Author's response to reviews

Title: Is There Much Variation in Variation? Revisiting Statistics of Small Area Analysis in Health Services Research

Authors:

Berta Ibañez (ibanez@bioef.org)
Julián Librero (julian.librero@uv.es)
Enrique Bernal-Delgado (ebernal.iacs@aragon.es)
Salvador Peiró (peiro_bor@gva.es)
Beatriz González López-Valcárcel (bvalcarcel@dmc.ulpgc.es)
Natalia Martínez (nmartinez.iacs@aragon.es)
Felipe Aizpuru (faizpuru@htxa.osakidetza.net)

Version: 2 Date: 21 December 2008

Author's response to reviews: see over
Dear Editor,

Please find enclosed our reflections on the comments by the reviewers and a revised version of our manuscript "Is There Much Variation in Variation? Revisiting Statistics of Small Area Variation in Health Services Research" [Manuscript ID 1086417261222518]", submitted for consideration in BMC HEALTH SERVICES RESEARCH.

We apologize for the delay in sending the new version of the manuscript. Including some of the reviewers' suggestions has involved substantial re-analysis, and making these changes has required more time than originally foreseen.

We very much appreciate the comments made by the referees. Their comments have been truly very useful in improving our paper.

Thank you for your interest.

Yours faithfully,

On behalf of all the authors,

Enrique Bernal-Delgado
Reviewers’ responses

In response to the concerns and suggestions made by the referees, we have made the following reflections and changes:

Reviewer: Peter Congdon; Comment #1. This is a generally worthwhile and interesting paper as far as it goes, and one which does some service in terms of popularising statistical methods. But in my view it needs to include a wider frame of reference, be more explicit in the models implicitly used, and to acknowledge some limitations in the framework adopted. I set out my issues with the paper below. I would regard responding to these points as involving major compulsory revisions.

We appreciate the comment about the divulging labour of our manuscript, as it is in fact one of the main (though unuttered) aims of our research. This paper is the first of three that we are working on trying to depict some weakness of the methodology currently used in “Medical Practice Variations” analysis (also named Small Area Variation Analysis –SAVA- in the health services research approaches), and attempt to overcome them using the more sophisticated methods that have been developed, for instance, in the so-called disease mapping setting.

The lack of recent works proposing and evaluating statistics in the SAVA methods is the main motivation of this paper. Most recent papers in geographical variation, even in this journal (see, for instance, Román et al, 2008) use the Extremal Quotient and the Coefficient of Variation as the only measures of geographical variation. Even in the Dartmouth Atlas which has contributed enormously to the advances in medical practice variations analysis and it is one of the few using the CSV statistic, supports its assertions on the Extremal Quotient statistics an absolute quantity which, depending on the instability of the rates, may not mean that much (see Wennberg et al., 2008). Graphical presentations of variability are normally limited to distribution graphs (see, for instance, Volinn et al, 1994, and many graphical displays of the Dartmouth Atlas showing the variability in the rates, and these are even used to visually compare variability in procedure rates, without mentioning that sometimes this apparent variability depends more on the

---

rate of the procedure and on the instability of low-populated areas than on the real variability itself.

We agree on the fact that a wider frame is needed, a framework that cope with some other possible casuistic such as the presence of spatial correlation structure of the data and the instability of some rate estimates when tackling with very small areas. However, given the low use of Statistics of Variation and models on geographical variation in health service research, we consider that to widen the frame in just one paper would not do any favour to the aim of divulging these techniques, since the article in the present form can be slightly difficult to follow for many interested readers but not very familiar with simulation studies.

Reviewer: Peter Congdon; Comment #2. WHAT DOES THE NULL HYPOTHESIS OF NO VARIATION REALLY MEAN AND WHERE DOES IT COME FROM?

A first concern is with regard to the oft-stated goal of the paper to assess a null hypothesis of no variation, or as stated on page 5, “determine whether [there is] higher than random expected variability” or on page 7 “whether variation between areas is higher than would be expected by chance”. The notion that any dataset has a particular “expected” level of variability, or one expected by chance, would be unfamiliar to statisticians. I would say that an expected level of variability is only possible to state under a particular density assumption.

Following on from this, the authors should in my view, acknowledge that the underlying rationale for the test statistics they apply is a health area equivalent of classical meta-analysis for normal data (e.g. for pooling effects over trials or studies) in which a common mean applies across studies. In particular the implicit baseline model in their paper (providing the “null hypothesis” of no variability) is a Poisson with a common mean across all areas or a binomial with a common probability.

Is it then true to say that what the paper is really assessing is the presence or otherwise of Poisson or binomial overdispersion (also known as extravariation)?

Coming at the problem a new (without being tied into the “Small Area Variation Analysis” template), one might equally propose as a baseline model for area health data a negative binomial or beta-binomial model. Could there still be more than “expected variability” under such an alternative density assumption.

We agree with the reviewer that a particular density assumption is needed to represent no variation. Our starting point from the SAVA framework was the set of works proposed by Diehr et al and the reasoning underlying the construction of the SCV statistic (see McPherson et al., 1982). All these works assumed, under the

---

null hypothesis of no variation, a Poisson distribution with a common risk for the counts of the whole region under study or a normal distribution with a common rate for the rates of that region. Within the different but close setting of disease mapping more recent works (see, for instance, Richardson et al, 2004, 6 Ugarte et al, 20087) design an scenario with the same homogeneous Poisson model to simulate an underlying homogeneous risk surface, and a set of regions are induced with an elevated risk to assess the ability of different techniques on detecting this elevated risk, maintaining the homogeneous risk surface in the remaining areas. Hence, here we have adopted the underlying distribution assumed in most studies in this framework to represent the null hypothesis of “homogeneous risk surfaces”.

Other alternatives such as the negative binomial, the beta-binomial or even the poisson-log normal models as baseline models could have been used as the reviewer suggest. In fact, they would have allowed to represent that even under the null hypothesis of no variation in medical practice, the level of health can vary across the whole area not only because of the age and sex structure variations in population, but also because other -unknown or not accounted- factors (see, for instance, Cain and Diehr, 19918). However, the use of these baseline models would have required a reliable estimate of which would have been the variability attributable to disease level variation to model specifications purpose, because all these models have additional parameters that represent the extra-variability allowed.

Unfortunately, variability in rates under “no variation in medical practice” can not be derived from any data set, as it does not exist a data set that represents “no variation in medical practice”. Recent works have tackled with the problem of ignoring variation in disease prevalence (see, for instance, Shwartz et al, 20059) by using information of outpatient registers to represent disease variability. These type of registers could be used to emulate variability in rates that is not attributable to medical practice, but unfortunately, at the moment the information available is not reliable enough in our setting.

In order to include this points, we have enlarged the analysis section to state explicitly which is the underlying model under the null hypothesis, where it comes from, and how the most frequently used methods assessing variability, such as the SCV, assume this underlying model in their calculations. Apart from that, we have

---


added in the paragraph devoted to explain the assessment of the null hypothesis which is the null model considered, and furthermore, in the discussion section we have added as a limitation that the model adopted to emulate the null hypothesis of no variation is one among others that could have been chosen, it does not consider the variability that may be present due to disease prevalence variation, and that other assumptions could have led to slightly different results.

Reviewer: Peter Congdon; Comment #3. ARE THERE DIFFERENT TYPES OF VARIABILITY.

The paper mentions on page 5 that “geography is destiny” and refers on page 7 to the Dean and Bohning statistic being applied to test “geographic patterns in neighbouring area studies”. On page 11 there is a reference to “geographic variability”.

This is really skating over the real relevance of geography. There is no recognition of the relevance of spatial correlation or of the distinction between spatially structured variability and pure white noise, well known in Bayesian disease mapping following the 1991 paper by Besag et al. In particular, what the authors term the “Empirical Bayes statistic” is estimating a composite of structured and unstructured variability.

So I think the authors should at least set all the SAVA references in this wider literature and methods context, and acknowledge really that their perspective has some limitation in not incorporating “space” or “geographic” effects. The SAVA method apparently treats areas as uncorrelated.

This is an important point we worked in before deriving the last version of this manuscript. As the reviewer point out, SAVA methods usually treat areas as uncorrelated, and the effect of this assumption on the estimation of variability in our setting was the first one to be analyzed. Being the popularization of methods one of our main aims, we did fit four methods adopted from the disease mapping setting to the six datasets analyzed here. This methods were the classical method (the Standardize Utility Ratio yi/ei, analogous to the Standardize Mortality Ratio), the Poisson discrete mixture model proposed by Schlatman and Bohning (1993)\textsuperscript{10}, the Poisson-lognormal model, also known as unexchangeable model or Independent Prior model (which is the baseline model for what we called here EB statistics, as we explain in this new version of the manuscript, see for instance MacNab et al, 2006)\textsuperscript{11}, and the so-called BYM model mentioned by the reviewer


(Besag et al, 1991)\textsuperscript{12} that allow for the spatial autocorrelation among risk areas. Together with the fitting of the models, the score test of spatial dependence (Ugarte et al, 2003)\textsuperscript{13} derived under the Leroux model (Leroux, 1999)\textsuperscript{14} was also used, and results showed that significant spatial autocorrelation was present in “Hip Fracture”, “Knee replacement” and “Pacemaker” data sets.

However, when comparing the risk estimates among the exchangeable and the BYM models, they were practically equivalent in the second decimal, and results presenting both the spatial and unexchangeable models in a map were identical. As an example, Figure 1 presents the results obtained for Pacemaker, which is the only procedure with both low prevalence rate and significant spatial autocorrelation, more sensible to unstable estimates under the classical model. This equivalence among model results occurs because the areas considered in Variation in Medical Practice studies are usually big enough as to obtain relatively stable estimates for the risk. For instance, for Pacemaker, the mean number of expected cases per area is about 40, and the advantage of considering neighbouring risks on estimating risks for each area do not have a practical effect in these studies.

On the other hand, the advantage of having the two sources of variability estimated separately (unstructured and spatially structured variability) did not contribute importantly to the interpretation of the results, as researchers in health services are more interesting in depicting how much of the observed variability can be attributed to medical practice itself (the physician’s practice style in each area) than in depicting how much of the observed variability can be attributed to spatial dependence. Taking into account all this, and given the characteristics of the current atlases in medical practice, we decided not to focus on the geographic correlation structure of the rates in this paper, although we are presently incorporating the BYM modelling in other applications which consider smaller geographical units such as Basic Healthcare Zones in stead of Healthcare Areas.

With regard to this aspect, we have added in the included paragraph on the analysis section some of the aforementioned relevant bibliography on spatial modelling, as well as some of the most widely used estimation procedures such as the full-bayes approach. Further, in the discussion section we have added why we did not accounted for it and mentioned the advantages of incorporating these models in studies of variation in medical practice.


Reviewer: Peter Congdon; Comment #4. HOW WOULD GENUINELY SMALL AREAS AFFECT THE FINDINGS. The authors refer to 147 areas covering 75% of Spain’s population. These are not truly small areas, as might be of concern to someone assessing variations within local health agencies. How would the power simulations (for example) be affected if the areas averaged 10,000 in population, rather than 200 thousand.

We agree with the reviewer that these are not truly “small areas” as they are understood in the disease mapping setting, although in health service research, small area variations are defined as “large differences in the rates of use of medical services (for instance, hospital admissions and surgical or diagnostic procedures) between geographic regions” (see Health Service Research group, 1992)\(^{15}\), and there is no mention to the mean population per area. Studies on medical practice variations are usually based on larger areas as the main focus is on practice style within hospital referral areas. In studying mortality or morbidity maps, most studies use smaller areas such as basic health areas or even census sections, because interest is usually on deriving the risk surface and to relate these outcomes with socio-demographic or environmental characteristics of the area, but this is not the main focus when hospital admissions are evaluated. For instance, the Spanish health areas are served for only one hospital that usually have higher market quotes in their area (between 80-95% of all hospital admission from the

area residents are realized in the hospital of its own area) and health services researchers are interested in the behaviour of this hospitals. In fact, health services researchers assume that morbidity differences in the adjusted by age and sex population of these large hospitals areas can explain some variation (for example the two fold hospitalization rates found in the hip fracture or in the myocardial infarction admissions), but not the more large differences shown for other conditions.

In order to explain how the power simulations would be affected if the population per area was lower, we have repeated the simulation study assuming that the population structure was the same as that observed divided by 10. We did not divide by 20 as you suggest because for low rate procedures such as lower extreme amputation, this would have produce a high quantity of regions with less than 2 cases per area, and for that case it is well known that even the BYM modelling would have not been appropriate. In this new setting, the power of the statistics is a little bit lower (see Figure 2), in that the detection of variability when it is present is slightly lower for a given alternative scenario because of the instability of the rates.

For instance, if with the six scenarios generated we achieved a power equal to one in S6 in the real setting for all statistics, in these new setting, other three scenarios have been generated to attain power equal to one, namely S7, S8 and S9, with 10, 20 and 40 regions with elevated risk equal to 3 respectively. Note that a vertical line on S6 has been added to facilitate the comparison among this results and previously derived findings. It is worth mentioning that the order of the statistics as a function of their power remains quite stable as compared to the real setting, being the EB, $X^2$, SCV5-95, Bohning and Dean the most powerful and the EQ5-95 and the CV the less powerful, and also de differences between a low-rate procedure and a high rate procedure remain the same.

Anyway, and as commented, SAV analyses in health services research usually use large hospital areas and, in fact, our areas about 100.000 persons are not so great in this context (for example, the Dartmouth Atlas usually use only 300 Hospital Referrals Regions for mapping all the United States).
We have added in the discussion section that our simulation scenario does not cover all the possibilities that the analyst may meet, and that some scenarios with a different population structure could give different results. Additionally, and taking into account that this limits the inference one may be attempted to do, the conclusion given in the abstract has been slightly changed.

Reviewer: Peter Congdon; Comment #5. IS THERE A NORM BEING USED FOR “LOW...HIGH” USAGE OR VARIABILITY? This ties in with my point above relating to excess variability. But how do the authors decide what is a low or high utilisation rate. Is this compared to the Spain average over all procedures?

We agree with the reviewer that it is difficult to sort procedures according to their variability when we do not have one aprioristic gold standard measure to make variability comparable among procedures, but the use of “low variability procedure”, or “high variability procedure” labels were useful to simplify the understanding of the text. Our experience with the whole set of procedures analyzed until now in the Spanish atlas project (about 35 procedures), together with some findings in the literature (for instance, hip fracture is frequently used as a procedure with very low variability) helped us to classify procedures according to their variability.

With regard to the use of “low” or “high” utilization rate labels, this was much easier because from the whole set of procedures that can be of interest, procedures such as hip fracture or knee replacement are among the most frequent as compared to the Spain average over all procedures, whereas the low extreme amputation and pacemaker can be considered as the low frequent as to be appropriately analyzed.
with these methods. Hence, having available a wide set of procedures to be analyzed, we selected these six as representative of cases one may find when analyzes medical practice in a setting similar to ours. We have added a comment on these in the description of the procedures selected.

Reviewer: Peter Congdon; Comment #5. DEFINE TERMS MORE. For the wide readership of BMC-HSR (including non-statisticians), I think the authors should provide rather more clarifying detail and definition for the reader and some references with regards to terms such as “parametric bootstrap” and “nonparametric sampling”.

According with the reviewer suggestion we have mentioned some details of these procedures in the section devoted to the assessment of the null hypothesis, although we have not extended that much in order to maintain the focus of the study.
The authors report that they excluded 5% of extreme standardized rates for each site. How sensitive are the results to the exclusion of these sites? Were the excluded sites different from included sites in important characteristics of interest?

For some of the statistics, the punctual estimations are diminished after the exclusion of these sites, particularly the EQ in the low rate procedures. For instance, for the extreme amputation procedure, this punctual estimation is very high (EQ=25.59) and it lows down to the value of 4.11 when excluding 5% of extreme rates for each tie, whereas for other procedures it is not possible to derive the EQ given the presence of zero cases in some areas.

In general terms, statistics can suffer a decrement of about one third or one fourth of their value when eliminating these extreme areas, but the effect that this exclusion has on the findings derived is practically negligible. For instance, the performance of the statistics, regarding both the comparison between the observed variability compared to that expected by chance and the width of the confidence intervals, remains unchanged, except for a slight decrement that can be observed in this interval amplitude after excluding the data (see Table 1). The same happens when variability under the null hypothesis is compared among procedures with different rates, as all statistics with and without the extreme values depict the “inflation” of variability in the low rate procedures (see Figure 1 of the manuscript). Hence, the effect of excluding the 5% of extreme rates in some statistics seems negligible with regard to the comparison between null and observed intervals, because the expected variability depicted is lower when excluding them both under the null hypothesis and under the observed variability.

The power of the statistics is in general slightly better when excluding extreme areas (see Figure 2 of the manuscript), particularly for the low-rate procedures, because usually the areas that have been eliminated are those corresponding to instable rates of low populated areas, which blur the underlying risk pattern. Hence, the only characteristic of interest that is shared by the excluded areas for each procedure is just the higher probability they have to present not very reliable risks or rates estimates. We would not recommend presenting results only with the data having excluded extreme rates, so that we present results both including and excluding these rates to assess the robustness of the findings. Some of these comments have been included in the result section to clarify these aspects.

Paragraph 1: The scenarios S1-S6 need to be clearly defined.
We have tried to clarify the definition of the scenarios simulated in this new version of the manuscript. The procedure is very close to that used in similar contexts by other authors, such as that proposed by Richardson et al., 200\textsuperscript{16} or that conducted in Ugarte et al, 2008.\textsuperscript{17} *In the new version the description of the scenarios is given in the second paragraph of the section devoted to the power assessment.* Apart from that, the simulation procedures under the null hypothesis have been rearranged in a box to facilitate the understanding, and the procedure can be extrapolated to the construction of the scenarios.

Reviewer: Freedom Nkululeko N Gumedze. Major comment #3. Page 17: Paragraph 3: The sentence “…, while the other three are based on a mixed model where the area-specific effect is the random effect”. These models need to be described in the Material and Methods section. This is also not clear from either Table 2 or 3.

We have added a paragraph in the analysis section explaining which are the underlying models assumed to derive the SCV and the EB statistics, and have tried to link the explanations with the baseline homogeneous Poisson model assumed under the null hypothesis, which is the bases for the $X^2$, DT and BT.

We have also mentioned the extensions of these models to accommodate other structures such as the spatial correlation that may be presented in data. Additional references have been included that adopt this models in similar settings.

Reviewer: Freedom Nkululeko N Gumedze. Major comment #4. Page 19: Paragraph 1: Could the authors explain why could not just use the R code to compute the EB estimates as they suggest in this paragraph.

We agree with the reviewer that the EB estimates can be derived using standard software (such us the command glmmPQL implemented in R) in the case of the exchangeable model, which is the one used here. Although the extended version to cope with spatial autocorrelation is not straightforward from these commands and authors usually write their own codes,\textsuperscript{18} as we mentioned in that sentence, there is no need to complicate procedures when spatial correlation is not included. *This sentence has been removed to avoid misunderstandings*, thank you for noticing this.

\textsuperscript{16} See footnote #6.
\textsuperscript{17} see footnote #7.
\textsuperscript{18} see footnote #11.
In effect, it is the estimate that is given in the Table 3. The cumulative percentage is not necessary, and although it was included to allow rapid derivation of the median and the interquartile range of the population without taking much space, it has been removed as you suggested.

Given that the underlying Poisson process for rare events (with a mean taking into account the population structure by means of the $e_i$) is usually adopted, $y_i$-Poisson($e_i r_i$), the null model would reduce the process to $y_i$-Poisson($e_i r$), where $r$ represents the homogeneous risk. The chi-squared test was derived in a much general context of association between variables and can be easily adapted to this context (see Cain and Dihe for a step-to-step derivation under different distributions, in particular for the Poisson distribution which is the version used here). The estimates obtained for a given procedure can be compared to a chi-squared distribution with J-1 degrees of freedom, and the main application is to compare the observed estimate with the 95-th quantile $X^2_{1-\alpha=0.95}$, so the higher the estimated value, the higher the distance between the observed and the hypothesis of no variability, and the higher the evidence of variability.

With regard to the Bohning statistics (1999)19, it was derived exclusively to test for Poisson extra variation in the context of disease mapping, and that is why in the formula it appears the SUR (or the original version, the Standardize Mortality Ratio, SMR). The formulae follows, under the null hypothesis, a normal distribution, and again the higher the statistics observed (compared to the quantile $z_{1-\alpha=0.95}$), the lower the probability of observing a discrepancy more than or equal than the observed under the homogeneity assumption, and so the higher the value, the higher the evidence of variability. Finally, what we called here “the Dean statistics”,20 it was derived to test for extra-Poisson variation in a general context in which other test of overdispersion against general alternatives such as beta-binomial models were provided. All these three statistics, therefore, were constructed under the null model poisson($e_i$) assumed here, and the performance of them when the null hypothesis is true had already been studied.

The use of these statistics to discriminate among different scenarios of variability had not been presented before, and only the power of the $X^2$ statistic has been evaluated previously. Perhaps its use on their ability to discriminate among

---

19 see footnote #7.


different levels of variability can be controversial as they were not derived to do so, but here we only present the results and assess the performance. Having a look on the $X^2$, Bohning and Dean test performance in the Figure 3 of the manuscript, it is clear that they are very dependent on the procedure rate (and equivalently on the population size), because all the three statistics have the difference of observed minus expected cases per area in the numerator, and the higher these difference (which is more probable in highly populated areas with high procedure rates) the higher the summands on the numerator. With this study we point out that although Diehr had recommended the use of the $X^2$ statistic because of the better performance, and others have maintained this recommendation (see To et al, 2003)\textsuperscript{22}, depending on the focus of each study caution is needed not only in this statistic but also in the Bohning and Dean statistics.

The analysis section has been extended to include these aspects to understand better the origin of these statistics, and also in the discussion section we have added (in the 4th paragraph) some comments about the interpretation of their performance.

\begin{footnotesize}
\begin{center}
\textit{Reviewer: Freedom Nkhululeko N Gumedze. Major comment #6. Table 2: Describe the calculation of SUR}
\end{center}
\end{footnotesize}

Linked with the previous explanation, if we assume a Poisson process to model the number of admission per county $y_i$, $y_i$~Poisson($e_i r_i$), and let $r_i$ be independent parameters to be estimated for each area, then the standardize utility ratio (SUR), homologous to the well-known standardize mortality ratio (SMR) in this setting, would be the maximum likelihood estimator of $r_i$, which is the quotient $y_i/e_i$ for all $i$ in 1,..., I. Here:

$$e_i = \sum_{j,k} n_{ijk} \left( D_{jk} / N_{jk} \right)$$

where $n_{ijk}$ is the population in area $i$, age group $j$ and sex stratum $k$, and $D_{jk}/N_{jk}$ is the age-sex specific rate for the whole region under study. \textit{The SUR derivation has been included in the paragraph included on the analysis section, and in Table 2 we have added that it is the quotient $y_i/e_i$}.

\begin{footnotesize}
\begin{center}
\textit{Reviewer: Freedom Nkhululeko N Gumedze. Major comment #7. Table 2 and 3: Describe the calculation of EB estimates}
\end{center}
\end{footnotesize}

Given that the EB estimates used here does not have a closed form, we did not have space enough as to write it in the Tables 2 and 3. However, the derivation of this estimate requires an iterative process that we did not mentioned in the

\textsuperscript{22} To, T Williams JL, Wu K, Theriault ME, Goel V. Comparison of methods to identify outliers observed in health services small area variation studies. Stat Methods Med Res. 2003;12(6):531-46.
previous version of the manuscript. Some points of the process have been added in the analysis section of this new version to clarify this aspect.

Reviewer: Freedom Nkhululeko N Gumedze. Minor Essential Revisions #1. Page 10: Line 10: The formula for the mean of the area specific proportion seem to be written as an exponent; is this correct or is this a typographical error?

It was a typographical effect after using the double-space for the format, but in this new version of the manuscript, we have reorganized the explanation of the bootstrapping procedure to make it clear, and this part has been included in Box 4.


In mathematics, the \#\{\} symbol represents the number of times \{\} occurs. Hence, \#\{EQr\geq EQobs\} denotes the number of simulated EQr greater than or equal to the observed statistic. Here R=2000, so it counts how many of the 2000 EQ values obtained from the 2000 samples simulated are greater than or equal to the observed EQ. Given the new format of this section in this version of the manuscript, this explanation appears inside the box corresponding to Step 4.

Reviewer: Freedom Nkhululeko N Gumedze. Minor Essential Revisions #3. Page 16: Discussion: First sentence: The word random should be replaced by the word chance.

Following the reviewer's suggestion, we have done the replacement.

Reviewer: Freedom Nkhululeko N Gumedze. Minor Essential Revisions #4. Page 16: Discussion: The phrase “These farness between intervals is...” should be replaced by the phrase “These confidence interval widths are...”

Following the reviewer's suggestion, the sentence has been replaced by another that tries to better explain what we were trying to highlight, which is not only the widths of the intervals, but also the distance between the upper limit under the null hypothesis and the lower limit of the confident interval derived form the observed data.
Following the reviewer’s suggestion, we have changed it in the new version.

We appreciate the comment about our paper. No changes were need in relation to this comment.

We have added this information in the new version.

The analysis of small area variations in hospitalization rates usually considers large areas because the interest here is to know if two patients belonging to different hospital referral areas with the same medical condition could have a different treatment, that is, if there exists “medical practice style” differences between areas for a given procedure. Hence, in contrast to other settings such as disease mapping for which very small areas are considered to depict whether
environmental or socio-demographic factors at a small geographic level may have an influence on mortality or morbidity, here larger areas are considered, and therefore more stable estimates are obtained for the rates and the risks per area. Areas with zero counts are rarely found, and areas with small number of counts (less than 5) only appear when analyzing extremely rare procedures. The simulation study was repeated using areas 10-times less populated than the real areas as suggested by other reviewer, and the main results derived in the manuscript still hold, although the power of the statistics was in general lower in this case. With regard to other conditions that cover other casuistic the analyst may meet such as spatial autocorrelation, we used some models as those reviewed in Ugarte et al (2006) (see responses to first reviewer) but the contribution of them to this particular topic of quantifying the variability when having such stable rates did not seem to be enough as to complicate the paper. However, they will be taken into account in other possible applications for the analysis of other spatial structures with sparser population.

Reviewer: Lawrence Ndekeneni Kazembe. Minor comments #2. page 12, line 7. The choice of RR 1.2 and 1.6 (line 9), needs some justification.

The choice of RR =1.2 and 1.6 was based on other published papers with similar simulation studies. That carried out by Diehr et al. (1992) in the context of medical practice variation showed that in their particular setting with 39 areas, using an increased risk of RR=1.6 for 2, 5 and 10 areas (about 5%, 13% and 25% of the areas), the 80% of the power was achieved for example for the SCV statistic. Given that we used 10, 20 and 40 out of the 147 with elevated risk (about 5, 15 and 25% of the total respectively), this increased risk of 1.6 could be used as reference. On the other hand, in other simulated scenarios of disease-mapping studies such as that presented in Richardson et al (2004) or Ugarte et al (2008), relative risks of RR=1.5, 2 and 3 are used.

The reason for using lower elevated risk areas in our case than these is that the real scenarios of interest in their settings have less populated areas with a low number of expected cases per area, and therefore higher risk is need to be detected with a given power in their case. Additionally, given that our highest variable scenario S6 attained a power equal to 1 for all statistics, it was not considered necessary to simulate others. In the case of a setting with less populated areas, as that used in response to the requirement of one of the referees,

25 see foodnote #6.
26 see foodnote #7.
three additional scenarios have been required (RR=3 for 10, 20 and 40 areas) to attain power equal to 1 for most statistics. A sentence stating that different degrees of induced variability (different values for RR) could have led to different results has been added in the new version.