Reviewer's report

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

Version: 3 Date: 2 February 2013

Reviewer: Hsien-Yen Chang

Reviewer's report:

The authors compared the capability of three different claims-based models (ACG-PM, DCG-HCC & CRG) in explaining prospective costs and identifying future high-cost users, when it was applied to data collected from the healthcare system in Basque Country. The results showed that the performance of these three models was comparable, and consistent with what was observed in other countries.

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.

Method

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

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 annualize costs for these enrollees.

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

How did the authors obtain standard costs?

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?

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).

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

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.

For table 1 the authors should include the main outcomes (costs, % of people above top 95/99 percentile).

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

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.

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

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

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

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

**Statistical review:** No, the manuscript does not need to be seen by a statistician.

**Declaration of competing interests:**

The reviewer is currently hired as a assistant scientist at John Hopkins University with part of the funding coming from the ACG team.