Exploring methods for transforming the c-statistic of a prediction model developed and validated on multiply imputed data

Cem Kalender
Student number: 2920734

Supervisors: Dr V.M.T. de Jong, H.I. Oberman, Dr G. Vink

Programme: master Applied Data Science, Utrecht University

July 1, 2022

Abstract

Prediction models play a crucial role in the clinical decision-making process, as they help physicians to estimate the risk of disease presence (diagnosis) or occurrence of a future event (prognosis). It is of utmost importance that prediction models are validated prior to clinical implementation. Because the data used for developing (and validating) prediction models are often replete with missing entries, the multiple imputation procedure (MI) can be used to account for missing data. However, interpretation of the discriminatory performance (c-statistic) of a prediction model built on multiply imputed data is complicated, as the c-statistic exhibits non-normality. This model-based simulation study explores transformation methods for c-statistic estimates prior to pooling across imputed data sets, and assesses methods for their ability to generate accurate pooled estimates in terms of bias and coverage. Candidate methods include standard pooling (using Rubin's rules), logit transformation prior to pooling, and arcsine square-root transformation prior to pooling. The findings show that when missing data are present and when θ_s is < 0.9, none of the investigated methods attained nominal coverage. However, both regular pooling and logit transforming estimates prior to pooling yield similarly satisfactory results in terms of coverage and bias if the true value for the c-statistic is $>= 0.9$. As a tentative recommendation, the logit method should be used due to the method's higher efficiency in terms of coverage.

1 Background

Prediction models play a crucial role in the clinical decision-making process, as they aid physicians with estimating a risk of disease presence (diagnosis) or occurrence of a future event (prognosis) based on a combination of patient characteristics (Vergouwe et al., 2010). Prediction models are constructed using data from subjects from a development set and ought to be validated to assess generalizability to novel groups of similar subjects.

Missing data is ubiquitous in clinical data sets and it impinges on the quality of the data needed to build prediction models. If unaccounted for, underlying systemic causes for missingness may distort statistical relationships (van Buuren, 2018), which may impact the performance of a prediction model. As such, improper handling of missing data in model development or validation can substantially undermine clinical decision-making, as it may culminate in biased inferences pertaining to the likelihood of the presence of a particular disease or other health status and/or its prognosis.

An often used technique for accommodating missing data is multiple imputation (MI), which involves generating multiply imputed data sets. The imputed values are calculated in such a manner that they vary slightly across data sets, as this emulates noise and reflects the uncertainty inherent in estimating an imputation value (van Buuren, 2018). Miles (2016) demonstrates how predictions can be obtained from models fit to multiply imputed data sets by first being estimated separately in each imputed data set, and then combined using Rubin's rules. However, Rubin's rules are based on asymptotic theory (Rubin, 2004), implying that inferences of about a population parameter (Q) are based on the normal approximation of $Q - \hat{Q} \sim N(0, U)$, where $\hat{Q}$ is the estimate of Q and U the variance for Q.

Although the normality condition is usually met for model coefficients (Enders, 2010; 220-221), it may be problematic for combining model performance measures such as the concordance (c-) statistic (AUC), which quantifies a model's discriminatory power (Hanley McNeil, 1982). It's distribution exhibits non-normality and asymmetry, with values bound between 0-1 and 0.5-1 in validation and development, respectively. Given these characteristics, undue credence to the normality assumption through the use of Rubin's rules for combining the c-statistic estimates may yield an inaccurate overall estimate of the performance of a prediction model developed using multiply imputed data.

While there is limited research on pooling methods for performance measures specifically, several studies on the distribution of the c-statistic provide grounds to suggest that specific transformations can countervail non-normality. Conceivably, such transformations may thus be useful prior to pooling. Snell et al. (2018) recommend logit transforming the c-statistic as an approximation of the normal distribution in the context of meta-analysis. Furthermore, Trikalinos et al. (2013) propose using "variance stabilizing transformations" such as the arcsine square-root transformations when meta-analyzing proportions and rates, respectively. On the other hand, in circumstances where transformations cannot be identified, alternative robust summary measures such as medians and ranges are proposed (Marshall et al., 2009). Yet, there appears to be little progress in the matter as current practice for combining performance measures is to pool without applying any transformations (Marshall et al., 2009).

Deploying the practical guideline of Vergouwe et al. (2010) for developing and validating a prediction model with missing data, this model-based simulation study explores transformation methods for the c-statistic prior to pooling. I appraise the methods for their ability to generate accurate estimates in terms of bias and coverage. Importantly, findings could ultimately help improve clinical decision-making because physicians can decide whether the estimated discriminatory performance of a model is sufficient in light of the context of clinical application. As such, the research question constitutes:

Which transformation method of the c-statistic supports unbiased inference of the discriminatory power of a prediction model developed using multiply imputed data sets?

In addition, this study aims to gain insight in how the transformations affect the distribution of the c-statistic estimates as well as the pooled standard errors, leading to the following exploratory research question:

How do the transformations compare with respect to the distribution of the c-statistic estimates and corresponding pooled standard errors?

This study is designed as follows. First, the methods section discusses the simulation set-up and provides an outline of the execution of the study (section 2). Thereafter, the results are presented (section 3), interpreted, and implications are discussed (section 4). Finally, concluding remarks as well as recommendations for further research are made (section 5).

2 Methods

2.1 Transformation methods for the c-statistic

Following Snell et al. (2018) and Marshall et al. (2009) I evaluate the logit and arcsine square-root transformations. As evident from figure 1, both transformations are essentially linear over the range of 0.3–0.7, with more curvature near the ends. However, the curvature of the logit transformation is much more pronounced compared to the arcsine transformation, implying that small differences on extreme c-statistics are more amplified on the logit scale.

2.2 Data-generating mechanism and simulation study

I use a parametric model for data generation, as it provides the researcher more control (Morris, White & Crowther, 2019): data can be generated based on specific variance-covariance matrices and data can be carefully amputed according to specific systematic causes for missingness. Given the disparate effects of both transformations near the ends of the distribution (Figure 1), I explore transformations supporting the application of Rubin's rules for pooling for 3 different ”true” c-statistic values. This methodological choice inherently reflects 3 different sub-studies (s), each of which comprises $R = 1000$ simulation rounds.

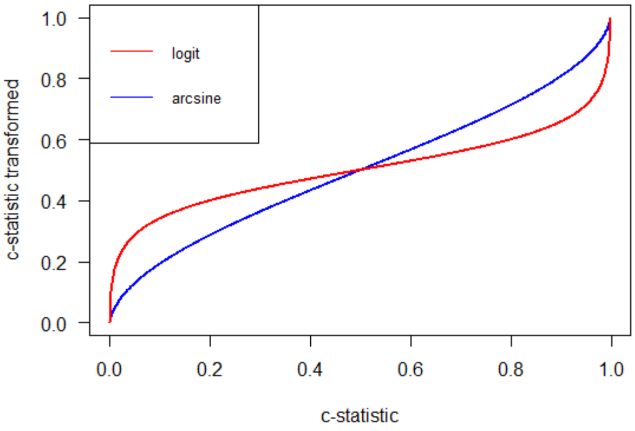


Figure 1: Logit and arcsine transformations of a proportion

Depending on s, the "true" c-statistic (θ_s) is approximately equal to 0.7, 0.8, or 0.9, respectively.

The following outlines the procedural steps of a single simulation round r within a single sub-study s, which is composed of 5 stages. Consider figure 2 for a schematic overview of this simulation study.

2.2.1 Stage 1. A multivariate, complete data set with a binary outcome variable is simulated

A multivariate, complete data set of n = 5,000 subjects is simulated. Let i, $i = 1, \ldots, n$, denote subject i. The complete data set contains two normally distributed variables $X_1 \sim$ N(10, 12) and $X_2 \sim$ N(6, 5), and a binary outcome variable $Y = (y_1, y_2, ..., y_n)$. The binary outcome is created by inverse logit transforming a linear combination of an intercept (β_0) and coefficients (β_1 and β_2):

$$p_i = \frac{e^{(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}}{1 + e^{(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i})}} \tag{1}$$

where X_i is the predictor value for subject i, $\beta_0 = 5$, $\beta_1 =$ -1, and $\beta_2 =$ -1. p_i is the probability of having a positive outcome (= 1) for subject i. Outcomes are assigned using a Bernoulli trial ($Y_i \sim B(p_i)$).

Additionally, a noise parameter (ψ) is used to introduce noise during data simulation. ψ constitutes a randomly sampled fraction of positively labeled observations that are transmuted into negative cases and vice versa. The usage of ψ allows for alternating θ_s depending on the the sub-study, with $\psi =$

$(0.25, 0.125, 0.05)$ corresponding to $\theta_s \approx 0.70$, 0.80 and 0.90, respectively to the third decimal place. To retrieve θ_s, the complete set is split into equal parts, Z_1 and Z_2. Data set Z_1 is used to develop a logistic regression model, which is validated on set Z_2.

2.2.2 Stage 2. The complete data set is amputed and split into incomplete versions of Z_1 and Z_2

Vergouwe et al. (2010) provide a practical guideline for developing and validating a prediction model with missing data. They recommend splitting the incomplete data into equally sized parts (a development and validation set), and performing imputation separately on both sets. This keeps the validation set blind to the outcome-covariate relationship in the development set (Wahl et al., 2016). This approach is reproduced here, meaning that the original complete data set generated in stage 1 is first made incomplete (amputated) and split again into incomplete "versions" of Z_1 and Z_2, rather than amputing Z_1 and Z_2 separately after calculating θ_s.

Approximately 30% of the covariate values are deleted conforming to an Missing At Random (MAR) mechanism using the `mice` R package. With MAR missingness, the missing values are random conditional on available information, meaning that the information about the missing data is in the observed data (Janssen, et al., 2010). Two following missingness patterns are implemented (simultaneously) to instantiate a situation of MAR. To exemplify, consider again variables X_1 and X_2 and outcome Y:

- X_1 is missing, conditional on X_2
- Y is missing, conditional on X_1

This particular combination of missingness patterns is ubiquitous in the clinical domain. The former presupposes a situation in which patients who are relatively healthier (as evidenced from X_2) might be less likely to undergo invasive tests, thus leading to missing values on those predictors (Groenwold et al., 2012). The latter implies a common clinical situation in which missing values on the outcome arise due to a loss to follow-up, which may be attributable to a patient characteristic. For instance, older people may be more likely to remain a participant in a cohort study than adolescents, culminating in more missing outcome values for adolescents. Within `mice`, the missingness patterns are specified as a pattern matrix, in which a pattern (k) is represented as a row and the variable on which k exerts influence as a column v (0 = missing, 1 = non-missing).

After defining k, values are deleted using a weight matrix. In order to appropriately implement a MAR mechanism, the weights of the variables that are made incomplete are set to zero, whereas variables causing missingness have a non-zero value (Van Buuren & Groothuis-Oudshoorn, 2011). Essentially, the values of the weight matrix are the coefficients of a linear regression equation of which the outcome is a weighted sum score. Suppose, subject i is a candidate for pattern k. Then, the weighted sum score (wss) is:

$$wss_i = w_{k,1} * X_{1,i} + w_{k,2} * X_{2,i} + ... + w_{k,v} * X_{v,i}, \tag{2}$$

where $\{X_{1,i}, X_{2,i}, ..., X_{v,i}\}$ is the set of variable values for subject i and $\{w_{k,1}, w_{k,2}, ..., w_{k,v}\}$ are the values on row (pattern) k of the weights matrix. Because I simulate 3 variables and 2 missing data patterns, $v = 3$ and $k \in \{1, 2\}$, respectively. I use matrix $\left(\begin{smallmatrix}0&5&0\\5&0&0\end{smallmatrix}\right)$ to assign weights to k. Furthermore, the proportion of missingness caused by the 2 missingness patterns is equal (i.e. 50-50%).

To verify that amputation has occurred correctly (i.e. according to MAR), the indicator method is applied once before execution of the full simulation study. The indicator method involves creating indicator variables for each variable with missing entries (e.g., 1 = missing, 0 = not missing). Logistic regression models are fit with the indicator variables as outcome and all the original variables as predictors. A statistically significant effect on the indicator variable presupposes that missingness is explained by the corresponding variable. Based on this, it was established that the amputation procedure occurred correctly.

2.2.3 Stage 3. The incomplete data sets are imputed by means of Bayesian linear regression and logistic regression imputation models

Several simulation studies have demonstrated that repeated imputations (J) can be as low as three for data with 20% missing entries (Van Buuren, Boshuizen & Knook, 1999). According to Vergouwe at al. (2010), there is little benefit in increasing J beyond 10 imputations. In view of this, the choice is made to set J to 10, thus giving 10 imputed development and validation sets.

Multiple imputation is implemented with the Multivariate Imputation by Chained Equations (MICE) procedure in R (Van Buuren & Groothuis-Oudshoorn, 2011). The general methodology suggested for imputation is to impute using the posterior predictive distribution of the missing data given the observed data and some estimate of the parameters (Liu, 2016). In light of this, MICE was done with Bayesian linear regression imputation for the missing predictors. With this, imputation uncertainty is accounted for by adding extra error variance to the predicted values from the linear regression model (Van Buuren, 2018). In addition, the uncertainty in estimating the regression coefficients of the imputation model is taken into account (Van Buuren, 2018). Missing outcome values were imputed using a logistic regression imputation model.

Appropriate imputation necessitates convergence of the MICE algorithm. A novel approach for diagnosis of (non-)convergence proposed by Oberman, van Buuren & Vink (2021) is to evaluate the potential scale reduction factor per iteration of the MICE algorithm. Convergence is achieved if this parameter does not improve over iterations. According to Oberman, van Buuren & Vink (2021), inferential validity is already achieved after 5 to 10 iterations. Adhering to this approach, a check for convergence is performed prior to the execution of the full simulation study. There were no cases of non-convergence of the MICE algorithm.

2.2.4 Stage 4. Develop models on the imputed data sets, pool the models using Rubin's rules

Following the imputation procedure, a logistic regression model is developed in each of the imputed development sets, thus yielding 10 models. The fitted models can be written as:

$$\hat{p}_{i,j} = \frac{e^{(\hat{\beta}_{0,j}+\hat{\beta}_{1,j}X_{1i,j}+\hat{\beta}_{2,j}X_{2i,j})}}{1+e^{(\hat{\beta}_{0,j}+\hat{\beta}_{1,j}X_{1i,j}+\hat{\beta}_{2,j}X_{2i,j})}}, \tag{3}$$

where $\hat{p}_{i,j}$ is the probability of a positive outcome for subject i by a prediction model built on the jth ($j = 1, \ldots, J$) imputed development set. $X_{i,j}$ is the predictor value for subject i in imputed development set j, and $\hat{\beta}_{0,j}$, $\hat{\beta}_{1,j}$, $\hat{\beta}_{2,j}$ are the estimates of β_0, β_1, β_2 in a prediction model built on the the jth imputed development set. To derive a single prediction model, the coefficient vectors and standard error vectors of the models need to be combined.

The computation of multiple imputation estimates and variances follows Rubin's rules (Rubin, 1987). Let $\hat{Q}_j$ and W_j denote an estimate (of a parameter) and it's variance, respectively, from a model developed on the jth imputed development set. The multiple imputation estimate ($\bar{Q}$) is calculated as:

$$\bar{Q} = \frac{1}{J}\sum_{j=1}^{J}\hat{Q}_j \tag{4}$$

The total error variance T of $\bar{Q}$ is obtained by a components-of-variance argument, giving:

$$T = W + (1+\frac{1}{J})B, \tag{5}$$

where W = within-imputation variance:

$$\frac{1}{J}\sum_{j=1}^{J} = W_j, \tag{6}$$

and where B = between-imputation variance:

$$\frac{1}{J-1}\sum_{j=1}^{J} = (\hat{Q}_j - \bar{Q})^2 \tag{7}$$

2.2.5 Stage 5. Estimate θ_s on j imputed validation sets. Apply transformations and pool $\hat{\theta}_s$

After deriving the pooled model, predictions are made on each of the imputed validation sets. Subsequently, 10 c-statistic estimates are calculated, which are computed as defined by DeLong et al. (1988) using the algorithm by Sun and Xu (2014), which is implemented in the `pROC` R package (Robin et al., 2021). The corresponding standard errors are computed by using the `auctestr` R package, which leverages the property that the c-statistic is equivalent to the Wilcoxon or Mann-Whitney U test statistic (Mason & Graham, 2002). The between-imputation variance simply constitutes the variance of $\hat{\theta}_s$ across the imputed validation sets.

The estimates and corresponding variances are pooled with three methods. The first method is to pool using Rubin's rules, without applying any transformation to the estimates and variances beforehand. The second approach is to logit transform the estimates using the `metamsic` R package (Debray & De Jong, 2021) prior to pooling. The third method uses an arcsine square-root transformation before pooling.

The 95% CIs for the transformed estimates are calculated on the respective pooled scale and then back-transformed. The 95% CIs of the pooled estimates are obtained using the total error variance (pooled SE) and a reference t-distribution with degrees of freedom:

$$\nu = (J - 1) * (1 + \frac{1}{r})^2, \tag{8}$$

where $r = (\frac{b + b/J}{w})$ denotes the relative increase in variance due to the missing values, with w and b denoting the within and between-variance of the estimates pooled into $\hat{\theta}_s$ at simulation round r, respectively. The CI of $\hat{\theta}_s$ is then of the form:

$$\hat{\theta}_s \pm t_v * se(\hat{\theta}_s), \tag{9}$$

where $se(\hat{\theta}_s)$ denotes the pooled SE of $\hat{\theta}_s$ and t_ν is the $\alpha/2$ quantile of the t-distribution with ν degrees of freedom.

2.2.6 Evaluate bias and coverage of the pooling methods

To juxtapose transformations prior to pooling with the standard method, I utilize two performance measures based on all R simulation rounds (remember, $R = 1000$). First, performance is determined by quantifying bias, indicating the amount by which the estimates exceed θ_s on average as quantified by the positive or negative mean difference:

$$Bias = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}_{s,r} - \theta_s), \tag{10}$$

where $\hat{\theta}_{s,r}$ denotes the pooled and back-transformed estimate of θ_s for simulation round r. The corresponding error of the bias is quantified by computing the Monte Carlo standard error (SE). The Monte Carlo SE quantifies simulation uncertainty as it provides an estimate of the SE of (estimated) performance due to using a finite simulated sample size (Morris, White & Crowther, 2019). It is calculated as:

$$MonteCarloSE(Bias) = \sqrt{\frac{1}{R(R-1)}\sum_{r=1}^{R}(\hat{\theta}_{s,r} - \bar{\theta}_s)^2}, \tag{11}$$

where $\bar{\theta}_s = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_{s,r})$. Furthermore, the performance of the methods are assessed in terms of coverage, which is determined by estimating the percentage of the 95% CIs that include θ_s over all simulation runs. The corresponding Monte Carlo SE of the coverage is calculated for every pooling method as:

$$MonteCarloSE(Coverage) = \sqrt{\frac{\widehat{cover}. * (1 - \widehat{cover}.)}{R}}, \tag{12}$$

where $\widehat{cover}.$ denotes the estimated coverage by a pooling method.

Finally, the (relative) impact of transformations may be apparent from the magnitude of the total error variance and how it compares to the pooled estimate. Therefore, methods are also assessed on how the relationship between the pooled estimate and the total error variance manifests, by means of a meta-regression stratified by sub-study and method.

The code implemented for this simulation study can be found on: https://github.com/Cem-Kalender/Exploring$_c - statistic_transformations$

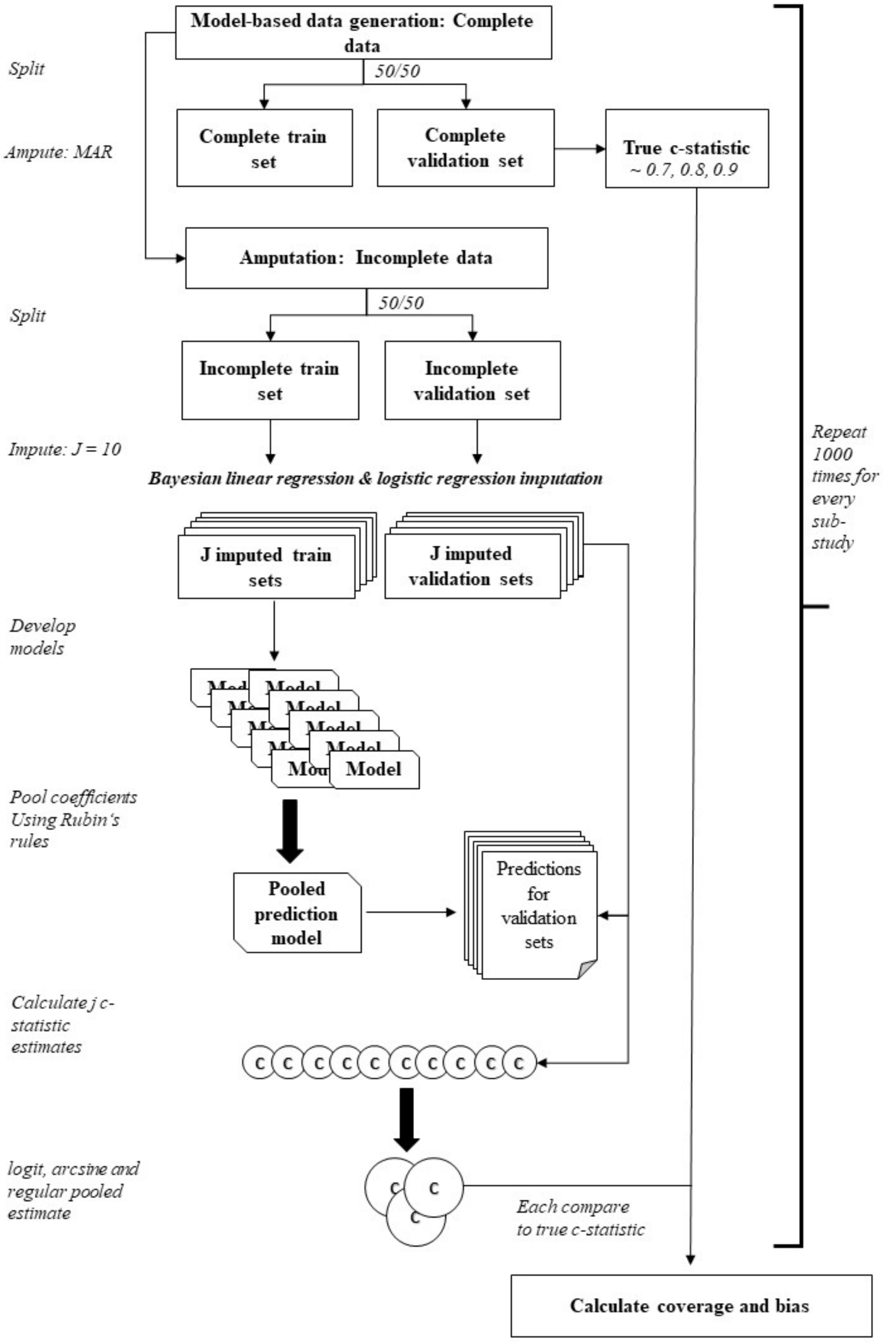

Model-based data generation: Complete data
Split
50/50
Complete train set
Complete validation set
True c-statistic
~ 0.7, 0.8, 0.9
Ampute: MAR
Amputation: Incomplete data
Split
50/50
Incomplete train set
Incomplete validation set
Impute: J = 10
Bayesian linear regression & logistic regression imputation
J imputed train sets
J imputed validation sets
Develop models
Model
Pool coefficients Using Rubin's rules
Pooled prediction model
Predictions for validation sets
Calculate j c-statistic estimates
c
logit, arcsine and regular pooled estimate
Each compare to true c-statistic
Repeat 1000 times for every sub-study
Calculate coverage and bias

Figure 2: Schematic overview of the simulation study

3 Results

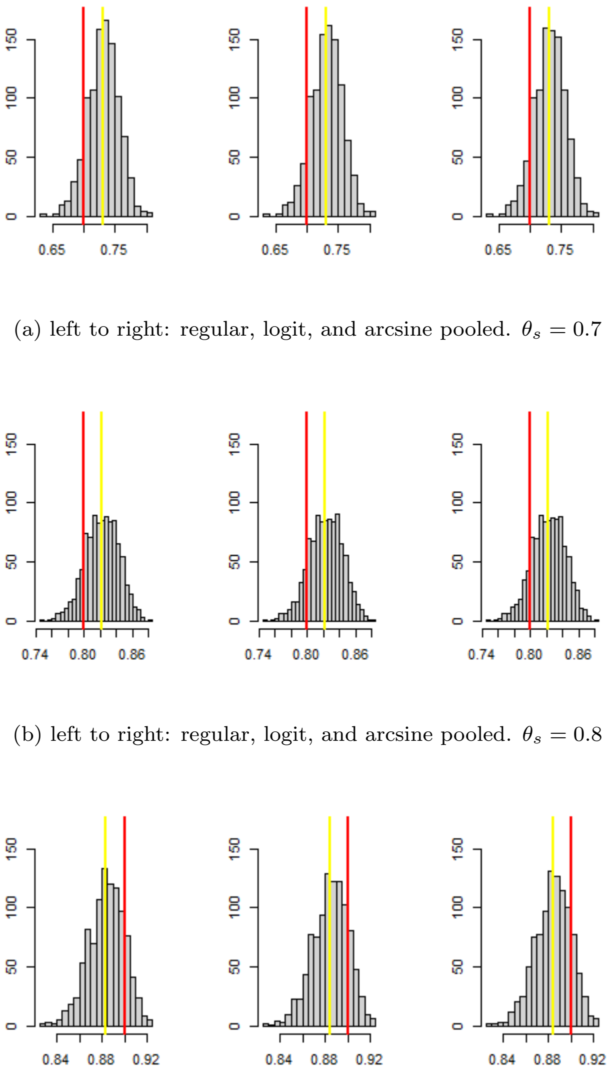
All simulation runs completed and there were no convergence issues. The results (Table 1) show that $\bar{\theta}_s$ is approximately equal between the methods (0.73, 0.82, 0.88 for $\theta_s = 0.7, 0.8, 0.9$, respectively). Furthermore, the similarly between the methods is also evident from the distributions of $\hat{\theta}_s$ (Figure 3).

Performance measure	Pooling method	$\theta_s = 0.7$	$\theta_s = 0.8$	$\theta_s = 0.9$
$\bar{\theta}_s$ (SE)	Regular	0.73 (0.00)	0.82 (0.00)	0.88 (0.00)
	Logit	0.73 (0.00)	0.82 (0.00)	0.88 (0.00)
	Arcsine	0.73 (0.00)	0.82 (0.00)	0.88 (0.00)
Bias (SE)	Regular	0.03 (0.01)	0.025 (0.01)	-0.01 (0.01)
	Logit	0.03 (0.01)	0.025 (0.01)	-0.01 (0.01)
	Arcsine	0.03 (0.01)	0.025 (0.01)	-0.01 (0.01)
Coverage (SE)	Regular	80.40% (1.25%)	82.20% (1.29%)	96.10% (0.61%)
	Logit	83.60% (1.72%)	87.10% (1.06%)	94.30% (0.73%)
	Arcsine	100% (0.00%)	100% (0.00%)	100% (0.00%)
95% CI width	Regular	0.12	0.10	0.08
	Logit	0.12	0.10	0.08
	Arcsine	0.36	0.30	0.24
Skewness	Regular	-0.23	-0.19	-0.34
	Logit	-0.23	-0.19	-0.34
	Arcsine	-0.23	-0.19	-0.34
Kurtosis	Regular	3.33	2.90	3.07
	Logit	3.27	2.92	3.09
	Arcsine	3.33	2.91	3.09

Table 1: The average of the pooled estimates of θ_s ($\bar{\theta}_s$), bias (SE) for pooling methods per sub-study, Coverage percentage (SE) of the 95% CI, average 95% CI width, skewness and kurtosis of the distribution of $\hat{\theta}_s$.

3.1 Distribution of (transformed and) pooled estimates

The distributions look fairly normal at a quick glance (Figure 3). However, the skewness and kurtosis (Table 1) indicate that for all methods, the skewness starts at -0.23 at $\theta_s = 0.7$, drops slightly to -0.19 at $\theta_s = 0.8$, only to increase to -0.34 at $\theta_s = 0.9$. In contrast, the kurtosis remains relatively constant as θ_s increases, only showing a slight drop when $\theta_s = 0.8$. There are no profound differences between the distributions with respect to the kurtosis, indicating relative overall consistency in the "heaviness" of the tails.



(a) left to right: regular, logit, and arcsine pooled. $\theta_s = 0.7$

(b) left to right: regular, logit, and arcsine pooled. $\theta_s = 0.8$

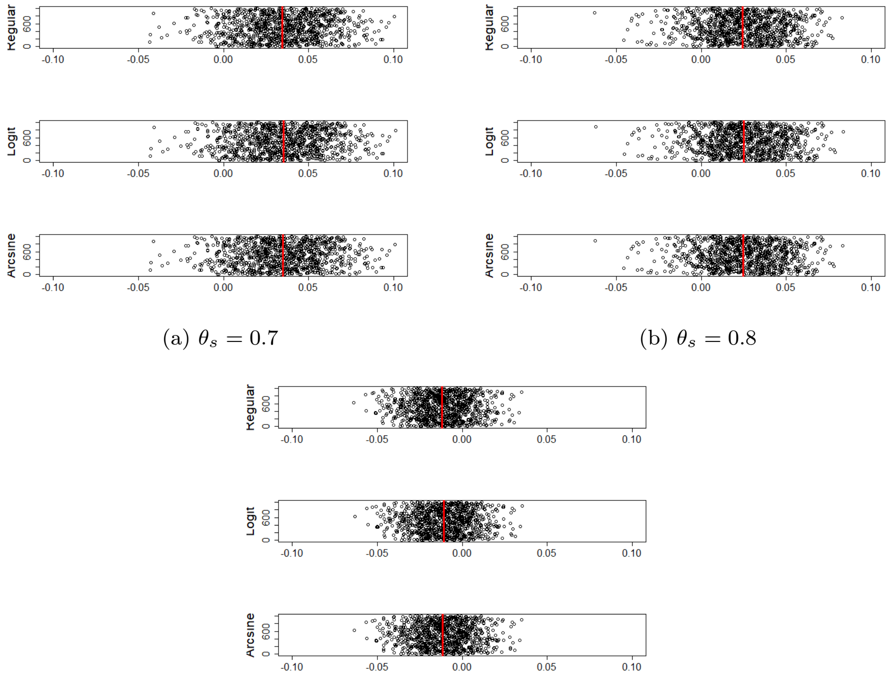
(c) left to right: regular, logit, and arcsine pooled. $\theta_s = 0.9$

Figure 3: Histograms showing the distribution of $\hat{\theta}_s$ for different pooling methods and values of θ_s. The yellow lines are the means of $\hat{\theta}_s$. The red lines show θ_s. The difference indicates the bias.

3.2 Bias of the pooling methods

Inspect table 1 for the bias of the methods as well as the corresponding Monte Carlo errors. There are practically no differences between the methods as they overestimate θ_s in equal rates when θ_s is 0.7 and 0.8: the bias is around 0.03 (0.01) for all methods when $\theta_s = 0.7$ and around 0.025 (0.01) when $\theta_s = 0.8$. When $\theta_s = 0.9$, the bias decreases and becomes practically zero.

Figure 3 visualizes the differences between $\hat{\theta}_{s,r}$ and θ_s per method. As evident from the narrower spread of points, the differences between $\hat{\theta}_{s,r}$ and θ_s become smaller as θ_s increases. This is applicable to all methods, with virtually no differences in terms of the distance between $\hat{\theta}_{s,r}$ and θ_s.



(a) $\theta_s = 0.7$

(b) $\theta_s = 0.8$

(c) $\theta_s = 0.9$

Figure 4: The difference between θ_s and $\hat{\theta}_s$ for different pooling methods. The red lines indicate the mean difference (bias) per method. The y-axis provides separation between the points.

3.3 Coverage of the pooling methods

At a quick glance, the results appear to indicate that the arcsine pooling method is the optimal approach in terms of coverage of the CI, while the other methods deliver satisfactory performance (Table 1). However, coverage has to be appraised in light of the width of the 95% CI per method to determine applicability and utility within a clinical context. Table 1 shows the average width of the 95% CI per method, by taking the mean difference between the lower and upper bounds of the 95%CI. When the 95% CI width is viewed in conjunction with the bias per method, it becomes evident that the arcsine method yields confidence intervals that could be deemed undesirable within a clinical or statistical context. For instance, note, that the average width of the 95% CI of the arcsine method at $\theta_s = 0.7$ is 0.36. Given that the same method yields a bias of 0.03166 at $\theta_s = 0.7$, the interval spans 0.55 and 0.91 ($0.73166 \pm 1/2 * 0.36$) on average. With the regular and logit methods, the average width of the 95% CI is much smaller (0.12), giving a span between 0.676 and 0.79 on average at θ_s = 0.7. The coverage of the logit method at a θ_s of 0.7 and 0.8 is 83.6% (1.72%) and 87.1% (1.06%), respectively. At the same values for θ_s, the coverage for the regular method is 80.4% (1.25%) and 82.2% (1.29%), respectively. Still, while the logit method outperforms the regular method when $\theta_s = 0.7$ and θ_s = 0.8, it is slightly outperformed by 1.8% when $\theta_s = 0.9$. Nevertheless, the logit method appears to be more efficient in a statistical sense, meaning that the rate at which the SE or CI becomes smaller as R approaches ∞ is higher for the logit method as compared to the regular method. This is evident from a lower difference in coverage between the lowest and highest value for θ_s (10.7%), whereas this difference is 15.7% for the regular pooling method.

3.4 Relationship between pooled estimates and pooled SEs

The relationship between $\hat{\theta}_s$ and the total error variance, $se(\hat{\theta}_s)$, is shown in figure 5. For different values of θ_s, both the regular and arcsine transformation method exhibit a strong negative Pearson's correlation (ρ), with the latter method showing the strongest ρ of the two regardless of the value of θ_s. In addition, the relationship between arcsine-transformed $\hat{\theta}_s$ and $se(\hat{\theta}_s)$ shows clear signs of heteroskedasticity. Conversely, $se(\hat{\theta}_s)$ appears to be weakly positively correlated with logit-transformed $\hat{\theta}_s$. As can be seen, when θ_s increases, ρ becomes significantly stronger for every method.

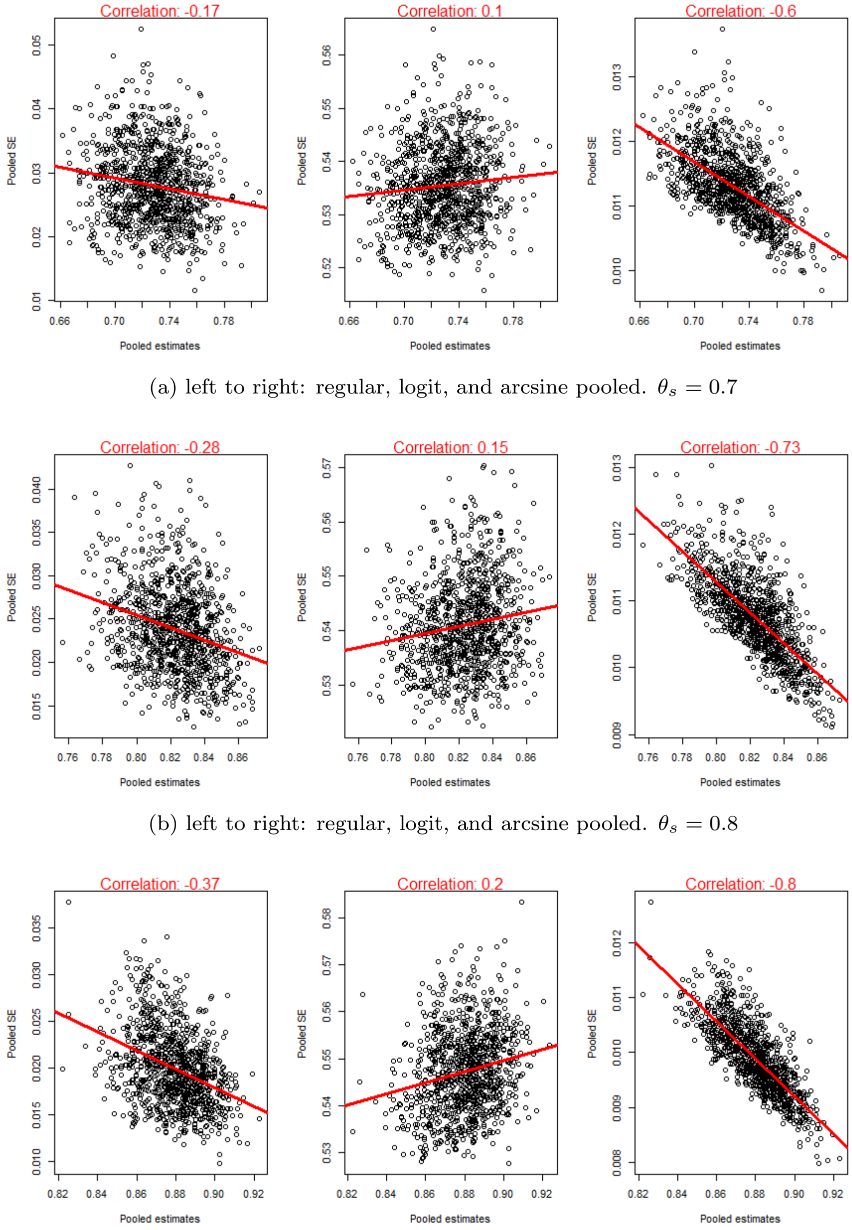
However, it should be noted that, despite these correlations, an increase in the pooled estimate is still associated with only a fractional decrease (increase in case of the logit method) in the pooled SE. Consider the results of a meta-regression stratified by sub-study and pooling method (Table 2). Despite a correlation of -0.37 for the regular pooled estimates at $\theta_s = 0.9$, the standardized difference in pooled SEs for $\hat{\theta}_s = 0.82$ and $\hat{\theta}_s = 0.88$ is a mere 0.02 (0.88*-0.37 - 0.82*-0.37). The standardized difference in arcsine-pooled SEs for the same pooled estimates constitutes a mere 0.047 (0.88*0.798 - 0.82*0.798), whereas for the logit-pooled SEs this difference is even more negligible (0.88*0.202 -

0.82*0.202 = 0.012). Furthermore, the meta-regression reveals that the variability observed in logit-pooled SE's are not well explained by the stratified regression model regardless of the value of θ_s (with $R^2 = 0.016, 0.027, 0.033$ for $\theta_s = 0.7$, 0.8, 0.9, respectively). On the other hand, the explained variance is noticeably high for the arcsine method and even increases as θ_s increases (with $R^2 = 0.394, 0.552, 0.681$). The regular pooling method sits in between these extremes with explained variances of 3.2, 9.0 and 18.6%, respectively.

Sub-study	Method	Estimate	SE	R^2
$\theta_s = 0.7$	Regular	-0.169***	0.031	0.032
	Logit	0.098*	0.032	0.016
	Arcsine	-0.596***	0.025	0.393
$\theta_s = 0.8$	Regular	-0.285***	0.030	0.090
	Logit	0.152***	0.031	0.027
	Arcsine	-0.733***	0.022	0.552
$\theta_s = 0.9$	Regular	-0.374***	0.029	0.186
	Logit	0.202***	0.031	0.033
	Arcsine	-0.798***	0.019	0.681

Note: ***p<0.001, **p<0.01, *p<0.05.

Table 2: The results of a meta-regression stratified by sub-study and pooling method, where $se(\hat{\theta}_s)$ are regressed on $\hat{\theta}_s$. Both $\hat{\theta}_s$ and $se(\hat{\theta}_s)$ have been standardized to allow comparison between beta's



(a) left to right: regular, logit, and arcsine pooled. $\theta_s = 0.7$

(b) left to right: regular, logit, and arcsine pooled. $\theta_s = 0.8$

(c) left to right: regular, logit, and arcsine pooled. $\theta_s = 0.9$

Figure 5: Scatter plots showing the correlation between pooled estimates $(\hat{\theta}_s)$ and errors $(se(\hat{\theta}_s))$. The red line is the regression line.

4 Discussion

4.1 Main findings and implications

Following the recommendations of Vergouwe et al. (2010) for developing and validating a prediction model built using multiply imputed data, a simulation study was conducted to evaluate transformation methods for the c-statistic in terms of bias and coverage. The main differences between the investigated methods reside in the coverage and the width of the 95% CIs of the c-statistic estimates, while they are practically identical with respect to bias. The main finding of this study is that, when missing data are present and when θ_s is $<=$ 0.9, none of the investigated methods attained nominal coverage. The arcsine method displays an actual coverage probability that is greater than the nominal coverage probability, regardless of the value of θ_s. On the other hand, the regular and the logit method are able to achieve nominal coverage only when θ_s is $>=$ 0.9. In this case, a strong improvement in bias at this value for θ_s might be a reason for the attainment of nominal coverage. To increase coverage at lower values for θ_s wider confidence intervals are needed. In this simulation study, the uncertainty inherent in the development sets (i.e. standard errors of the coefficients of the pooled model) are not propagated to the estimates made on the validation sets, thus resulting in narrower 95% CIs of the pooled estimates for lower values of θ_s. Despite of this drawback, the logit method clearly outperforms the regular method when $\theta_s < 0.9$, and is more efficient in terms of coverage. However, this in itself can hardly be seen as irrevocable evidence of the superiority of the logit method.

With respect to the average width of the 95% CI, the regular and logit methods are on par with each other. However, the findings show that calculating the confidence interval using arcsine-transformed pooled estimates is problematic. This is apparent from the abnormally wide average width of the 95% CI, culminating in a coverage of 100% for all values of θ_s. This latter finding must be interpreted in light of the claims made by Wilson et al. (2013) that arcsine-transformation of proportional data might give rise to nonsensical values if extrapolated. This problem arises from the fact that an arcsine square-root transformation essentially normalizes values to fall between 0 and π (Wilson & Hardy, 2002), whereas extrapolation requires transformed values to be monotonic to avoid the possibility of nonsensical values (Warton & Hui, 2011). The arcsine transformation does not satisfy this requirement due to the sine function's periodicity but the logit transformation does (Warton & Hui, 2011). Conceivably, the non-monotonicity caused by the arcsine transformation might be the reason that arithmetic operations yield nonsensical results (e.g. Rubin's rules for pooling, calculating the 95% CIs). Clearly, this is a highly undesirable trait of the arcsine transformation as more precision is typically expected within a clinical context given that lives may be at stake.

As an additional exploratory research question, this study aimed to explore how transformations affect the distribution of the pooled estimates as well the pooled standard errors. The findings indicate virtually no differences with respect to the distribution of the pooled estimates. Rather, the estimates display normality regardless of the pooling method implemented. Several speculations are in order as to the cause of this finding. According to Snell et al. (2018), who conducted a meta-analysis to summarise performance measures, non-normality of the c-statistic is caused by extreme scenarios of predictor effect heterogeneity and is best accounted for through logit transforming estimates. This is in line with Austin & Steyerberg (2012), who showed that the c-statistic depends on both the log odds ratio and variance of a continuous explanatory variable. In contrast, in this study, little predictor effect heterogeneity is expected given that every pooled estimate is essentially derived from the same parametric data, with differences manifesting mainly as consequence of the implemented imputation procedure. This could thus be a reason for the apparent normality.

Alternatively, the imputation methods implemented in this study could also be a reason for the apparent normality of the pooled estimates. This simulation study used a Bayesian linear regression imputation method for missing continuous values and a logistic regression imputation method for missing outcome values, both of which are parametric methods. As this study simulated parametric data, the distributional assumptions of the Bayesian linear regression imputation method were met. Additionally, the assumption of congenial imputation models to account for the MAR missingness mechanism was met (Meng, 1994; van Buuren, 2018). As a result, the parametric imputation methods were well able to impute the missing data (van Buuren, 2018: Addo, 2018). This may in turn have yielded c-statistic estimates with minimal variance, thereby possibly affecting the distribution of the estimates to a degree that logit transformations are of minimal use. Therefore, I recommend future research to evaluate transformations for the c-statistic while also experimenting with a range of different imputation methods.

Finally, the findings suggest that the relationship between the pooled estimates and pooled standard errors manifest differently for pooling methods. In case of regular and arcsine pooled estimates, the findings show that the pooled SE becomes smaller as the pooled estimate gets larger (positively biased) whereas the opposite is true for the logit-pooled estimates. It is beyond the scope of this study to go into detail as to how this occurred. At the very least, it can be suspected that an interaction between the transformations and Rubin's rules for pooling is the cause. Regardless, the effect of the correlations are inconsequential as they do not translate to the performance measures to any meaningful extent.

4.2 Limitations and future research

The methodological choices of this study limit the generalizability of the findings to cases in which predictor data is normally distributed and missingness in exclusively MAR. When data is missing based on different systematic causes or when predictor data is non-normally distributed, results might deviate substantially. Within clinical settings, a variety of systemic causes of missingness impinge on data quality. Some data are inherently missing not at random (MNAR) and it is by definition not possible to account for systematic differences between the missing values and the observed values using solely the observed data (Sterne et al., 2009). While there exist methods that leverage the measured data to correct for MNAR, these make strong assumptions. In such cases, multiple imputation on its own is not sufficient nor able to accommodate for missing data appropriately, implying that validation can be subject to substantial bias and therefore can result in misleading conclusions regarding the performance of a prediction model built using multiply imputed data. Nonetheless, it may still be better to use MI with these assumptions than to ignore missing data. Another constraint shared with Vergouwe et al. (2018), is that this study did not consider internal validation during model development. Wahl et al. (2016) research several internal validation strategies and recommend deploying internal validation followed by MI on the training and test set separately, in order to correct for optimism when estimating performance. As such, for future research it is recommended to replicate the approach of Wahl et al. (2016) and appraise the regular and logit transformations on the same performance measures evaluated here.

5 Conclusion

In conclusion, this simulation study has shown that when missing data are present and when θ_s is < 0.9, none of the investigated transformation methods attained nominal coverage. However, both regular pooling and logit-transforming estimates prior to pooling yield satisfactory results in terms of coverage and bias if θ_s is $>= 0.9$.

A practical implication of the findings, then, is that additional scrutiny is warranted when c-statistic estimates (of a clinical prediction model built using multiply imputed data) fall around 0.7-0.8, whereas estimates around or above 0.9 can be trusted. However, such high values for discriminatory performance are rare in practice. In light of the findings, a tentative recommendation is to use the logit method due to the method's higher efficiency in terms of coverage as compared to the regular method.

As a scientific contribution, this work lays the groundwork for further research into when and what transformations should be performed on other model performance measures that are considered non-normal. By sharing the code implemented in this simulation study, I have demonstrated my commitment to accountability and transparency, meaning that the conduct of this study is ethical and in concordance with the tenets of open science.

Acknowledgements

I would like to express my gratitude to my two supervisors, Hanne Oberman and Dr Valentijn de Jong, who guided me throughout this thesis project and helped me improve the manuscript.

References

Addo, E. D. (2018). Performance comparison of imputation algorithms on missing at random data (Doctoral dissertation, East Tennessee State University).

Austin, P. C., & Steyerberg, E. W. (2012). Interpreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable. BMC medical research methodology, 12(1), 1-8.

Van Buuren, S. (2018). Flexible imputation of missing data. CRC press.

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45, 1-67.

Van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine, 18(6), 681-694.

Debray T.P.A De Jong V.M.T. (2021). Package ‘metamisc’.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 837-845.

Enders, C. K. (2010). Applied missing data analysis. Guilford press, 220-221.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention science, 8(3), 206-213.

Groenwold, R. H., White, I. R., Donders, A. R. T., Carpenter, J. R., Altman, D. G., Moons, K. G. (2012). Missing covariate data in clinical research: when and when not to use the missing-indicator method for analysis. Cmaj, 184(11), 1265-1269.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29-36.

Janssen, K. J., Donders, A. R. T., Harrell Jr, F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., & Moons, K. G. (2010). Missing covariate data in medical research: to impute is better than to ignore. Journal of clinical epidemiology, 63(7), 721-727.

Marshall, A., Altman, D. G., Holder, R. L., & Royston, P. (2009). Combining estimates of interest in prognostic 498 modelling studies after multiple imputation: current practice and guidelines. BMC Med Res 499 Methodol, 9(57), 500.

Mason, S. J., & Graham, N. E. (2002). Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 128(584), 2145-2166.

Meng, X. L. 1994. "Multiple Imputation with Uncongenial Sources of Input (with Discusson)." Statistical Science 9 (4): 538–73.

Miles, A. (2016). Obtaining predictions from models fit to multiply imputed data. Sociological Methods Research, 45(1), 175-185.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. Statistics in Medicine, 38(11), 2074-2102.

Oberman, H. I., van Buuren, S., & Vink, G. (2021). Missing the point: Non-convergence in iterative imputation algorithms. arXiv preprint arXiv:2110.11951.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., ... & Robin, M. X. (2021). Package 'pROC' (Vol. 56, pp. 1-71). 2012-09-10 09: 34.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley Sons.

Rubin, D. B. (1987). Multiple imputation for survey nonresponse.

Snell, K. I., Ensor, J., Debray, T. P., Moons, K. G., & Riley, R. D. (2018). Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?. Statistical Methods in Medical Research, 27(11), 3505-3522.

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ... & Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj, 338.

Sun, X., & Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters, 21(11), 1389-1393.

Trikalinos, T. A., Trow, P., & Schmid, C. H. (2013). Simulation-based comparison of methods for meta-analysis of proportions and rates. Agency for Healthcare Research and Quality (US).

Vergouwe, Y., Royston, P., Moons, K. G., & Altman, D. G. (2010). Development and validation of a prediction model with missing predictor data: a practical approach. Journal of Clinical Epidemiology, 63(2), 205-214.

Wahl, S., Boulesteix, A. L., Zierer, A., Thorand, B., & van de Wiel, M. A. (2016). Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Medical Research Methodology, 16(1), 1-18.

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. Ecology, 92(1), 3-10.

Wilson, K., & Hardy, I. C. (2002). Statistical analysis of sex ratios: an introduction. Sex ratios: concepts and research methods, 1, 48-92.

Wilson, E., Underwood, M., Puckrin, O., Letto, K., Doyle, R., Caravan, H., ... & Bassett, K. (2013). The arcsine transformation: has the time come for retirement. Unpubl. manuscript, Meml. Univ. Newfoundland, Newfoundl. Labrador, Canada.