Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., Wood, A. M., & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *British Medical Journal*, *339*, 157–160.

Tilling, K., Williamson, E., Spratt, M., Sterne, J. A. C., & Carpenter, J. R. (2016). Appropriate inclusion of interactions was needed to avoid bias in multiple imputation. *Journal of Clinical Epidemiology*, *80*, 107–115.

Tompsett, D. M., Leacy, F., Moreno-Betancur, M., Heron, J., & White, I. R. (2018). On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. *Statistics in Medicine*, *37*, 2338–2353. https://doi.org/10.1002/sim.7643.

van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*, 219–242.

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd edn). Chapman and Hall.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

Vansteelandt, S., Carpenter, J. R., & Kenward, M. G. (2010). Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology*, *6*(1), 37–48.

Von Hippel, P. T. (2009). How to impute interactions, squares and other transformed variables. *Sociological Methodology*, *39*, 265–291.

Welch, C., Petersen, I., Bartlett, J. W., White, I. R., MArston, L., Morris, R. W., Nazareth, I., Walters, K., & Carpenter, J. R. (2014). Evaluation of two-fold fully conditional specification multiple imputation for longitudinal electronic health record data. *Statistics in Medicine*, *33*, 3725–3737.

White, I. R., Daniel, R., & Royston, P. (2010). Avoiding bias due to perfect prediction in multiple imputation of incomplete categorical variables. *Computational Statistics and Data Analysis*, *54*, 2267–2275.

Wood, A. M., White, I. R., & Thompson, S. G. (2004). Are missing outcome data adequately handled? a review of published randomized controlled trials in major medical journals. *Clinical Trials*, *1*, 368–376.

Yucel, R. M. (2011). Random covariances and mixed-effects models for imputing multivariate multilevel continuous data. *Statistical Modelling*, *11*, 351–370.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Carpenter JR, Smuk M. Missing data: A statistical framework for practice. *Biometrical Journal*. 2021;*63*:915–947. https://doi.org/10.1002/bimj.202000196

APPENDIX A

A.1 | Consequence of MAR

For unit (individual) i, let $R_i = 1$ if X_i is observed, and 0 otherwise. Algebraically, the definition of MAR (Table 1) means $f(R_i|X_i, Y_i, Z_i) = f(R_i|Y_i, Z_i)$. Using the definition of conditional probability, this implies that the distribution of the partially observed variable, X, in the observed data, that is

$$\begin{aligned} f(X_i|Y_i, Z_i, R_i = 1) &= \frac{f(R_i = 1, X_i, Y_i, Z_i)}{f(R_i = 1, Y_i, Z_i)} \\ &= \frac{f(R_i = 1|X_i, Y_i, Z_i) f(X_i, Y_i, Z_i)}{f(R_i = 1|Y_i, Z_i) f(Y_i, Z_i)} \\ &= \frac{f(X_i, Y_i, Z_i)}{f(Y_i, Z_i)} \\ &= f(X_i|Y_i, Z_i), \end{aligned} \tag{A.1}$$

that is the distribution of X given Y, Z in the population. It is worth emphasising that this shows that MAR means that the distribution of X given Y, Z is the same *whether or not X is observed*.Therefore, under MAR, we can estimate the distribution of X given Y, Z in the observed data and use this (implicitly or explicitly) to impute the missing values of X.

A.2 | Criteria for validity of complete records for logistic regression

To obtain the results in Table 2, consider the odds ratio relating Y to binary X_1 at a fixed value of X_2. Suppose that the probability of a complete record depends on Y and X_2. Then the odds ratio in the complete records is

$$
\begin{aligned}
&\left\{\frac{\Pr(Y=1|X_1=1,X_2=x_2,R=1)}{\Pr(Y=0|X_1=1,X_2=x_2,R=1)}\right\}\times\left\{\frac{\Pr(Y=0|X_1=0,X_2=x_2,R=1)}{\Pr(Y=1|X_1=0,X_2=x_2,R=1)}\right\}\\
=&\left\{\frac{\Pr(R=1|Y=1,X_1=1,X_2=x_2)\Pr(Y=1,X_1=1,X_2=x_2)}{\Pr(X_1=1,X_2=x_2,R=1)}\right\}\\
&\times\left\{\frac{\Pr(X_1=1,X_2=x_2,R=1)}{\Pr(R=1|Y=0,X_1=1,X_2=x_2)\Pr(Y=0,X_1=1,X_2=x_2)}\right\}\\
&\times\left\{\frac{\Pr(R=1|Y=0,X_1=0,X_2=x_2)\Pr(Y=0,X_1=0,X_2=x_2)}{\Pr(X_1=0,X_2=x_2,R=1)}\right\}\\
&\times\left\{\frac{\Pr(X_1=0,X_2=x_2,R=1)}{\Pr(R=1|Y=1,X_1=0,X_2=x_2)\Pr(Y=1,X_1=0,X_2=x_2)}\right\}\\
=&\left\{\frac{\Pr(Y=1|X_1=1,X_2=x_2)}{\Pr(Y=0|X_1=1,X_2=x_2)}\right\}\times\left\{\frac{\Pr(Y=0|X_1=0,X_2=x_2)}{\Pr(Y=1|X_1=0,X_2=x_2)}\right\},
\end{aligned}
\tag{A.2}
$$

in other words the odds ratio in the population, as the probability of a complete record depends on Y and X_2, so $\Pr(R=1|Y=y,X_1=x,X_2=x_2)=\Pr(R=1|Y=y,X_2=x_2)$.

This is simply a version of the same argument that justifies the use of logistic regression for case-control studies; there selection depends on case/control status (Y), but not on exposure (X), and so the estimate of the odds ratio relating exposure to outcome is valid. The validity of complete records in logistic regression is explored in more detail by Bartlett et al. (2015a), using simulations and an example.