

- Response: Mortality in low mortality DRG has two main differences with regard to the other studied PSI. First, it is a quasi-sentinel event and therefore, it would be inappropriate to adjust its “rates”. Second, it is the most infrequent one, and it is difficult to find hospitals with more or less cases than expected, because of its large standard errors. These two questions have already been commented in the methods, results and discussion sections.

To take into consideration the comment by the reviewer, we have added some insight when discussing the cluster effect in those more complex hospitals, and within the closing remarks. As for the first, the original text in pg #17 was substituted by the following “As observed, except in the case of MLM where heterogeneity across hospitals was the underlying reason for results (just 4 out of 47 hospitals were statistically above the expected in this second analysis) (...)”. As for the second, a specific remark was included in the conclusion section; thus: “However, ability to flag hospitals beyond the expected was limited in Mortality in Low-Mortality DRGs due to its larger standard errors”.

Query #6: In the discussion section, the authors should document the limitations of their study and of their results.

- Response: As a validation study, the article was designed as a Q&A exercise, responding to specific questions related to the measurement properties of several PSI -namely empirical properties. So, it is not the classical cross-section observational study looking at estimating the true rate of adverse events, which as observed by the reviewer requires a specific discussion on the internal validity of the estimates.

Taking into account this perspective, we think that, as an empirical validation exercise, the whole article is a debate around determining in what way the empirical properties of five PSI, in a particular dataset, allow potential users a proper utilization. So, this article –and specifically the discussion- is all about documenting potential flaws, e.g. whether the findings are spurious or not, the effect of extra-variance attributed to heterogeneity across hospitals, the effect of secondary diagnoses coding intensity, the effect of outliers, the improvement of poor risk adjustment by using natural-pooling risk when analysing complex hospitals, etc.

Nevertheless, some additional comments were made on the relevance of carrying out time-series analysis to know more on PSI empirical properties (beyond the scope of our article), and, on the need of taking into account criterion-validity information, as a requirement for PSI appropriate use.
Query #7: The authors should document the version of AHRQ PSI they are using.

- Response: Thanks for the comment. As mentioned we used a Spanish version of the AHRQ indicators. So all the analyses are based on the cases retrieved by using the “Spanish algorithms”. We could allow the reviewers to access the algorithms if needed. In any case, a reference to the ARQH version 4.1 has been added to the methods section.