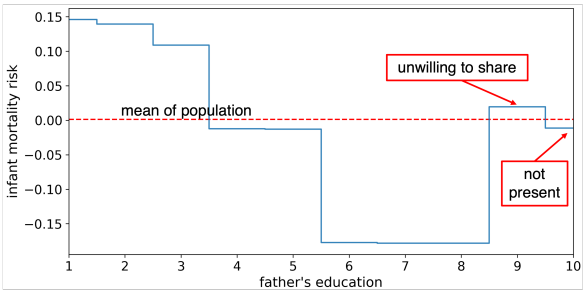


Figure 7: Impact of father's education on infant mortality risk, 2013.

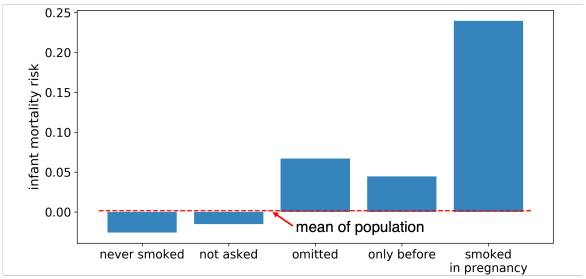


Figure 8: Impact of smoking before and during pregnancy on infant mortality risk, 2013.

Appendix A. Testing for MCAR with EBM: Case Study

In some cases we have information about the mechanism generating missing values and the likelihood that a similar mechanism will generate data in the future.

As an example, consider CDC Birth Cohort Linked Birth – Infant Death Data Files United States Department of Health and Human Services (US DHHS) et al.. The dataset describes pregnancy and birth variables for all live births in the U.S. together with an indication of an infant's death before the first birthday. The dataset is collected using two certificates: 1989 Revision of the U.S. Standard Certificate of Live Birth (unrevised) and the 2003 revision of the U.S. Standard Certificate of Live Birth (revised). As a result of the delayed, phased transition to the 2003 Certificate, the cohorts from 2004 to 2015 include data for reporting areas that use the newer 2003 revision along with data for reporting areas that still use the older 1989 Certificate (unrevised), with later years having a larger fraction of data corresponding to the 2003 revision. Values for variables that are present only in the 2003 certificate will be missing for areas using the earlier, 1989 certificate. In 2013, 10% of records come from such areas, the fraction is declining year to year and we can expect it to be even smaller in subsequent years.

Figure 7 shows the impact of father's education on infant mortality risk according to an EBM model trained on 2013 data. Values from 1 to 8 correspond to different levels of educational attainment, with 1 indicating 8th grade or less and 8 a doctorate or professional degree. The risk is high for levels 1-3, drops to just below the average risk for levels 3-4 (some college and associate degree) and even further for BA/BS, MA/MS and doctorate (levels 6-8)[2]. Level 9 indicates unwillingness to share this information and 10 corresponds to 10% of records where this variable was not present (version 1989). Level 9 is associated with slightly elevated risk; we may guess that fathers unwilling to share are more likely to be lower on the education scale. Level 10 is associated with risk slightly below average, which is surprising at first glance. Unlike for Level 9, the mechanism according to which the information is withheld is independent of the value of the variable in question (namely, the geographical area using an older version of the certificate). However, if the populations using the two certificate versions were coming from the same distribution, we would expect average risk (0 on the shape function) for this group. The MCAR test from Section 4.1.1 indicates these groups are statistically different from each other, suggesting social, demographic or other differences between these populations.

A similar picture emerges when we look at infant mortality as a function of mother smoking before and during pregnancy. The risk is highest for mothers who smoked during pregnancy, slightly elevated for those who smoked before pregnancy and lowest for mothers who never smoked. Risk for mothers who didn't share this information ('omitted') is clearly elevated. The group for whom the value is missing (older 1989 certificate, denoted 'not asked') has risk slightly lower than average (0). Again, risk different from average indicates a distribution shift with respect to the rest of the population, and we see that 'omitted' is different from 'not asked'.

If we were to train an infant mortality risk model on 2013 data and use it for prediction on data from subsequent years, we could run into the problem of

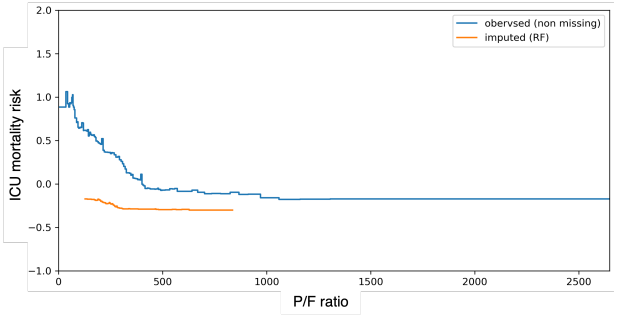
2. Parents' education is the best proxy we have in the dataset for family's income.

values missing for an even lower fraction of all records and possibly coming from a distribution even more shifted with respect to the distribution of the majority of the records. Our model would likely predict the risk less accurately for this segment of the population.

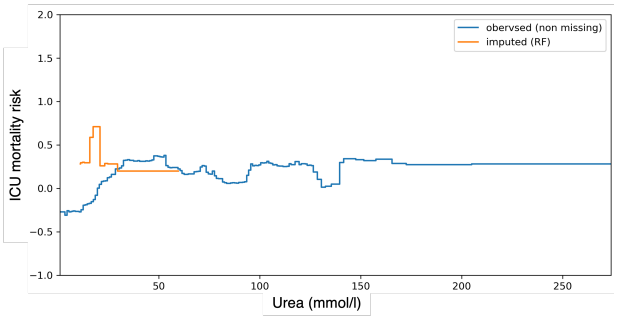
Appendix B. Visualizing the Effect of Imputation

As mentioned in Section 5, the advanced imputation methods can significantly change the learned shape functions and such changes can sometimes be problematic. To help visualize the effect of imputation and identify potential problems in advance, we propose to separate the components of the missing group and the observed group in the EBM shape functions. To separate these two components, instead of directly imputing the missing values with the output of the imputation algorithm, we add a large offset to these imputed values so that the imputed values do not have overlap with the observed values. For example, in our experiments, we add max feature value plus 1 to the imputed values. This can be viewed as a trick to squeeze the feature and its missingness indicator variable into one dimension. Training EBMs on such separated feature values, the shape function will be a concatenation of the two curves corresponding to the observed group and the missing group. Also, because we know the offset we added to the imputed value, we can subtract it during visualization, and show the two curves on the same plot and original x-axis.

Figure 9 shows the EBM shape functions of the imputed group and the observed group separated using the method proposed above. Figure 9(a) shows that the risk of the RF imputed group is much lower than the risk of the observed group which corroborates what we found in Figure 6(a). Similarly, the effects of the imputed group in Figure 9(b) also differ significantly from the observed group, which explains why there exist spikes in the RF imputed EBM shape function in Figure 6(b). Using interpretable methods like EBMs allows one to understand the consequence of different imputation methods that otherwise would be invisible.



(a) Shape functions for P/F ratio



(b) Shape functions for Urea

Figure 9: EBM shape functions when the effects of imputation group (imputed by MissForest, denoted as RF imputed) and observed (non missing) groups are separated. The plots suggests how the two groups are different in terms of predicting the ICU mortality risk, and suggests how MissForest imputation might result in problematic models.