The Medical Journal of Australia

Supporting Information

Supplementary methods and results

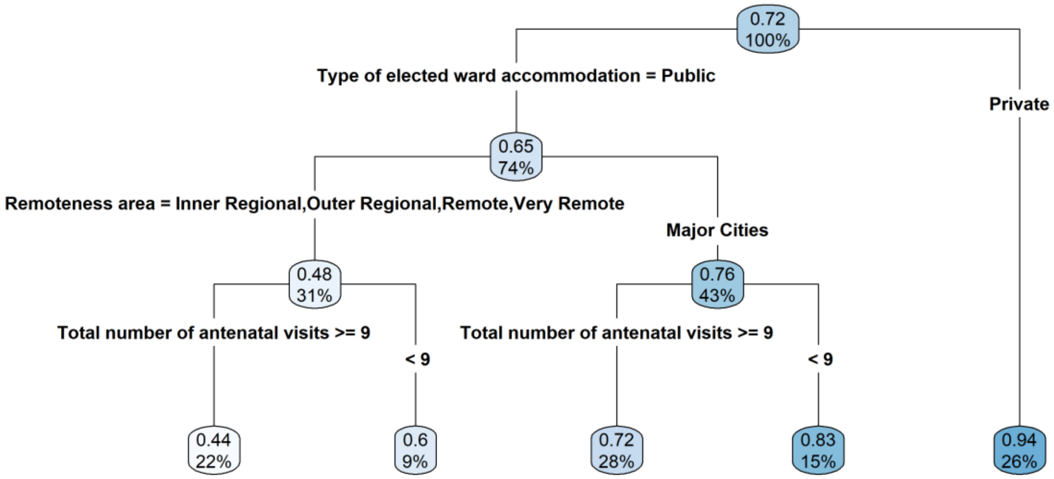
Appendix to: Meloncelli NJL, Barnett AG, Cameron CM, et al. Gestational diabetes mellitus screening and diagnosis criteria before and during the COVID-19 pandemic: a retrospective pre–post study. *Med J Aust* 2023; doi: 10.5694/mja2.52129.

Classification tree for missing glucose data

The aim of this analysis is to compare the characteristics of women who did and did not have glucose data. Given the large number of potential predictors, we decided to use a classification tree as this is a useful approach for selecting predictors from a larger subset. A classification tree with a binary outcome is an alternative to the more familiar multiple logistic regression. Trees make predictions about individuals by using a series of binary splits that aim to create leaves where individuals in the same leaf are homogenous, whilst individuals between leaves are heterogeneous. Trees and logistic regression can both be used to make binary predictions about individuals that can be compared with the observed binary outcome. One important difference of trees to standard logistic regression is that trees serially select the next most predictive variable until the predictive performance cannot be improved or a pre-specified tree depth is reached. Hence a tree may not use every available predictor and therefore also performs variable selection. Trees are protected against over-fitting by using cross-validation to assess the predictive performance. For a detailed discussion of trees see “Classification and Regression Trees", 1984, Routledge, by Leo Breiman, Jerome Friedman, Charles J. Stone, R.A. Olshen. ISBN 9781315139470.

We fitted the tree using the rpart package (version 4.1.19) in R. We used 10-fold cross validation, with a minimum split size of 20 individuals, minimum leaf size of 7 individuals, and maximum tree depth of 30 splits.

Figure 1. Classification tree for predicting if women had missing glucose data. Each node in the tree shows the proportion with missing data (top) and the overall percentage of the sample (bottom)



We used a classification tree to estimate for whom glucose tests were missing. The tree included the following 27 predictors: pregnancy sequence id, admission date, Indigenous status, accommodation status, mother’s birth year, antenatal transfer flag, previous pregnancies flag, number of previous pregnancies, last birth method status, number of previous caesarean deliveries, discharge status, total number of antenatal visits, weeks gestation at 1st antenatal visit, alcohol use before 20 weeks gestation, alcohol use after 20 weeks gestation, illicit drug use screen, smoked before 20 weeks gestation, smoked after 20 weeks gestation, estimated confinement date, body mass index, antenatal care flag, medical condition flag, pregnancy complications flag, source, baby's birth year, GDM diagnosis, ASGS remoteness area 2016.

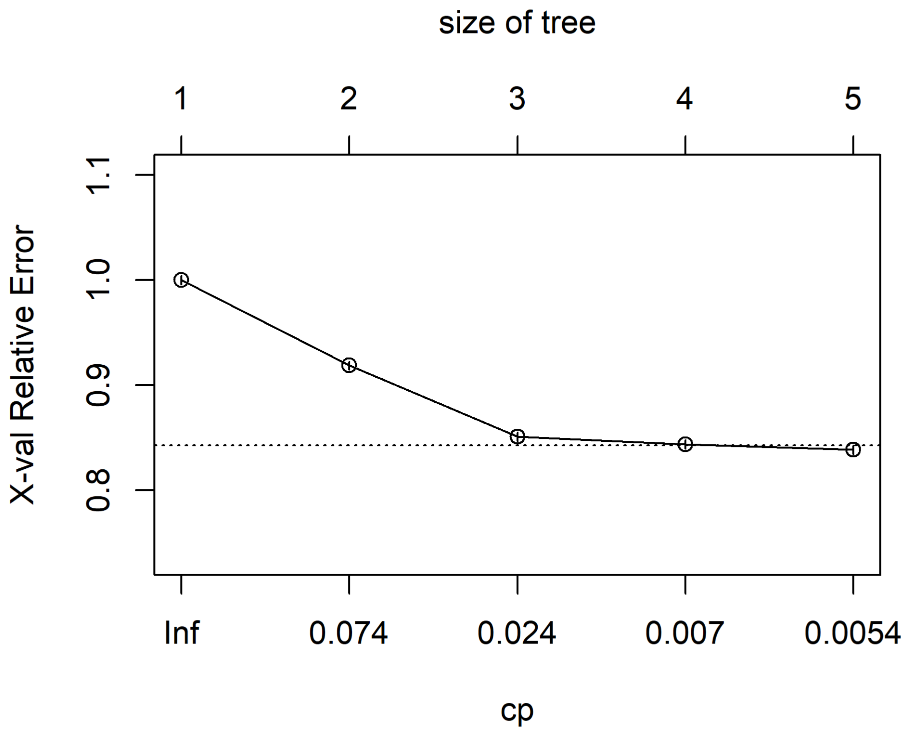
The final tree included three predictors: antenatal care type (public, private), number of antenatal visits, and remoteness area.

The leaves in the tree show the proportion of women for whom glucose data were missing and their proportion of the entire sample. The right-most leaf, for example, data were missing for 94% of women who received private antenatal care, and comprised 26% of the sample.

Glucose data were missing for a much larger proportion of women living in major cities and those who had fewer than nine antenatal visits.

The total number of pregnancies included in the analysis was 57,891. The proportion of women for whom glucose data were not available was 0.72 (44,209 women).

Figure 2. Cross-validated error for the tree



The plot above confirms that the cross-validated error decreased as the tree complexity (cp) increased. It also shows the reduction in model performance with increasing complexity, with the diminishing returns for additional predictors.