Author’s response to reviews

Title: Predictive risk modelling in the Spanish population: a cross-sectional study

Authors:

Juan F Orueta (jon.orueta@osakidetza.net)
Roberto Nuño-Solinis (nuno@bioef.org)
Maider Mateos (maidermateos@gmail.com)
Itziar Vergara (itziar.verbaramitxeltorena@osakidetza.net)
Gonzalo Grandes (gonzalo.grandesodriozola@osakidetza.net)
Santiago Esnaola (sesnaola@ej-gv.es)

Version: 4 Date: 14 May 2013

Author’s response to reviews: see over
Reviewer: A C-L

Minor revisions:

1) The authors analyse the explanatory power of socio-economic variables extracted from 2001 census data and they conclude that inclusion of these variables led to only marginal improvements. Still, the critical impact of socio-economic status both on multimorbidity and health services utilization has been widely described. I would recommend the authors to discuss the limitations of the variable that they employed to measure deprivation in terms of the time elapsed since extraction as well as its ecological nature.

The deprivation index was categorised into quintiles. Although the socioeconomic variables on which the deprivation index is based could have changed since the census was taken, it is very plausible that the relative position of each census tract has not changed in that period of time.

The ecological character of the deprivation index could underestimate the contribution of individual socioeconomic characteristics. Socioeconomic indicators of a geographic area refer to contextual factors that determine the health of its inhabitants independently of individual factors. Furthermore, when constructed from individual characteristics, area indicators can capture constructs that refer to the area as well as to the individual (Diez Roux AV.: The study of group-level factors in epidemiology: rethinking variables, study designs, and analytical approaches. Epidemiol Rev 2004, 26:104-11.). Hence, in the absence of individual socioeconomic indicators, we cannot distinguish the compositional effect (of individual) from the contextual effect (of the region).

Therefore, in the discussion of the manuscript, we have added a sentence stating that:

"Finally, a third limitation refers to the social variable used (deprivation index) which, given its ecological character, may underestimate the contribution of individual socioeconomic characteristics."

2) In two recent studies, it was stated that the count of chronic diseases (Huntley et al, Ann Fam Med 2012) or prescriptions (Brilleman et al, Fam Pract 2012) showed a predictive capacity almost as high as those derived from proprietary risk adjustment systems which entail important costs to end-users. It would be appreciated if the authors could discuss this stressing the added value of their findings.

Although some simple measures have been used as a disease burden measure and, in some cases, to predict mortality, number of visits or other variables, we are unaware of their use in predicting total healthcare cost. Therefore, given our limited space, we would prefer not to add this issue in the discussion.

3) Within the limitations section, the authors mention that the over-reporting of diagnoses and/or prescriptions should not be a problem as records of primary care tend to suffer from a lack of sensitivity but not specificity. However, it has been suggested that only active diagnostic episodes should be considered for
risk-adjustment purposes in order to avoid an artificial amplification of patients’ morbidity burden (Sicras et al, Aten Primaria 2006). The authors may want to add this to the limitations.

As the reviewer points out, it is very difficult to determine whether a problem diagnosed in a patient continues to be active in a subsequent period using administrative databases.

In our case, we have only included the diagnoses for which patients contacted the Basque Health Service, according to the criteria of the professional responsible for their care, that is, because they are registered in the hospital discharge report or entered into primary care medical records.

The diagnoses accepted in primary care correspond only to those episodes in which a doctor included one or more annotations during the observation period (between 01/09/2007 and 31/08/2008). Therefore, these health problems were inevitably active. Even in the case of illnesses that last throughout the patient's life, episodes from previous years (active or not) were not taken into account unless an annotation had been made by the physician during the study period.

To clarify this in the manuscript, the following sentence has been added in the methods section:

“In order to avoid the possible inclusion of long-term diseases that were not currently active, a diagnosis of chronic disease recorded in the PC-EMR was not accepted unless, and according to the annotations of the doctors responsible, it was considered reason for encounter during the study period”.

4) Another possible limitation could be related to the fact that no split half method was applied, which involves dividing the total sample in two and correlating the results, as an acknowledged way of assessing the reliability of a test (Meenan et al, Med Care 2003; Forrest et al, Am J Manag Care, 2009). The authors may want to comment on this.

At the end of page 10, we explained that:

"To avoid overfitting problems, a 5-fold cross validation was carried out."

In the new version of the manuscript this topic has been extended, indicating that:

"To avoid overfitting problems and to confirm that the results do not depend on the sample, a fivefold cross validation was carried out. Thus, the sample was split into five random subsamples and model fitting was performed five times, considering four of these subsamples as the training set, and the remaining one as the test set, each time. The statistics employed to evaluate and compare the performance of the different models were derived from this cross validation."

5) Many studies have focused on the predictive capacity of risk-adjustment systems regarding health resource utilization, both in the US and Europe. However, the applicability of such tools for the actual improvement of population health outcomes in public health systems offering universal coverage and with a solid primary health care network has been less frequently evaluated. It would be of high interest if the authors could further elaborate on this issue within the discussion section.
In the discussion we explain that, in our setting, stratification may allow the identification of patients who will present high healthcare needs in the future and whose deterioration can be avoided by implementing proactive activities. While this point might be worth further development, we would prefer not to prolong the discussion section, thus maintaining the length of the manuscript at around 3,500 words.

**Reviewer's report: H-Y C**

**Reviewer's report:**

**Introduction**

The authors might want to provide more information of the healthcare system in Basque, such as its enrollees, qualification, coverage, financing and payment system, so that the readers can have a better sense of the background and the implications of this study.

In the methods section we originally stated that "Osakidetza is a public healthcare organisation that provides universal coverage, funded through regional general taxation". We have now added a new sentence, saying "All health professionals in Osakidetza are salaried. Each citizen is included on the list of one primary-care physician, who acts as the gatekeeper to other levels of care."

We would prefer to maintain the length of the manuscript at around 3,500 words. Although this description of Osakidetza is very brief, we think that it will allow readers to obtain a basic idea of its characteristics.

**Method**

It seems the enrollment was only required in the first year (at least 6 month); how about the second year?

Although such information was included in the results section, we recognise that the wording could be confusing. Therefore, this explanation has been modified and moved to the methods section, where it is now described that those citizens who died during the second year were included, but those who withdrew from Osakidetza for other reasons were not.

How did the authors take into account that some study subjects did not have a full second-year enrollment to construct costs? The authors might want to annualise costs for these enrollees.

Since such patients are those who died in that year and the last months of life are usually those of the highest cost, we have not annualised costs.

Why did the authors only include people assigned to a doctor? Is being assigned to a doctor required for enrollment in the healthcare system?

As indicated above, all persons enrolled in Osakidetza are assigned a primary care physician.

How did the authors obtain standard costs?
We have added an explanation indicating that the standardised cost corresponds to the "average cost of each service provided to a patient treated in Osakidetza, according to calculations made by the aforementioned organisation."

Costs associated with many types of services were excluded. Did the authors have the access to the record of payment to each claims from the Basque health insurance? Can the authors estimate what the percentage of total cost was not included due to such exclusion?

We have added a sentence stating that "it was estimated that the total cost of the excluded services corresponds to 28.2% of the overall budget."

The authors should give more detailed information on the main independent variables (ACG-PM, CRG, & DCG-HCC) in the manuscript instead of the appendix, especially for the ACG-PM model since it is just a subset of the ACG system and many readers might not be aware of the difference between the ACG-PM and other component of the ACG system (for example, on page 10 among the 180 markers for diagnosis included in the ACG-PM, 34 are actually from the mutually exclusive categories called ACGs).

We agree that it is necessary to provide readers with a description of the case-mix systems studied. In our opinion, to include it in the methods section will prolong excessively its length and we would prefer to submit it as an appendix. On the other hand, the text of the manuscript has been modified, now stating that only 34 ACG categories are used in the predictive model.

Please confirm that the C-statistic was also derived from the cross-validation set.

In the new version of the manuscript this topic has been extended, indicating that:

“To avoid overfitting problems and to confirm that the results do not depend on the sample, a fivefold cross validation was carried out. Thus, the sample was split into five random subsamples and model fitting was performed five times, considering four of these subsamples as the training set, and the remaining one as the test set, each time. The statistics employed to evaluate and compare the performance of the different models were derived from this cross validation.”

Results
The authors might want to include # of independent variables in the table. Readers cannot make fair comparison without this information because the model performance usually increases as more variables are included.

We agree with the reviewer: readers need to know the number of variables included in each model. However, it is very difficult to make up a table containing all the required cells. Once again, we believe that such information is best presented as Appendix-1

For table 1 the authors should include the main outcomes (costs, % of people above top 95/99 percentile).
We believe that information regarding costs and contact numbers of patients is very important. Since our description is very extensive, we have decided to present it in detail as an appendix.

On page 13 table5 and 6 should be table 4 and 5.

Such mistake has been corrected

Discussion
The authors might want to discuss the impact of including people with incomplete enrollment in the second year and the exclusion of certain types of costs on the model performance.

As we have included a sentence in the methods section explaining that patients who did not complete the second year are only those who died in that year, we believe that it is not necessary to repeat this in the discussion.

On page 18 is it 260 ACG group or ACG-PM group?

We have tried to indicate that the case-mix systems used in our study categorised the diagnoses, which can describe the morbidity of a population and identify patients with health problems. In the case of the ACG system, diagnoses are classified into 260 EDCs (Expanded Diagnoses Clusters). While it is true that not all EDCs are included in the predictive model (this is explained in Appendix-1), we find it useful to keep this figure to help readers understand the level of granularity that allows this case-mix for the description of pathologies. On the other hand, we prefer to keep the acronym ACG-PM, since it is employed in other parts of the manuscript.

On page 19, the word/sentence for citation 30 is missing.

This mistake has been corrected

On page 30, the statement that “…in the total adult population…” may not be correct since those aged 14-17 were also included.

The term "adult population" has been changed to "population over 14 years of age"