Reviewer's report

Title: Use of generalised additive models to categorise continuous variables in clinical prediction

Version: 2 Date: 12 January 2013

Reviewer: Lin Yee Hin

Reviewer's report:

\documentclass[12pt,a4paper]{article}
\RequirePackage{amsmath,amssymb,graphicx,txfonts,mathrsfs}
\begin{document}
\begin{center}
{\bf Reviewer's report for ``Use of generalised additive models to categorise continuous variables in clinical prediction''}
\end{center}
\section{Major Compulsory Revisions}
\begin{enumerate}
\item Methods section, Theoretical methods subsection, Generalized Additive Models sub-subsection, paragraph 1, line 1:

Rephrase ``A GAM[11] is a Generalized Linear Model in which part of the linear predictor is specified in terms of a sum of smooth functions of the covariates."

GAM is an extension of the Generalized Linear Model where the modeling of the mean functions relaxes the assumption of linearity, albeit additivity of the mean function pertaining to the covariates are assumed. Whilst the mean functions of some covariates may be assumed to be linear, the non-linear mean functions are modeled using spline-based methods.

\item Methods section, Theoretical methods subsection, Generalized Additive Models sub-subsection, paragraph 2, line 8:

Rephrase the sentence ``Smoothing splines [13] keep the dimension of the basis fixed ..."

The dimension of the basis/knots used in smoothing spline is fixed conditional on the unique values of the covariate in question. It is not ``kept" at a size ``a little larger than is believed could reasonably be necessary" as one cannot adjust the
knot sequence used in smoothing spline.


Methods section, Theoretical methods subsection, Generalized Additive Models sub-subsection, paragraph 2, line 12:

Amend the sentence "Knot optimisation can be avoided ..".

This is not true. Granted, a rule-of-thumb knot selection strategy has been suggested by Ruppert (2002), the reference quoted in the immediately preceeding point. Without knot optimisation strategy, loss of efficiency in terms of mean average squared error is a price to pay unless some knowledge is available about the covariate in question to facilitate selection of a suitable knot sequence.

Methods section, Theoretical methods subsection, Categorisation methodology sub-subsection, paragraph 5:

It will be helpful if the authors explain how they would handle the situation when $(x_{01_{inf}}, x_{01_{sup}})$ and $(x_{01_{inf}}, x_{01_{sup}})$ overlap.

Methods section, Implementation Methods subsection, Application to the IRYSS-COPD Study database sub-subsection, paragraph 3, line 1:

The phrase "... the GLM selected was the logistic .." need to be rectified. It is GAM that is being used, not GLM. It is the link function used that corresponds to the logistic distribution due to the dichotomy of the response variable and this should be made clear instead.

Methods section, Theoretical methods subsection, Categorisation methodology sub-subsection, paragraph 2:

It will be useful to point out to the potential readers that $f(X)$ is the centered mean function where the centering coefficient is $\alpha_0$ and that the non-influence category constructed around $f(X)=0$ refers to the average value of the covariate.

The sentence "The line $f(X) = 0$ represents the point at which the outcome does not depend on X." may cause some confusion in this context and should be clarified. When $f(X) = 0$, $g(\mu) = \alpha_0$ and there will be no need to model the covariate $X$.

In addition, the terms "non-influence", "low-influence" and "high-influence" need to be explained explicitly. If my understanding is correct,
``non-influence'' refers to the region circa $g^{-1}(\alpha_0)$, 
``low-influence'' refers to the region circa $g^{-1}(\alpha_0 - c)$, \forall c > 0$, while
``high-influence'' refers to the region circa $g^{-1}(\alpha_0 + c)$.

Since the categorisation strategy maps a continuous variable into an ordinal variable, the use of the term ``non-influence'' runs the risk of giving the wrong impression that the ``non-influence'' region is associated with a lower mean value compared to the ``low-influence'' region while in fact it is the other way round. Since $\alpha_0$ is the centering coefficient, the ``non-influence'' is more like ``average risk''.

If the authors concur with this view, the use of the term ``non-influence'' throughout the entire manuscript need to be replaced by an alternative term.

\item Results section, Validation subsection, paragraph 1:

Pertaining to the assessment of model fit using AIC and prediction power using AUC of ROC, it will be informative to make explicit the following details:
\begin{enumerate}
\item Is the prediction power evaluation using AUC of ROC an in-sample assessment or out-of-sample assessment? It is known that in-sample assessment runs the risk of over-estimating the prediction power of the CPR.
\item If an out-of-sample assessment is carried out, what are the sample sizes and how are they partitioned into model development sample and model validation sample?
\item Is the AIC reported based on the model development sample alone?
\end{enumerate}

\item Discussion section paragraph 2:

The suggestion circa the sentence ``Clinical assumptions were made to recover missing data but this could only be done in the case of categorised variables ("high-risk", "non-risk" or "low-risk" patients).'' that missing value can be recovered by categorisation is an inaccurate description. Categorisation is an information reduction procedure, such as mapping a discrete ensemble of observed data that is originally recorded as a continuous variable to an ordinal variable as in the present context. It cannot ``recover" missing data.

Instead, categorisation approach is particularly advantageous because clinical parameters such as RR may have been recorded by the clinicians in some instances as a ordinal variable in the likes of eupneic/taquibneic while in other instances as a continuous variable in the likes of the actual numerical value of the respiratory rate per minute.

The adoption of categorisation of these parameters that are partially available as
continuous variable and partially available available as ordinal variable facilitate reconciliation of information. The sacrifice of some information in the subset of data recorded as continuous variable to avoid exclusion of the subset of data recorded as ordinal variable is a worthwhile trade-off. In addition, the clinical parameter is question will be categorised to construct CPR anyway.

Would the preceding paragraph be an accurate description of the message the authors intend to convey? If so, I suggest removing the claim of recovering missing values.

\begin{enumerate}
\item Discussion section paragraph 2:

The authors stated that ``When developing the proposed methodology, we considered the method suggested by Hin et. al. (1999) [8] as the first approach to our designated objective: their proposal consists of dichotomising the predictor variable by using GAMs, taking a value for which there is no risk as the cut point.''

That is not an accurate depiction of the recommendation in Hin et. al. (1999) [8]. Referring to the said paper on page 1103 paragraph 4 from above in \textit{Statistics in Medicine} issue 18 that contain the said paper, the first sentence in that paragraph states that ``... to centralize the fitted function to zero.'' while the fourth sentence states that ``...based on the fact that the average of the fitted value is zero.'' This effectively means that \( f(x) = 0 \) is taken as the \textit{average} in the sense that \( f(x) < 0 \) reduce the logit value below the average while \( f(x) > 0 \) increases the logit value beyond the average.

Therefore the recommendation in Hin et. al., viewed from the authors' perspective, is to use a threshold that corresponds to ``average risk'' instead of ``no risk''. In the special case of logistic regression, no risk corresponds to a logit value of \( -\infty \).

\end{enumerate}

\section{Minor Essential Revisions}

\begin{enumerate}
\item Background section, last paragraph, line 4:

Replace ``The third section ...'' with ``The Results section ...''.

\end{enumerate}

\section{Discretionary Revisions}
\begin{enumerate}
\item Methods section, Theoretical methods subsection, paragraph 1, line 2: Suggest removing "... , when the response variable is non-normal, ..." as the assumption of distribution family Y belongs to has been explained later on in the same sentence.
\end{enumerate}

\end{document}

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

Quality of written English: Needs some language corrections before being published

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

Declaration of competing interests:

I declare that I have no competing interests.