PIM3 North American Development

[calibration, validation and comparison to PIM3]

Gerardo Soto-Campos[1], Lahn Straney[2], Tom Rice[1], and Randall Wetzel[1,3]

[1]VPS, LLC, [2]Monash University, [3]Children's Hospital of Los Angeles

Introduction

Scoring systems which were developed within a particular population, or a particular health care system, may require recalibration to properly risk-adjust outcomes when applied in a different population[1,2]. Recalibration implies that the set of coefficients associated with a particular set of risk factors predicting a specific outcome, such as mortality, are actually recomputed using the data from the new population. Leteurtre et al[1] found when PIM2[3] was recalibrated with French and British PICUs the French recalibration did not fit the British population (Hosmer-Lemeshow p-value was 0.02), but the British recalibration did calibrate well with the French population (Hosmer-Lemeshow p-value = 0.36). In fact, Straney et al[4] also recommend periodic recalibration when using new data to ensure applicability of these models. We describe the methodology and results found when reestimating the coefficients associated with the set of risk factors in Straney's PIM3 model[4], using a large North American PICU dataset.

Methodology

We first applied PIM3 published coefficients[4] using the North American VPS dataset and found good discrimination[i] (AUROC curve[5] was 0.889), however the p-value of the Hosmer-Lemeshow calibration test[6] was less than 0.05, which could be interpreted as a lack of fit of the model with the North American data. This suggested a need for recalibration.

We first estimated the discriminative power of the model studied by a standard model validation approach. This was achieved by computing the area under the receiver operating curve[5]. Next, we assessed the goodness of fit of the model, i.e. calibration, using a conventional χ^2 analysis. In this approach deciles of the predicted risk are compared against the corresponding actual outcome, the Hosmer-Lemeshow test. The utility of this test comes from estimating the experimental χ^2 statistic of the model, and compares it against the published χ^2distribution with eight degrees of freedom. If the p-value of the test is larger than 0.05 one assumes the model calibrates properly. Conversely, if the p-value is smaller than 0.05 it is inferred that the outcome predicted by the model does not properly fit the actual outcome, and that the actual and predicted outcomes, in the data used to test the hypothesis, are statistically significantly different.

Traditionally the Hosmer-Lemeshow test[6] has been the dominant methodology to study model calibration in health care analytics. However, the Hosmer-Lemeshow test runs into serious inconsistencies[7] when the number of samples utilized to calibrate models is large (around N>25,000).

One alternative often used to avoid the inconsistencies generated by applying Hosmer-Lemeshow tests to large samples is the so called GiViTI calibration belt[8,9,10,11]. In this approach, a confidence band for the

calibration is built and graphically displayed showing good or bad calibration depending on whether the upper and/or lower regions of this confidence interval intersect and/or cross the 45 degree line obtained when plotting observed versus predicted mortality deciles of risk. This line is called the *bisector*, and it corresponds to the ideal situation of a perfect model fitting all the outcomes studied in the data.

The GiViTI calibration belt mathematically depends on a complex polynomial expansion relating the expected to observed probabilities of the outcome[8,9,10,11]. In practical terms, calibration with the GiViTI calibration belt corresponds to the case where the confidence intervals of the belt, do not overlap with the *bisector* line. Figures 1,2,3 below exemplify this situation.

We used 278,011 consecutive records from the North American VPS dataset from July 1, 2013 to March 31,2017 distributed among 118 VPS units. 246,656 records from July 1, 2013 – December 31,2016 were used for the recalibration and 32,355 records from January 1,2017 – March 31,2017 were utilized for the validation process. Each record collected PIM3 variables as defined in the original PIM3 paper by Straney et al[4]. That is, *pupils fixed*, *elective admission*, *mechanical ventilation in the first hour*, *absolute value of base excess*, *systolic blood pressure*, *ratio of* FiO_2/PaO_2, *recovery post-procedures*, as well as diagnoses for *low-risk*, *high-risk*, and *very high-risk*. Using these risk factors, two models were constructed: 1. a **fixed effects** model, and; 2. a **mixed effects** model[12]. It is important to remember that mixed effects models contain a fixed effect contribution plus a random contribution. In our case the fixed effect was the set of risk factors from Straney's original PIM3 model[4], and the random effect was the the geographical region in the United States or Canada. In particular, we included these regions: North East, Mid West, South, West, and Canada.

We computed the probability of death with the set of risk factors for PIM3, and the model described by Straney et al[4]. The discrimination of both models was assessed with the standard area under the ROC curve, as well as the calibration using both, the Hosmer-Lemeshow[6] test and by the GiViTI calibration belt[8,9,10,11] since the calibration belt may be better suited for situations when the number of records exceeds 25,000 (our sample size 246,656). For the Hosmer-Lemeshow test an indication of good calibration would be a p-value larger than 0.05. And for the GiViTI calibration belt a graphical representation of a confidence area around the bisector.

Results

We performed the validation studying two data sets. One, case (a), where the entire set of 246,656 records was used to recalibrate a new set of PIM3 coefficients and was also used in the validation of the new model. The second case (b), used an independent sample of size N= 32,355 records, of VPS data collected during January 1,2017 to March 31,2017.

Table 1. Area under the ROC curve for fixed and mixed effect PIM3 models for a validation set of 32, 355 cases. The random effect of the mixed effects model was geographical region in North American PICUs.

Model	AUROC
Straney's model	0.889
Fixed Effects	0.895
Random Effects	0.894

Calibration

The Hosmer-Lemeshow test was applied by taking twenty consecutive samples, with replacement, of size 25,000 from the whole set of 246,656. For these twenty samples, the Hosmer Lemeshow p-value for Straney's original model was **always** less than 0.05. However, in the case of the newly built set of random effects coefficients, fourteen out of twenty cases had a p-value larger than 0.05. This may be due to the fact that a sample of size 25,000 is already very large, and in this instance, according to Paul[7], it could be already pushing the limits of reliability for this test.

The results using the calibration belt with the GiViTI R package are shown in Figures 1, 2, and 3, below

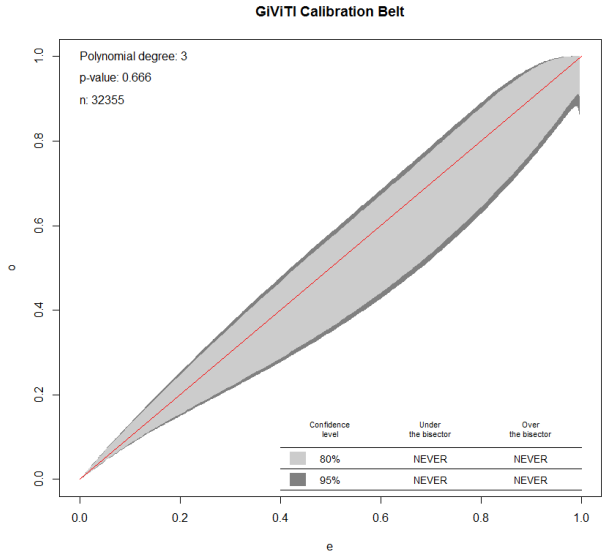

Figure 1. GiViTI calibration belt for mixed effects model using the validation test of 32,455 samples independent of the dataset used to train the coefficients in the new PIM3 model.

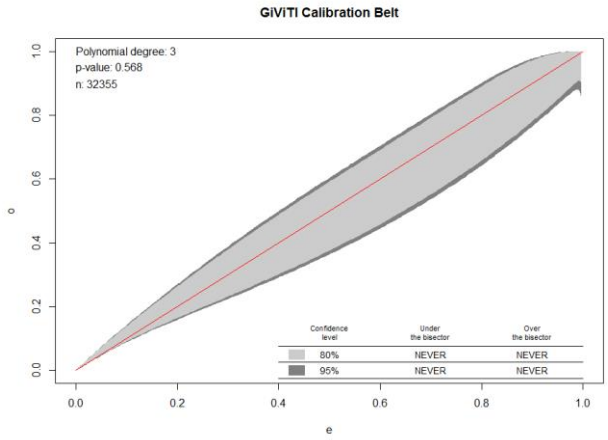

Figure 2. Giviti calibration belt for fixed effects model using the validation set (N=32,455), and the retrained set of coefficients for the new PIM3 model.

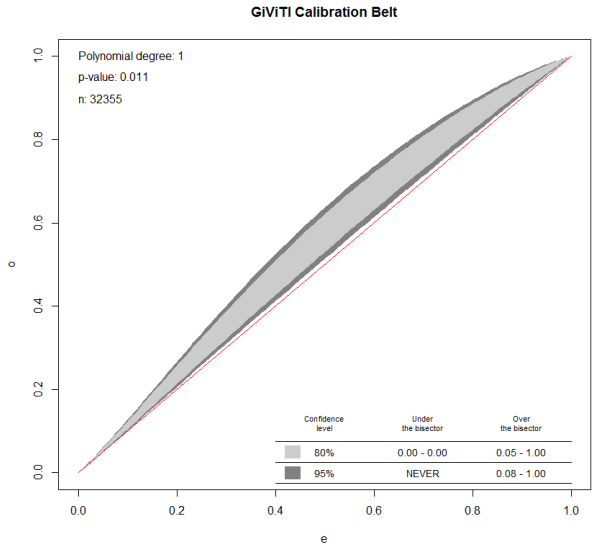

Figure 3. GiViTI calibration belt using Straney's PIM3 set of coefficients published in the literature[4]. The data used to generate this calibration belt was the validation set, which consisted of 32,355 records from all the VPS units in Q2-2017.

Conclusions

Our primary objective for the project was validation of PIM3 for use in North America utilizing the VPS database and accepted statistical techniques found in the medical literature. We retrained the coefficients of the PIM3 set of risk factors published by Straney et al in the original PIM3 paper[4] using North American VPS data from Q3-2013 to Q2-2017. Utilizing the methodology in Straney's paper[4], we built two models: 1. a fixed effects, i.e. Straney's risk factors, and 2. a mixed effects model with fixed

effects equal to Straney's risk factors fixed, and a random component corresponding to geographic region in VPS North American PICUs.

Table 1 presents the results for the area under the ROC curve for the new models using the validation set (N=32,455). These results show good discriminative power. One should notice that Straney's original, also had good discrimination, AROC=0.889 with the validation set (N=32,455) and AROC=0.895 with the training set (N = 246,656).

We validated our model using internal and external validation methods. First, internally using the entire set of 246,656 records from July 1, 2013 to December 16, 2016. Second, externally using the prospectively collected data from January 1, 2017 to March 31, 2017 (N=32, 355).

In the case of calibration, our results show that Straney's original PIM3 model does not pass the Hosmer-Lemeshow test, and does not pass the GiViTI calibration belt test, see Figure 3. On the other hand, the newly retrained fixed and mixed effects models pass the GiViTI calibration belt test. With respect to the Hosmer-Lemeshow test, the mixed effects model performs much better than the fixed effects model when using random samples of size 25,000. The mixed effects model is closer in methodology to Straney's original approach. VPS will utilize and adopt the newly retrained mixed effects model for inclusion as VPS version of PIM3, and propose it be called PIM3 North America[ii] (PIM3 NA).

It is important to notice that since we recalibrated Straney's original model[4] with a North American population, applying it outside North America may produce discrepancies similar to the case reported by Leteurtre et al[1], when PIM2 recalibrated with French patients was used in British cohorts. Within North America, since the VPS database exclusively includes members from the USA and Canada, PIM3 NA is appropriate for comparing North American ICUs to one another.

To capture time dependencies or drift in the VPS cohort, we will periodically recalibrate the coefficients in PIM3 North America.

References

1. Leteurtre S, Grandbastien B, Leclerc F, et al, International comparison of the performance of the paediatric index of mortality (PIM) 2 score in two national data sets, Intensive Care Med (2012) 38: 1372-1380.
2. Brady AR, Harrison D, Black S, et al, Assessment and Optimization of Mortality Prediction Tools for Admission to Pediatric Intensive Care in the United Kingdom, PEDIATRICS (2006) 117: e733-e742.
3. Slater A et al, PIM2: a revised version of the Paediatric Index of Mortality, Intensive Care Med (2003) 29: 278-285.
4. Straney L, Clemens A, Parslow RC, et al, Paediatric Index of Mortality 3: An Updated Model for Predicting Mortality in Pediatric Intensive Care, Pediatric Critical Care Medicine (2013) 14: 673-681.
5. James G et al, An Introduction to Statistical Learning (Springer, New York, 2013).

6. Hosmer DV and Lemeshow S, Applied Logistic Regression (John Wiley & Sons, New York, NY, 1989).
7. Paul P, Pennell ML, and Lemeshow S, Standardizing the power of the Hosmer-Lemeshow goodness of fit test in large data sets, Statistics in Medicine (2012) 32: 67-80.
8. Finazzi S, et al, Calibration Belt for Quality-of-Care Assessment Based on Dichotomous Outcomes, PLoS ONE (2011) 6:e16110.
9. Package 'givitiR' (CRAN, https://cran.r-project.org/web/packages/givitiR/givitiR.pdf).
10. Gruppo Italiano per la Valuatazione degli Interventi in Terapia Instensiva (http://www.giviti.marionegri.it/).
11. Wasserman J, A Comparison of the GiViTI Calibration Belt to Hosmer-Lemeshow Goodness of Fit, Masters Thesis, Ohio State University, 2012.
12. Kleinman K and Horton NJ, SAS and R Data management, Statistical Analysis, and Graphics (CRC Press, San Francisco, CA, 2010).

[i] Discrimination in this context is referred to how well the model separated true positives outcome predictions from the false positive ones. It is quantified with the area under the curve of a receiver operating curve, and it is denoted with the acronym AROC. The closer to 1.0 the AROC is, the better discriminatory power the model has.
[ii] A more in depth study will be written and published separately summarizing all this research.