Commentary: Reporting standards are needed for evaluations of risk reclassification

Margaret S Pepe1,2* and Holly Janes1

1Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA and 2Biostatistics Department, University of Washington, Seattle, WA, USA

*Corresponding author. Biostatistics and Biomathematics Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA. E-mail: mspepe@u.washington.edu

Accepted 14 April 2011

New approaches have been developed in recent years to quantify the improvement in prediction performance gained by adding a novel marker to a set of baseline predictors of risk. The paper by Tzoulaki et al.1 concerns risk reclassification techniques and focuses specifically on the net reclassification improvement (NRI) index. Their review shows that use of risk reclassification analysis is extremely common in practice, with 51 papers using the technique published in only 3 years since its introduction. Unfortunately and alarmingly, the review shows that the quality of reporting is dismal. Investigators seem confused about the roles and interpretations of risk reclassification metrics. Guidance on how to report results of risk reclassification analysis would be helpful to authors, reviewers and the field in general.

The risk reclassification table was first introduced by Cook.2 The table is constructed by choosing clinically meaningful risk categories and cross-classifying individuals according to their risks calculated with the baseline risk model and with the expanded risk model. The top panel of Table 1 provides an illustration. Cook and Ridker3 developed a whole analysis strategy around the risk reclassification table including new hypothesis tests and a new metric called ‘percent correct reclassification’. However, the value of these analysis techniques is doubtful and results can be misleading.4 Pencina et al.5 argued that the reclassification table itself was problematic, at least as proposed by Cook, because it did not distinguish between subjects with events (cases) and subjects without events (controls). They suggested constructing separate event and non-event reclassification tables as shown in the middle and bottom panels of Table 1. Entries above the diagonal correspond to risks that are higher with the expanded vs baseline model, representing improved prediction for subjects with events. Correspondingly, entries below the diagonal represent worse prediction for them. The event-NRI is the difference between the proportions of subjects above vs below the diagonal in the event reclassification table. Using a similar logic, the non-event-NRI is calculated from the non-event reclassification table by taking the difference between the proportions of subjects below vs above the diagonal. The NRI summary index that gained immediate popularity in the literature following Pencina’s paper is the sum,

\[
\text{NRI} = \text{event-NRI} + \text{non-event-NRI}.
\]

A prerequisite for considering risk reclassification is that the risk models are well calibrated, in the sense that the observed event rates for subgroups defined by the predictors in the models are close to the values calculated from the models. A poorly calibrated risk model is considered invalid for calculating risk as a function of the modelled predictors. It is of great concern therefore that almost half of the papers reporting risk reclassification results do not report assessment of model calibration. A second basic premise for considering risk reclassification is that the chosen categories of risk are clinically meaningful in the sense that changing risk categories has clinical consequences. The review indicates that only 27% of papers provided justification for the particular risk categories used. This is a very poor state of affairs.

Even if the risk models are valid and the risk categories chosen are clinically relevant, is NRI a good way of summarizing improvement in risk reclassification performance? We do not find this single numeric summary very enlightening. Calculated as 17.4% in our example, the NRI seems to fall short of the task of gauging whether or not a substantial improvement has been obtained. Somewhat more revealing are its components, event-NRI and non-event-NRI. If only two risk categories were involved, the event-NRI is the increase in the proportion of subjects with events that are classified as high risk by the predictors and correspondingly the non-event-NRI is the increase in the proportion of subjects without events who are deemed at low risk. These are simple useful...
summarizes. However, with more than two categories the interpretations are far less appealing because all upward movements of risk category are counted equally and all downward movements are counted equally. Yet, the clinical implications are usually not equal. For example, moving from the lowest to highest or moving from the lowest to intermediate categories typically has very different consequences. Perhaps a single numeric summary is not needed or at least should not be the main focus of analysis. One alternative suggestion is to report the net changes in proportions of subjects classified in each of the risk categories. These are (−2.1%, −6.3%, 8.4%) for subjects with events and (7.1%, −6.6%, −0.5%) for subjects without events in Table 1. In other words, of subjects with events, 8.4% more are in the high-risk category and 2.1% fewer are in the low-risk category, whereas of subjects without events, 7.1% more are in the low-risk category and 0.5% fewer are in the high-risk category. These are simple summaries of reclassification performance that seem more clinically relevant than the NRI index of 17.4%.

Although risk reclassification analysis with the NRI has taken off like wildfire in applications, it is not yet a highly developed rigorous statistical technique. Unfortunately, this point is not widely appreciated and it is not acknowledged in the review. In particular, statistical techniques used for hypothesis testing and for calculating confidence intervals for the NRI have not been validated. Of concern is the fact that current techniques do not account for sampling variability in model parameter estimates and this has been shown to render techniques for hypothesis testing grossly invalid when applied to the change in area under the ROC curve (AUC) statistic and to the integrated discrimination improvement (IDI) statistic. Similar issues are likely to pertain to the NRI. Furthermore, concerns were raised about use of the NRI in a series of commentaries on the original 2008 paper. Important modifications to the NRI statistic have been suggested by Pencina et al. since, but many of the original concerns remain including the technical concerns. In our opinion, the jury is still out on if and how the NRI should be applied in practice and reliance on it to summarize improvement in prediction performance seems premature for now.

A consensus set of guidelines for reporting of genetic risk prediction studies has appeared recently in several journals as ‘The GRIPS Statement’. This is a welcome step forward and, since most of the guidelines apply to general risk prediction studies, journals would do well to adopt them more broadly. With regard to analysis, the guidelines stress assessment of model calibration. We can hope therefore that future papers will be more consistent about reporting measures of model fit than were found in the review by Tzoulaki et al. However, no specific guidance is offered on how to quantify improvement in prediction performance by adding genetic factors to a risk prediction model, reflecting the lack of consensus among experts in the field at present. Perhaps in the future, a consensus for quantifying improvement in risk prediction performance will be developed and could be added to the GRIPS statement to offer more complete guidance to investigators, readers and journal editors.

Conflict of interest: None declared.

Funding
This work was supported by the National Institutes of Health [R01 GM054438 and U24 CA086368].

References
5 Pencina MJ, D’Agostino RB Sr, D’Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a


