Exploration of Misssingness

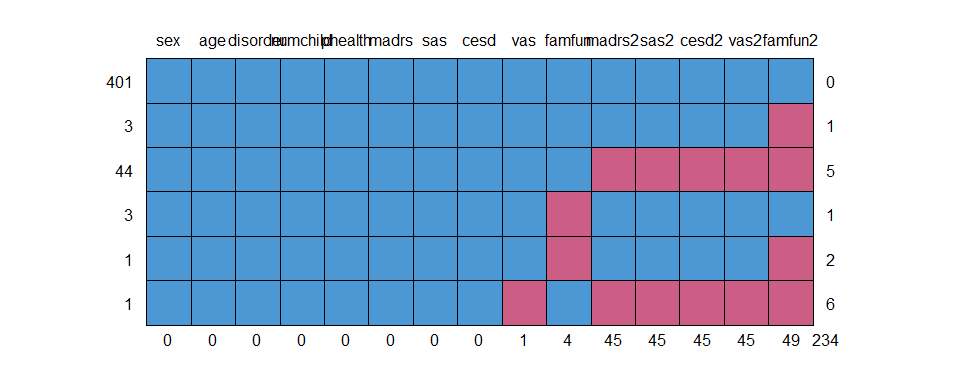
Florian Stijven

2023-09-19

In this document, we explore the missingness in the Browne data following the multiple imputation approach we have used. First, we summarize the different missingness patterns that appear in these data. Second, we examine how well the missing variables can be predicted from the observed variables in the two imputation models that we have used: (i) global imputation, and (ii) imputation per arm.

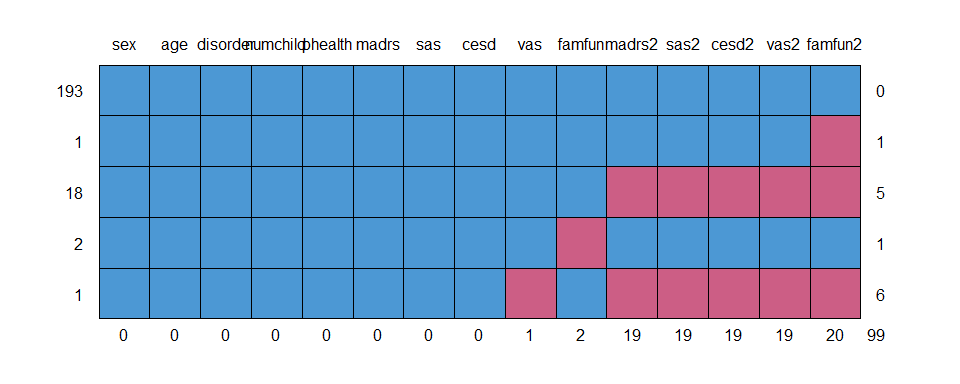
# Missingness Patterns

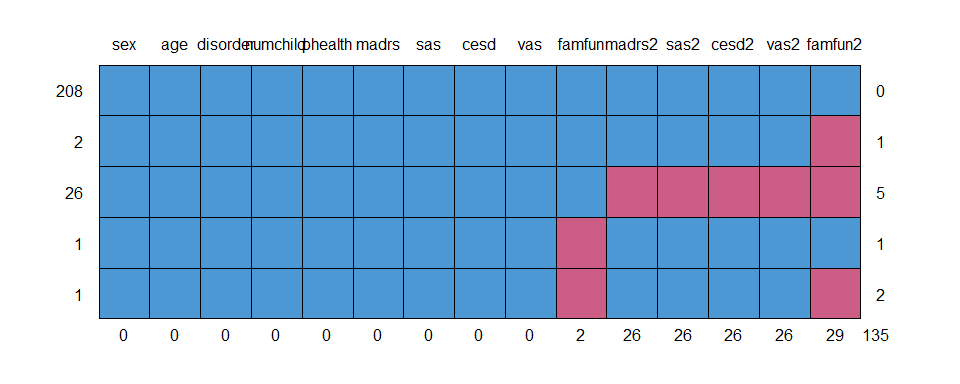
In the next figure, the missing data patterns are summarized for the patients in the “Sertraline alone” and “Sertraline and IPT” groups. The first column provides the frequency of each pattern. The last column lists the number of missing entries per pattern. The bottom row provides the number of missing entries per variable, and the total number of missing cells.



The above figure shows that almost all missingness is due to missingness in the outcome variables measured 6 months after randomization. In addition, most missingness is due to drop out before the 6 months measurement occasion. Therefore, a missing outcome at 6 months is generally imputed using only the baseline covariates, *but not the other outcomes at 6 months*. The accuracy of the imputations will thus depend mostly on how well the baseline covariates can predict the outcomes at 6 months.

The same graph is repeated for the treatment groups separately.





# Accuracy of Imputations

In this section, we will examine how accurately missing observations can be imputed in the two imputation models we have considered: (i) global imputation and (ii) imputation per arm. We will examine two aspects of this accuracy:

1. How well can the missing values be predicted from the observed values? This is quantified in the adjusted R-squared values for the prediction models for each of the 5 outcomes measured at 6 months.
2. How much uncertainty is there in individual predictions? This is quantified by the standard error for the predicted mean outcome for patients with missing outcomes at 6 months. We additionally consider the width of the associated prediction intervals.

To achieve the above goals, we fit two linear regression models corresponding to the two imputation models.

* The first linear regression model uses all baseline covariates in the linear predictor *without* interaction terms. This corresponds to global imputation.
* The second linear regression model uses all baseline covariates *and* interaction terms with treatment in the linear predictor. This corresponds to imputation per arm.

In the next table, we summarize the adjusted R-squared values for the linear regression models. We make two important observations:

1. The adjusted R-squared values are relatively low. This means that the missing values cannot be very accurately predicted from the observed values. The accuracy is especially low for VAS. For FAMFUN and SAS, the accuracy is actually moderate.
2. There is almost no difference in adjusted R-squared values between the two linear regression models. Missing values can thus *not* be more accurately imputed in the imputation per arm model as compared to global imputation.

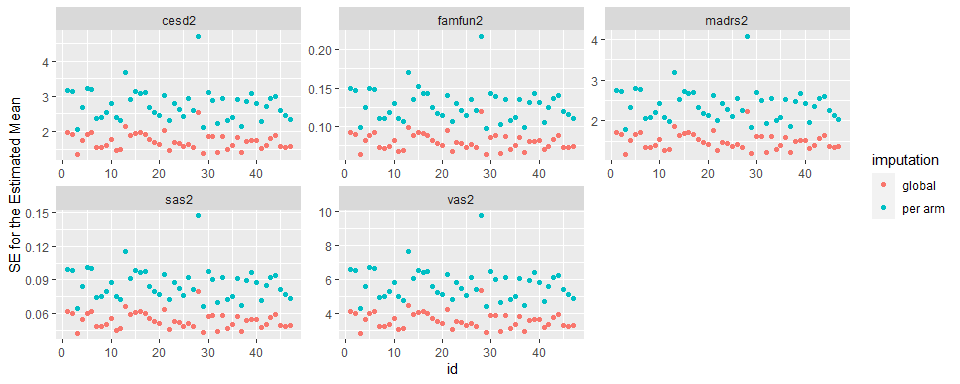
Adjusted R-squared values for linear regression models corresponding to global imputation and imputation per arm.

| outcome | global | per arm |
| --- | --- | --- |
| cesd2 | 0.264 | 0.255 |
| famfun2 | 0.369 | 0.372 |
| madrs2 | 0.226 | 0.221 |
| sas2 | 0.355 | 0.347 |
| vas2 | 0.152 | 0.148 |

We can thus conclude that there is no real gain in accuracy when doing imputation per arm. But is there an increase in uncertainty when doing imputation per arm? In what follows, we consider the standard errors for the predicted means for patients with missing outcome values at 6 months (and completely observed baseline covariates).

We only give the results for patients in the first two treatment groups because these patients were used for estimating the optimal treatment regimes.

The next plot shows these estimated standard errors for both linear regression models. Note that this only captures uncertainty in the model parameters.



To capture the full prediction uncertainty, we consider the widths of the prediction intervals for the same predictions as above. This figure shows that a small degree of uncertainty is added to the predictions if we do imputation per arm.

