**SUPPLEMENTARY MATERIAL**

**Appendix A: Calibration in-the-large, calibration slope, ICI, E50 and E90**

Two measures that are usually used as summary statistics of the calibration plot [Moons et al. 2015] are an intercept , called *calibration in the large,* and a slope *,* also known as *calibration slope.* These two measures describe deviation of the calibration curve from the ideal 45º line characterized by an intercept zero and slope one. Perfect calibration corresponds to =0, and =1. A values ofsmaller than one indicates that forecast probabilities are too extreme: low predictions are too low and high predictions are too high (*i.e.* the P-CRA model underestimates the actual risk of new caries when it outputs low forecast probabilities, and overestimates the actual risk of new caries when it outputs high forecast probabilities). On the other hand, a value of  greater than one indicates that forecast probabilities are not sufficiently extreme: low predictions are too high and high predictions are too low (*i.e.* the CRA model overestimates the actual risk of new caries when it outputs low forecast probabilities, and underestimates the actual risk of new caries when it outputs high forecast probabilities).

The calibration in the large, , compares the mean forecast probabilities with the mean observed risk (i.e. with the incidence of caries in the sample). A positive (negative) value indicates that the mean forecast probabilities are smaller (greater) than the observed risk. If the calibration slope is one ore close to one, the calibration in the large indicates the extent to which the forecast probabilities are systematically too low (>0) or too high (<0).

As a complementary set of numerical indices to characterize the calibration curve and to quantify its deviation from the ideal 45º line, the Calibration Index (ICI) and two related metrics E50 and E90, are used in the paper. The ICI is a weighted average of the absolute difference between the calibration curve and the diagonal line of perfect calibration, where the absolute differences are weighted by the density function of forecast probabilities. The larger is the value of the ICI the larger is the (weighted) distance of the smoothed calibration curve from the ideal 45º line. The E50, and E90 metrics represent the median, and the 90th percentile, respectively, of the absolute difference between (smoothed) observed and predicted probabilities of new caries. They all have a simple and appealing interpretation as summary statistics of the absolute differences between observed frequencies and predicted probabilities of new caries.

**Appendix B: What can be said about calibration using the information in the published P-CRA models literature?**

In order to appreciate the added value of the proposed framework with respect to current reporting of P-CRA models predictive performance it is instructive to consider what could be said about calibration from the statistical analysis reported in published research on P-CRA models. The only information somehow related with calibration, sometimes published in the literature (see Table 1), consists of the relative frequencies of subjects that developed new caries at follow-up stratified by the P-CRA model risk group and the number of subjects in each risk group. For example, in [Campus et al. 2012], the authors partition the children in the study in 5 risk groups according to the Cariogram forecast probabilities of new caries at baseline. The first group corresponds to children with forecast probabilities of new caries in the interval [0, 0.2); the last group corresponds to children with forecast probabilities of new caries in the interval [0.8, 1]. If the frequency of subjects with new caries per group was available, it would be possible to construct an approximate standard calibration plot in which the number of *strata* (groups) equals the number of the model risk groups (see Figure S1b). This standard calibration plot, however, provides a very imprecise assessment of the calibration of the model. In addition to the usual limitations of standard calibration plots discussed above the standard calibration plot in Figure S1b, suffers from three additional drawbacks. First of all, the number of strata is very small thus providing a very rough description of the calibration performance of the Cariogram across the range of predicted probabilities of new caries. In addition, and most importantly, the information of the mean predicted probability of new caries in each risk group is not available. Thus, for each interval defining a risk group, the middle point of the interval is taken as representative value of the forecast probabilities that follow in the interval. This procedure, however, might be misleading if the distribution of forecast probabilities in each stratum is not uniform. In these cases, the middle points of the intervals might be very poor representative of the forecast probabilities in the intervals and the corresponding standard calibration plot should be interpreted with caution. The according to approximate standard calibration plot in Figure S1b, the Cariogram forecast probabilities for the bin [0.8, 1] match perfectly the observed frequencies of new caries (the corresponding point lies on the ideal 45º line). However the smoothed calibration plot in Figure 1a, shows that the distribution of the forecast probability in the bin [0.8, 1] is not uniform at all (the forecast probabilities are concentrated around 0.8) and that the Cariogram, in fact, underestimates the actual chances of new caries when it outputs forecast probability larger than 0.6.

Finally no uncertainty measure can be provided about the observed frequencies and the predicted probabilities in FigureS1b which makes very difficult to decide whether how, and how much, the calibration curve deviates from the ideal 45º line.

It is important to note that these limitations are common to all published research on predictive performance of P-CRA models. In particular, the same analysis of reporting and potential assessment of calibration based on published research performed and described in this paper can be replicated exactly for all the studies in Table 1 (see Figure S2-S3)



**Figure S1**: (a) Distributions of the forecast probabilities of new caries for the children that did and did not developed new caries in [Campus et al., 2012]. The distributions are represented as rescaled histograms. Note that for the histogram of children that did not developed new caries the y-axis has been reversed. (b) Approximate standard calibration plot for the data in [Campus et al., 2012]. Five bins of the same size, and middle point of the bins have been used to build the plot. Dots represent the observed versus predicted probabilities per bin with the sample size of the bin in parenthesis (and are joined by a solid line). The dashed line represents the ideal 45º line of perfect calibration,



**Figure S2:** Approximate calibration plots for the prospective studies in Table 1. Study population: preschool and school children. Only studies with at least four risk categories have been considered. NUS-CRA denotes the National University of Singapore Caries Risk Assessment model. Dots represent the observed versus predicted probabilities per risk category with the sample size of the risk category reported next to each point. The solid gray and black lines joint the dots corresponding to the Cariogram and the NUS-CRA model respectively. The ideal 45º line of perfect calibration is represented as a dashed line.



**Figure S3:** Approximate calibration plot for the prospective studies in Table 1. Study population: young adults and elderly. Only studies with at least four risk categories have been considered. Dots represent the observed versus predicted probabilities per risk category with the sample size of the risk category reported next to each point. The solid gray and black lines joint the dots corresponding to the Cariogram and the NUS-CRA model respectively. The ideal 45º line of perfect calibration is represented as a dashed line.