Author's response to reviews

Title: Using machine learning algorithms to guide rehabilitation planning for home care clients

Authors:

Mu Zhu (m3zhu@uwaterloo.ca)
Zhanyang Zhang (z29zhang@uwaterloo.ca)
John P Hirdes (hirdes@uwaterloo.ca)
Paul Stolee (stolee@uwaterloo.ca)

Version: 2 Date: 30 October 2007

Author's response to reviews: see over
Attention: Iratxe Puebla, Senior Assistant Editor, BMC-series journals

Dear Ms. Puebla,

Re: MS 1985425591157745: Using machine learning algorithms to guide rehabilitation planning for home care clients.

We have completed the requested revisions to the subject manuscript, and are re-submitting the paper for your consideration. We found the reviewers’ comments to be very helpful, and feel the paper is much improved as a result. The following is a point-by-point response to the suggestions made by the reviewers (our responses are in italicized text):

Reviewer: Enzo Grossi

Major Compulsory Revisions

Data set description.

The focus of the paper are the variables collected in the interRAI/minimum data set. The variables at the end belong to human subjects of whom no mention is made. The paper is based on huge data set concerning more than 24000 patients. Some descriptive statistics of this population according to the 19 variables collected could help the reader in putting the results in the appropriate clinical context.

We agree that additional descriptive information on the subjects should have been included. We have included information on mean age, % female, and % with Alzheimer disease or other dementia. We have also added three columns of descriptive statistics to Table 1:

- **% = 1:** overall percent of patients in our dataset whose recoded covariate is equal to 1, i.e., for whom this covariate was present.
- **Chi-square:** Chi-square statistic for testing the correlation between each covariate and the response.
- **Corr:** Pearson correlation between each covariate and the response.

We understand the reviewer’s suggestion to add the correlation index values. We were concerned however, that strictly speaking, the chi-square statistic may be more appropriate. In the end, we included both, with the addition of the following explanatory note on page 6:
“Since the response is binary and covariates here are also recoded as binary, the Pearson correlation is not exactly the right correlation coefficient to use. The chi-square statistic is more commonly used. However, the usual Pearson correlation coefficients are still included here for the following reason. Since we have a sample size of about 25,000, the chi-square statistics are all very large, reflecting the well-known caveat of classical hypothesis testing that one can reject any null hypothesis with a large enough sample size. The absolute magnitude of these chi-square statistics should not be interpreted in the usual way, but their relative magnitude is still meaningful.”

Methods and Results.

Since the contribution offered by the article is mainly methodological and is focusing on the strength of specific statistical approaches, the description of the validation protocol is a fundamental procedure to verify the models’ ability to generalize the results reached in the real world.

It is not clear from the description if the confusion matrix data described are just the result of interpolation or describe validation data. In this case we need to know: which protocol has been adopted like formal cross validation training-testing protocol (like so-called 5x2 cross-validation training-testing protocol described by Dietterich in 1998), resampling methods such as bootstrap, or K-fold, in which we subdivide the data into K roughly equal-sized parts, then repeat the modeling process K times, leaving one section out each time for validation purposes.

We have substituted the Data Analysis section related to Study 1 in the previous manuscript with two sections. The first is a brief description of the Support Vector Machine. The second is a section on “Performance Evaluation” in which we have provided the additional detail requested by the reviewer.

In addition the target variable should be precisely defined (mobility in bed?)

Additional description has been added to Table I where the definition of the covariate was unclear.

In the tables 2,4,5,7 the overall accuracy rate should be added to false+ and false- rates to facilitate the comparison among different techniques.

Overall error rate has been added to Tables II, IV, V and VII.

Minor Essential Revisions

Title page: the title could be more direct as for example: role of machine learning algorithms in optimization of rehabilitation protocols. The e-mail addresses for all authors’ must be included in the title page.
We appreciate the suggestion, but upon reflection we would prefer to keep our original title. We feel the proposed alternative might overstate our findings (specifically the reference to optimization). E-mail addresses for all authors are now included.

Authors’ contribution: I suggest to add the following sentence: “All authors read and approved the final manuscript.”

This sentence has been added.

Figure legends: The legends should be included in the main text file immediately following the references, rather than being a part of the figure file.

This has been done.

Discretionary Revisions

Style and language. The style of the article is rather narrative. I recommend to be more short and snappy.

We agree with the recommendation, but given the requests for additional detail, were unable to shorten the paper.

Thank you for your continued consideration of our manuscript. We again look forward to the results of your review.

Yours sincerely,

Paul Stolee, Ph.D.
Associate Professor and Graham Trust Research Chair in Health Informatics
University of Waterloo, Waterloo, Canada
Reviewer: Kwokleung Chang

Major Compulsory Revisions

In general, the manuscript is well written and organized. Here are a few major comments:

1) In the abstract, it mentions that “a client is defined as having rehabilitation potential … i) improvement in ADL functioning, or ii) discharge home.”

In the “Data” paragraph on page 5, it mentions that “…true rehabilitation potential (y) can be reliably assessed from linked health service…”

It should be more clearly explained how the “true rehabilitation potential” is defined and why it is defined that way.

*Thank you for pointing out this discrepancy between the abstract and the body of the manuscript. We have more clearly explained our definition of “true rehabilitation potential”.*

2) It should be explained more how cross-validation (CV) for parameter tuning is done. Were all eight CCAC datasets used in the CV? Were only some samples from each dataset involved in the CV? Were these samples reused in the performance evaluation of the SVM?

*As noted above in relation to Dr. Grossi’s comment about the validation protocol, we have substituted the Data Analysis section related to Study 1 in the previous manuscript with two sections. The first is a brief description of the Support Vector Machine. The second is a section on “Performance Evaluation” in which we have provided the additional detail requested by the reviewer.*

3) On page 11, why is the ratio profile scores defined as max(r>r>1)/max(r<r<1)? Min(r>r>1)/max(r<r<1) seems to be a more reasonable score. Or how about median(r>r>1)/median(r<r<1)?

*We understand the reviewer’s suggestion, but feel that the ratio profile score we chose is defensible. We have added a discussion of this on page 13. This addition may be more suitable as a footnote, but we will leave that to the editor’s discretion.*

Minor Essential Revisions

4) In tables 2, 3, 4, 5 and 7, the results are presented with too many number of decimal places. They give the impression that the results are accurate down to 1/10000.

*We have revised the figures in the tables to include only two decimal places.*