Reviewer's report

Title: Using Relative Survival Measures for Cross-Sectional and Longitudinal Benchmarks of Countries, States, and Districts - The BenchRelSurv- and BenchRelSurvPlot-Macros

Version: 1 Date: 11 October 2012

Reviewer: J. David Beatty

Reviewer's report:

Overall this is an excellent manuscript that extends work published several years ago. The authors have made excellent revisions such as utilization of survival rather than mortality and creation of indices that generate higher scores for better outcomes. They have provided the software for doing these analyses. I have not personally assessed the software. My recommendations are as follows:

1. The second paragraph of the Introduction refers primarily to the work which was done by our group. There are several areas that I feel do not represent the thought process and work in an adequate and balanced fashion. I would like to make a specific recommendation for changes and have shown them in the paragraph below by ‘tracking’ the changes to simplify the editors and authors evaluation.

First approaches on how the number of patients per disease phase / stage can be linked to the corresponding chance of survival are based on Beatty et al. [6,7]. They developed a series of benchmark-algorithms that address different aspects of health care. In preliminary work, Beatty and colleagues suggested described a screening index based upon the the sum of the products of the stage number (0-4) and the number of cases at that stage divided by the total number of cases. Using the stage number in the calculation was considered arbitrary and it was replaced by the national 5-year mortality for that stage termed the case-mix index for the institution or region. This was then standardized by comparison to the national case-mix index and termed the standardized case-mix index. They also described a standardized work-up index to address the issue of ‘upstaging’ of cases, the so-called Will Rogers-Phenomenon, and a standardized treatment index to evaluate outcome using institutional or regional mortality (stage and overall) compared to national mortalities. The product of the 3 as well as standardized indices (case-mix-, work-up- and treatment)-indices and was a summarizing index called overall performance evaluation (OPE). They used these indices and evaluations to compare different regions over the same time interval (cross-sectional analysis perspective) and the same institution across different time intervals (longitudinal analysis perspective). However, the screening-index was not standardized, reflected a negative association with the observation period, and, thus, was not included in the OPE. In addition, the OPE had a logical weakness, because the standardized case-mix index was present in the denominator as well as the numerator and cancelled itself out. Finally, the
approach was based on disease-specific cause of death statistics which required
the modeling of competing risks according to the method of Gooley [8]. This
method reduces case numbers substantially which is of special importance for
small geographic units that already have small case numbers as it is. This
outcome-indicator also does not include background-mortality.

2. In the Standardized Indices paragraph, line 4 the word “defined” is repeated.

3. The formula for the eSSI as defined would have zero in the denominator for
both the observed and expected calculation. From the description provided, I
believe that this index should be defined as:

\[ eSSI = \frac{O}{E^*} = \frac{\sum (N_i \times i/N)O}{\sum (N_i \times i/N)E^*} \]

If correct, then the program should also be checked to be certain that it is correct.

However, I think that the eSSI as intended provides the only major problem with
the proposed indices. The eSSI provides more weight to the advanced stages of
disease (stages 3 and 4) because the number of cases in the stage is multiplied
by the stage number. Thus, as the proportion of observed early disease cases
(stages 0, 1 and 2) increases, an early surrogate indicator of improved screening,
the eSSI number will decrease. Further, if the observed screening is better than
the reference, the eSSI will be less than 1. The negative association of eSSI with
screening efficacy could be avoided by using the definition \( eSSI = E^*/O \). By using
survival rather than mortality, the authors have achieved a positive association
for the other indices and I agree with them that associating higher scores with
higher performance is generally easier to understand and embrace.

I note that the first paragraph of Strengths refers to the eSSI as “expressed as a
reciprocal”. If this is correct, then this second problem with eSSI is solved by
defining eSSI as follows:

\[ eSSI = \frac{E^*}{O} = \frac{\sum (N_i \times i/N)E^*}{\sum (N_i \times i/N)O} \]

4. In the Conceptual pitfalls for longitudinal analyses section, the major emphasis
is on the ‘stage-migration’ and the Will Rogers Phenomenon. This is also a pitfall
of cross-sectional analyses. Unfortunately, the authors do not emphasize the
strength of the approach that they have taken which addresses this pitfall. In
addition, they have not presented the other basic principles being utilized in
assessment of the programs. I would suggest that the discussion or conclusions
should include several sentences such as:

In the approach being used in this report, stage-migration is not a pitfall, but
rather is an asset because it is being measured. The SWI is based upon the
premise that the more critical the work-up, the more upstaging will occur and the
better the survival at each stage. However, unless the stage migration alters the
treatments administered, it will not impact the overall survival, only the survival at
each individual stage. The eSSI and the SCI are based on the premise that the
more effective the screening, the more the distribution of cases will be skewed
toward the earlier stages of disease, again an asset. The STI is based upon the
premise that the better the overall survival corrected for the case-mix, the more effective the treatment being administered. The utilization of each of these indices as a benchmark provides a means of identifying specific areas of program strength and weakness. The combination of these indices to create ROPI provides a benchmark for assessment of overall program quality.