Reviewer's report

Title: Information for decision making from imperfect national data: tracking major changes in health care use in Kenya using geostatistics

Version: 1 Date: 1 August 2007

Reviewer: Basile Chaix

Reviewer's report:

Review of manuscript

This paper is of interest, and methodologically sound. Readers will appreciate the effort of the authors to explain in a didactical way their complicated model. As they provide the appropriate references for the technical details, I believe it is enough for the present work.

A central question I have is that it is not clear what the author's methodology in fact adds in the present application. Certainly, more accurate estimates are obtained than when using crude means; but it is not clear here whether different conclusions would have been made on the temporal trends if crude means have been used. Is there a difference in terms of the conclusions provided to the policymakers?

An important methodological comment I have is related to the opposition made by the authors between the geostatistical paradigm and the multiple imputation approach. Would not the best approach be to combine these two perspectives? Obviously, the facilities that are underreporting information are different than those that correctly report, and the propensity to report is certainly related to observable characteristics of the facilities, that may also be correlated with the number of diagnoses and with the type of diagnoses. Would not that be possible to make a prediction based on both the characteristics of the facility and its space-time position? For example, the facility type (dispensaries, health centres, hospitals) should predict the number and type of diagnoses. If not taking into account the characteristics of facilities, the interpolation procedure may be based on the selected and specific facilities that are reporting.

Was the size of the facility directly taken into account to predict the number of cases in case of missing data? I guess that predicting the value for a facility a given month based on the values for this facility in other months is a way to take into account the facility size. But is it sufficient?

The authors mainly focused on temporal variations. I expect that correctly predicting spatial variations would be much more difficult with this method.

Of course there is temporal autocorrelation in the number of cases for a given facility (which is probably explained by the facility size). But was there some spatial autocorrelation? On what spatial scale (in kms)? What was the range of
correlation? In case of absence of spatial autocorrelation, borrowing strength from the neighbours would be inefficient.

Could not there be a bias related to the fact of using only the facilities that were already operational in 1996, not the others? Changes in the number of patients in these facilities may be due to the opening of other facilities around them, which would diminish their patient load.

It is assumed that once a facility is reporting data a given month, it is reporting all its cases exhaustively. There may also be temporal variations in the percentage of cases reported in each hospital, and these temporal trends in the percentage of cases reported when reporting may be common to various facilities (for reasons related to the HMIS system), affecting the overall trends.

It would be desirable to have uncertainty bands around the predictions. I guess the model could provide 95% credible intervals.

**Which journal?**: Appropriate or potentially appropriate for BMC Medicine: an article of importance in its field

**What next?**: Accept for publication in BMC Medicine after minor essential revisions

**Quality of written English**: Acceptable

**Statistical review**: No

**Declaration of competing interests**: I declare that I have no competing interests.