Imputation of ordinal outcomes: a comparison of approaches in traumatic brain injury

Kevin Kunzmann MSc[1], Lorenz Wernisch PhD[1], Sylvia Richardson PhD[1], Ewout W. Steyerberg PhD[2,3], Hester Lingsma PhD[4], Ari Ercole MD, PhD[5], Andrew I.R. Maas MD, PhD[6], David Menon MD, PhD[5], and Lindsay Wilson PhD[7]

[1] MRC Biostatistics Unit, Cambridge Institute of Public Health, University of Cambridge, Cambridge, United Kingdom [2] Dept of Public Health, Erasmus MC, Rotterdam, the Netherlands; [3]Dept of Biomedical Data Sciences, LUMC, Leiden, the Netherlands [4]Center for Medical Decision Sciences, Department of Public Health, Erasmus MC– University Medical Center Rotterdam, Rotterdam, the Netherlands [5]Division of Anaesthesia, University of Cambridge, Addenbrooke's Hospital, Cambridge, United Kingdom [6] Department of Neurosurgery, Antwerp University Hospital and University of Antwerp, Edegem, Belgium [7]Division of Psychology, University of Stirling, Stirling, United Kingdom

Running title: GOSe imputation in TBI
ToC title: Outcomes imputation: comparison of approaches in traumatic brain injury

Corresponding author
Kevin Kunzmann,
MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, United Kingdom

Lorenz Wernisch,
MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, United Kingdom

Sylvia Richardson,
MRC Biostatistics Unit, University of Cambridge, Cambridge Institute of Public Health, Forvie Site, Robinson Way, Cambridge Biomedical Campus, Cambridge CB2 0SR, United Kingdom

Ewout W. Steyerberg,
Dept of Public Health, Erasmus MC - University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

Hester Lingsma,
Dept of Public Health, Erasmus MC - University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, The Netherlands

Ari Ercole,
University Division of Anaesthesia, University of Cambridge, Box 93, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ

Andrew I.R. Maas,
Department of Neurosurgery, Antwerp University Hospital and University of Antwerp, UZA Wilrijkstraat 10, 2650 Edegem, Belgium

David Menon,
University Division of Anaesthesia, University of Cambridge, Box 93, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ

Lindsay Wilson,
Division of Psychology, University of Stirling Stirling FK9 4LA Scotland, United Kingdom

Abstract: Loss to follow-up and missing outcomes data are important issues for longitudinal observational studies and clinical trials in traumatic brain injury. One popular solution to missing 6-month outcomes has been to use the last observation carry forward (LOCF). The purpose of the current study was to compare the performance of model-based single-imputation methods with that of the LOCF approach. We hypothesized that model-based methods would perform better as they potentially make better use of available outcome data. The CENTER-TBI study (n = 4509) included longitudinal outcome collection at 2 weeks, 3 months, 6 months, and 12 months post injury; a total of 8185 GOSe observations were included in the database. We compared single imputation of 6-month outcomes using LOCF, a MI panel imputation, mixed effect model, a Gaussian process regression, and a multi-state model. Model performance was assessed via cross-validation on the subset of individuals with a valid GOSe value within 180 +/- 14 days post-injury (n = 1083). All models were fit on the entire available data after removing the 180 +/- 14 days post-injury observations from the respective test fold. The LOCF method showed lower accuracy (i.e. poorer agreement between imputed and observed values) than model-based methods of imputation, and showed a strong negative bias (i.e. it imputed lower than observed outcomes). Accuracy and bias for the model-based approaches were similar to one another, with the multi-state model having the best overall performance. All methods of imputation showed variation across different outcome categories, with better performance for more frequent outcomes. We conclude that model-based methods of single imputation have substantial performance advantages over LOCF in addition to providing more complete outcome data.

Keywords: traumatic brain injury, missing data, imputation, GOSe

Introduction

Assessments of global functional outcome such as the Glasgow Outcome Scale (GOS) and the Glasgow Outcome Scale extended (GOSe) are used across the full spectrum of recovery, and have popularity as endpoints in traumatic brain injury.[1,2] However, missing outcome data is a common problem in TBI research, and for longitudinal studies completion rates at six months can be lower than 70%.[3] This is important, since complete-case analyses may introduce bias and at least reduce power. [4]

Last observation carried forward (LOCF) is a recommended single-imputation method for dealing with missing data in TBI research clinical trials because it is conservative with respect to evaluation of the intervention.[5] One recognized version of this approach is to substitute the three-month outcome for missing six-months data.[6,7] Although LOCF is easy to understand and implement, the technique is suboptimal in several respects. Firstly, it is biased in that it ignores potential time trends in GOS(e) trajectories. Secondly, application of LOCF is inefficient, since it neglects data observed briefly after the target time window. For example, a GOS(e) value recorded at 200 days post-injury is likely to be more informative about the status at 180 days post-injury than a value observed 90 days post-injury. Finally, the *ad-hoc* nature of the LOCF method implies that there is no probabilistic model, and thus no measure of uncertainty concerning the imputed values. This also implies that it is impossible to include additional covariates to further reduce bias introduced by the imputation method and that LOCF cannot be used to obtain multiply imputed data sets by design. Statistical Imputation of patient outcomes is gradually gaining acceptance in the TBI field as a method of dealing with missing data. Recent longitudinal studies have successfully employed techniques for both single[6,8,9] and multiple imputation.[10–13]

Model-based imputation may not only be of value in case of missing outcomes, but also for dealing with effects of broad time windows for assessments. The variation in timing of outcome assessments for patients with TBI varies between studies. Some studies define very

stringent time windows (e.g. +/- 2 weeks; https://tracktbi.ucsf.edu/researchers), but in some contexts this can lead to a substantial amount of missing data).[3] Consequently other studies have defined more pragmatic protocol windows (e.g. -1 month to +2 months, see also[14]). While the wider windows enable more complete data collection, they suffer from the problem that outcome can be evolving over this period, and an outcome assessment obtained at five months (the beginning of this window) in one subject may not be strictly comparable with outcomes obtained just before eight months (the end of the window) in another subject. Consequently, even where outcomes are available within pragmatic protocol windows, there may be a benefit from being able to impute an outcome more precisely at the 180 day (6 month) time point.

In this manuscript, four model-based imputation strategies for GOSe at 6 months (=180 days) post-injury in the longitudinal CENTER-TBI study[14] are compared with LOCF with respect to their single-imputation performance. The focus on single-imputation is due to the fact that the imputed values are to be integrated in the CENTER-TBI database and used in subsequent analyses by investigators. We examine four different model-based approaches – a panel imputation approach using multiple imputation via chained equation (MICE), a mixed-effects model, a Gaussian process regression, and a multi-state model - for imputing cross-sectional GOSe at 6 months exploiting the longitudinal GOSe measurements. Each model is fit in a version with and without baseline covariates.

Materials and Methods

Study population

The CENTER-TBI project methods and design are described in detail elsewhere.[14] Participants with TBI were recruited into three strata: (a) patients attending the emergency room, (b) patients admitted to hospital but not intensive care, and (c) patients admitted to intensive care. Follow-up of participants was scheduled per protocol at 2 weeks, 3 months, and 6 months in group (a) and at 3 months, 6 months, and 12 months in groups (b) and (c). The protocol time window for the 6 months GOSe was between -1 and +2 months from the 6-months time point (5-8 months post injury). Outcome assessments at all timepoints included the GOSe. The GOSe has the following categories: (1) dead, (2) vegetative state, (3) lower severe disability, (4) upper severe disability, (5) lower moderate disability, (6) upper moderate disability, (7) lower good recovery, (8) upper good recovery. The GOSe was collected using structured interviews[15] and patient/carer questionnaires[16]. Since the latter does not identify vegetative patients as a separate category, the vegetative state and lower severe disability were combined in one group.

The study population for this empirical methods comparison are all individuals from the CENTER-TBI database (total of n = 4509) whose GOSe status was recorded at least once within the first 18 months and who were still alive 180 days post-injury (n = 3343). The rationale for conducting the comparison conditional on 6-months survival is simply that the GOSe can only be missing at 6-months if the individuals are still alive since GOSe would be known to be “dead” otherwise. Data for the CENTER-TBI study were collected through the Quesgen electronic-case report form (Quesgen Systems Inc, USA), hosted on the International Neuroinformatics Coordinating Facility (INCF) platform and extracted via the INCF

Neurobot tool (https://neurobot.incf.org/). Release 1.1 of the database was used (cf. Appendix for details). Basic summary statistics for population characteristics are listed in Table 1.

We decided to use only those GOSe observations obtained between injury and 18 months post injury, since extremely late follow-ups were considered uninformative for the index follow-up time point of 6 months post injury. This led to a total of 8185 GOSe observations of the study population being available for the analyses. For 1151 (34%) individuals, GOSe observations at 180 +/- 14 days post injury were available and 2394 (72%) individuals had GOSe observations within the per-protocol window of 5-8 months post injury. The distribution of GOSe sampling times and both absolute and relative frequencies of the respective GOSe categories are shown in Figure 1. True observation times were mapped to categories by rounding to the closest time point, i.e., the '6 months' category contains observations up to 9 months post-injury. Thus, the figures include a small proportion of GOSE 1 representing patients who died between 6 and 9 months.

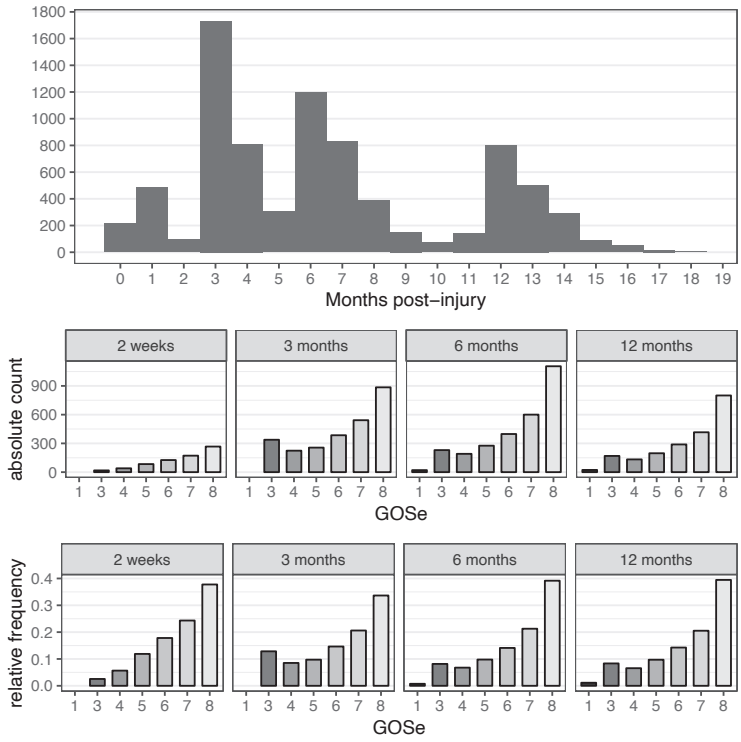

Figure 1: GOSe sampling time distribution and distribution at per-protocol time points (actual date rounded to nearest assessment window).

Imputation methods

We compared LOCF to a MICE panel regression approach (MI), a mixed effect model (MM), a Gaussian process regression (GP), and a multi-state model (MSM). For all model-based approaches we additionally explored variants including the key IMPACT[7] predictors as covariates. These are age, GCS motor score, pupil reactivity (0, 1, 2), hypoxia, hypotension, Marshall CT classification, traumatic subarachnoid hemorrhage, epidural hematoma, glucose, and hemoglobin.

Last-observation-carried-forward

Since LOCF is widely used to impute missing outcomes in TBI studies,[6,7,9] it served as the comparator method. Here, LOCF was defined as the last GOSe observation before the imputation time point of 180 days post-injury. LOCF is not a model-based method and, by definition, only permits the imputation of a GOSe value for subjects where at least one value was available within the first 180 days post injury. We accounted for this lack of complete coverage under LOCF by performing all performance comparisons including LOCF only on the subset of individuals for which a LOCF-imputed value can be obtained.

Model-based methods

Model-based imputation approaches offer richer output (probabilistic imputation, multiple imputation) and may reduce the LOCF-inherent bias. We compared the performance of four model-based approaches to that of LOCF.

The MICE regression approach (MI) is a standard approach to multiple imputation that defines regression models for each missing variable in a matrix.[17] By iterating over each varia-

ble that contains missing values, and resampling missing values from the corresponding regression model while holding all other variables fixed, a steady-state canultimately be reached, and a set of imputed datasets can be generated. Since our goal is single imputation, we reduced the set of imputed values to a prediction by taking the most frequently imputed GOSe value. The frequency distribution of the imputed GOSe values can be used as probabilistic prediction in very much the same way as the probabilistic output of other model-based methods. To incorporate the longitudinal aspect of GOSe, we jointly imputed GOSe at 2 weeks, 3 months, 6 months and 12 months jointly. This means that the GOSe at 2 weeks, 3 months, and 12 months act as covariates in the regression model for the 6-months GOSe. Mixed effects models (MM) are a widely used approach in longitudinal data analysis and models individual deviations from the population mean trajectory. [18] The mixed effects model used for the GOSe imputation incorporates time as non-linear covariate via a spline to be able to capture non-linear dynamics of GOSe over time in the population. The mixed effects model was fitted using Bayesian methods to allow the inclusion of patient-specific quadratic random effect (see Appendix for details). An alternative non-linear regression model for longitudinal data is Gaussian process regression (GP) which allows flexible modelling of both the individual GOSE trajectories as well as the population mean in a Bayesian non-parametric way.[19] Both the employed mixed effects model as well as the Gaussian process regression model are non-linear regression techniques for longitudinal data. While these are powerful tools to model longitudinal trajectories, they do not explicitly model the probability of transitions between GOSe states. Since the number of observations per individual is limited in our data set (1 to 4 GOSe observations per individual), an approach explicitly modelling transition probabilities might be more suitable to capture the dynamics of the GOSe trajectories. To explore this further, a Markov multi-state model (MSM) was considered.[20,21]

All models were fitted using either none or all IMPACT predictors except for the MSM model which only used age due to issues with numerical stability. Computational intensity is hard to compare since it depends on the exact hardware used. All methods except MSM can at least partially be run in parallel. On a Mac Book Pro 2019 the required time to fit each of the

models on the entire available CENTER TBI data (with IMPACT covariates) was 13 minutes (MI), 26 minutes (MSM), 86 minutes (MM), and 112 minutes (GP). Although these differences are substantial, they are within one order of magnitude and would not preclude any of the methods in practice. Further details on the implementations are given in the Appendix.
All models, irrespective of the fact whether they are Bayesian or frequentist, produce probabilistic outputs, i.e., a discrete probability distribution over the possible GOSe values at 6 months for each individual. Although we propose only to store these probabilities along with the most likely GOSe value at 6 months, multiple imputations can be obtained post-hoc by resampling from the discrete probability distribution of each individual via inverse transform sampling.[22] The functions required to sample from a discrete probability distribution are available in any statistical software package.
All four model-based approaches allow unbiased inference under a 'missing at random' (MAR) mechanism [23] Here, MAR means that the fact whether or not a GOSe observation is missing is independent of the true functional outcome status of the individual. GOSe is an interview-based assessment that can also be completed by a proxy. The main operational challenge for consistent collection of longitudinal GOSe for observational studies thus lies in the organisation and scheduling of the interviews. A MAR assumption for CENTER data was thus deemed plausible, albeit is not testable.[23]

Performance assessment

Model performance was assessed via three-fold cross validation on the subset of individuals with a valid GOSe value within 180 +/- 14 days post-injury (n = 1083). All models were fit on the entire available data after removing the 180 +/- 14 days post-injury observation from the respective test fold thus mimicking a missing completely at random missing data mechanism. The distribution of GOSe values in the three test sets was well balanced, (cf. Appendix, Figure A.1). All confusion matrices are reported as averages over the three-fold cross

validation test sets. The column fraction confusion matrices are normalized within each category of observed GOSe value and are thus estimates of confusion probability conditional on the observed GOSe. Performance was assessed using the absolute-count and the normalized (proportions) confusion matrices as well as bias, mean absolute error (MAE), and root mean squared error (RMSE). Bias is calculated by averaging the signed differences between observed and imputed values. A negative value of bias signifies that predicted values are lower overall than observed values, and a positive value means that they are higher. If differences cancel each other out then bias can be zero even if the the predictions are inaccurate. MAE employs the unsigned differences, and it therefore gives a measure of accuracy irrespective of whether imputed outcomes are higher or lower than observed. In the calculation of RMSE the differences are squared, which penalizes large deviations from the target value more strongly than small ones. For example, the MAE will be 0.5 if 50% of imputed values agree with observed values and 50% differ by one category. The same MAE will arise if 75% agree exactly and 25% disagree by two categories. In the former case the RMSE will be 0.71 and in the latter 1.0.

These metrics have some limitations with ordinal data, and we therefore also considered directional bias (d-bias), which was calculated as the difference between the model-fitted probability of exceeding the observed value minus the model-fitted probability of undershooting the observed GOSe as an alternative measure of bias. It is important to note that the scale of the directional bias is not directly comparable to the one of the other three quantities. All measures were calculated in the dataset that was conditional on the ground-truth (observed 6-months GOSe) as well as averaged over the entire test set.

LOCF, by design, cannot provide imputed values when there are no observations before 180 days post injury. A valid comparison of LOCF with the other methods must therefore be based on the set of individuals for whom an LOCF imputation is possible. Overall, 118 out of 1083 test cases (10.9%) could not be imputed with the LOCF approach. In the entire study population, 345 individuals (10.3%) did not have data that would permit an LOCF imputation.

The subset used for comparison of the imputation approaches with the LOCF approach was similar to the overall dataset (cf. Table 1).

Results

The overall performance of all fitted models in terms of bias, d-bias, MAE, and RMSE is depicted in Figure 2 both conditional on LOCF being applicable (gray) and, excluding LOCF, on the entire test set (black). Values are reported as mean over the three cross-validation folds and error bars indicate +/- 1.96 standard errors.

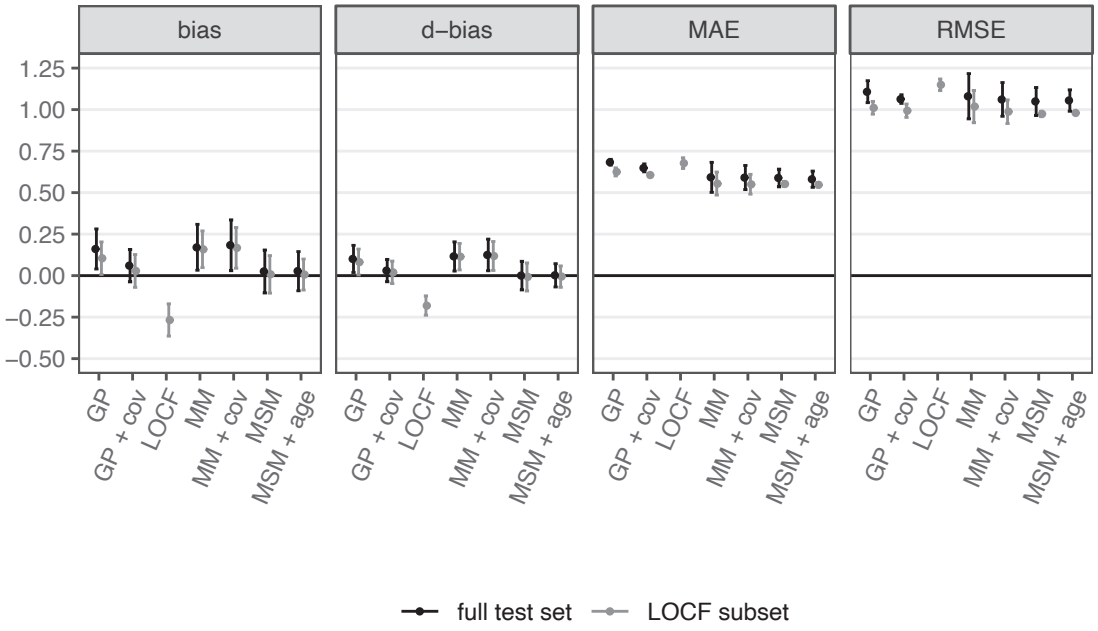

Figure 2: Cross-validated overall performance for all fitted models on the LOCF subset (allowing LOCF) and the entire test set (LOCF performance not shown).

Several key findings are worth highlighting. Firstly, LOCF is overall negatively biased, i.e., on average it imputes lower-than-observed GOSe values. This reflects a population average trend towards continued recovery within the first 6 months post injury. The fact that both ways of measuring bias qualitatively agree, suggests that application of these metrics is reasonable for the data. In terms of MAE and RMSE, LOCF also has worst performance, but differences between methods are less pronounced than for measures of bias. Notably, the RMSE difference between LOCF and the other methods is slightly larger than the MAE difference which indicates that LOCF tends to produce more large deviations, i.e., across several GOSe categories.

Second, including baseline covariates only produces clinically meaningful impact in the case of the GP regression model. The MI, MM, and MSM models perform more or less the same irrespective of adjustment for baseline covariates. This indicates that the additional predictive value of baseline covariates over the information contained in at least one observed GOSe value is limited. Furthermore, both variants of the MI model and the mixed effects model fail to correct the overall bias of the imputed values.

We proceed with a detailed analysis of a subset of models both in direct comparison with LOCF and in the entire data set including those cases where LOCF is not applicable. In the following we only consider the baseline-adjusted Gaussian process model ('GP + cov'), the MI model without baseline covariates, the mixed effect model without baseline covariates ('MM'), and the multi-state model without baseline covariates ('MSM'). The rationale behind dropping baseline adjustment for MI, MM, and MSM being that the additional complexity does not substantially alter overall performance. On the other hand, the GP model benefits from the inclusion of the IMPACT baseline covariates.

Detailed comparison conditional on LOCF subset

We first consider the results for the set of test cases which allow LOCF imputation (n = 965). Both the raw count as well as the relative (by left-out observed GOSe) confusion matrices are presented in Figure 3. The GOSe scale is restricted to 3+ since the imputation is conditional on an observed GOSe larger than 1 (deaths are known and no imputation necessary) and GOSe 2 was not distinguished as a separate category.

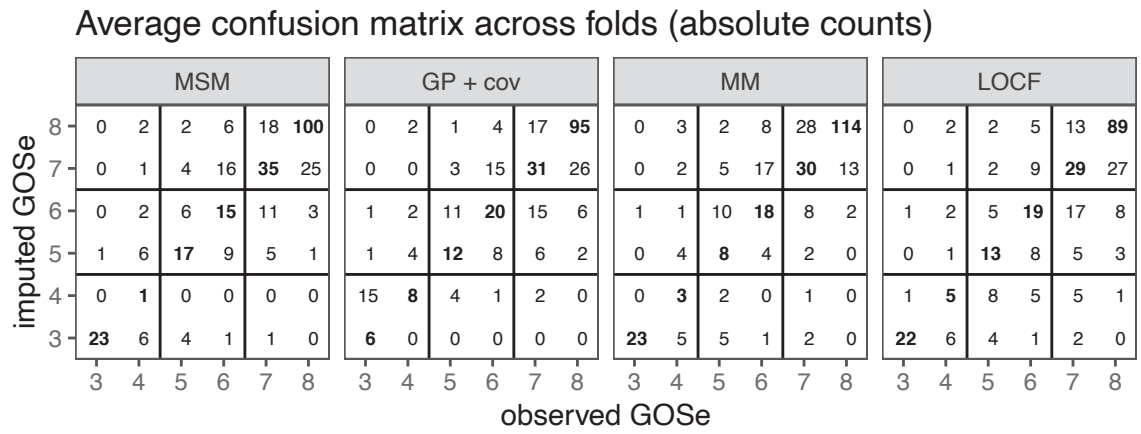

Figure 3: Confusion matrices averaged across folds (LOCF subset only); absolute counts. Note that imputation is carried out with the relevant observed values (test set) of the GOSe removed

The absolute-count confusion matrices show that most imputed values are within +/- one GOSE categories of the observed ones and this yields an RMSE of approximately 1. However, they also reflect the category imbalance (cf. Figures 1) in the study population. The performance conditional on the (in practice removed) observed GOSe value clearly shows that imputation for the most infrequent category 4 is hardest. This is true across the range of methods considered. Both the MSM and the MM models account for this difficulty by almost never imputing a GOSe of 4. Instead, the respective cases tend to be imputed to GOSe 3 or 5.

To better understand the overall performance assessment in Figure 2, we also consider the performance conditional on the respective ground-truth (i.e. the observed GOSe categories in the test sets). The results are shown in Figure 4 (vertical bars are +/-1.96 standard error of the mean).

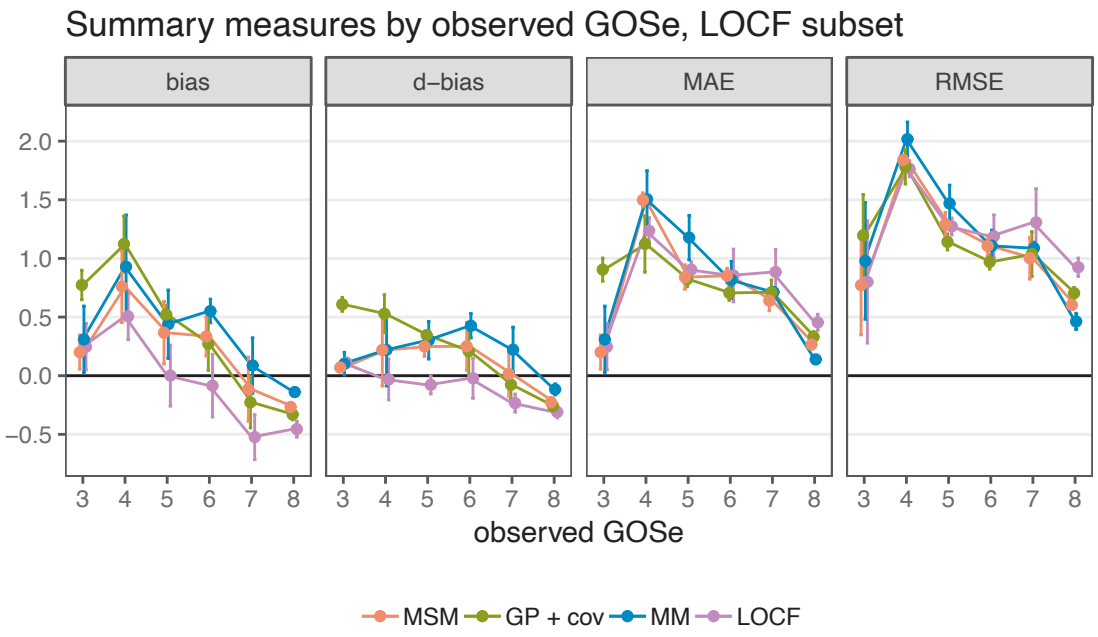

Figure 4: Performance measures by observed GOSe; LOCF subset only.

Just as with overall performance, differences are most pronounced in terms of bias. Interestingly, the assessment conditional on LOCF being feasible reveals differences between bias as the difference between mean imputed and mean observed values and the difference in the probability to over- or undershoot the observed value. Again, the category imbalance in the GOSe distribution explains the fact that all model-based approaches tend to perform better for the most frequent categories 6, 7, and 8 while sacrificing performance for the less frequent categories 4 and 5 as compared to LOCF. With respect to bias, all methods exhibit a certain regression to the mean effect since low categories tend to be confused with better (higher) GOSe on average while high observed GOSe values are subject to a negative bias (at GOSe 7 and 8). Since LOCF does not take the category imbalance into account and since it exhibits a relatively large negative bias at the most frequent GOSe values, it is overall negatively biased. The conditional assessment of the GP regressions bias profile reveals overall unbiasedness, but this is the consequence of the relatively high positive and negative biases conditional on low/high GOSe values canceling each other out in the overall population. The MI, MSM, and MM models are fairly similar with respect to accuracy but MSM clearly dominates with respect to bias. Note that irrespective of the exact definition of bias

used, MSM dominates the other model-based approaches. Comparing LOCF and MSM, there is a slight advantage of MSM in terms of accuracy for the majority classes 3, 7, 8 which explain the overall difference shown in Figure 2. With respect to bias, MSM also performs better than LOCF for the most frequently observed categories, but the extent of this improvement depends on the performance measure.

Detailed comparison on full test set

LOCF was not considered in the analysis of the full dataset, since no LOCF was available for subjects with a first recorded outcome assessment more than six months post-TBI, and this renders a meaningful comparison across the entire dataset impossible. The qualitative performance of the three remaining imputation approaches in the complete dataset was similar to their performance in the subset of data used for comparison with LOCF (cf. Figures 5 and 6).

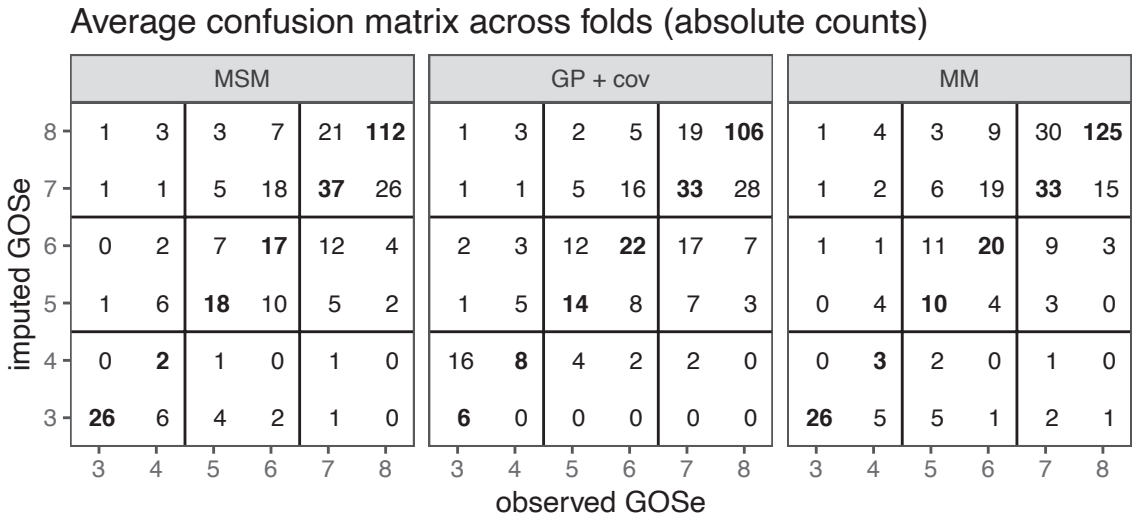

Figure 5: Confusion matrices averaged across folds (entire test set, no LOCF); absolute counts. Note that imputation is carried out with the relevant observed values (test set) of the GOSe removed.

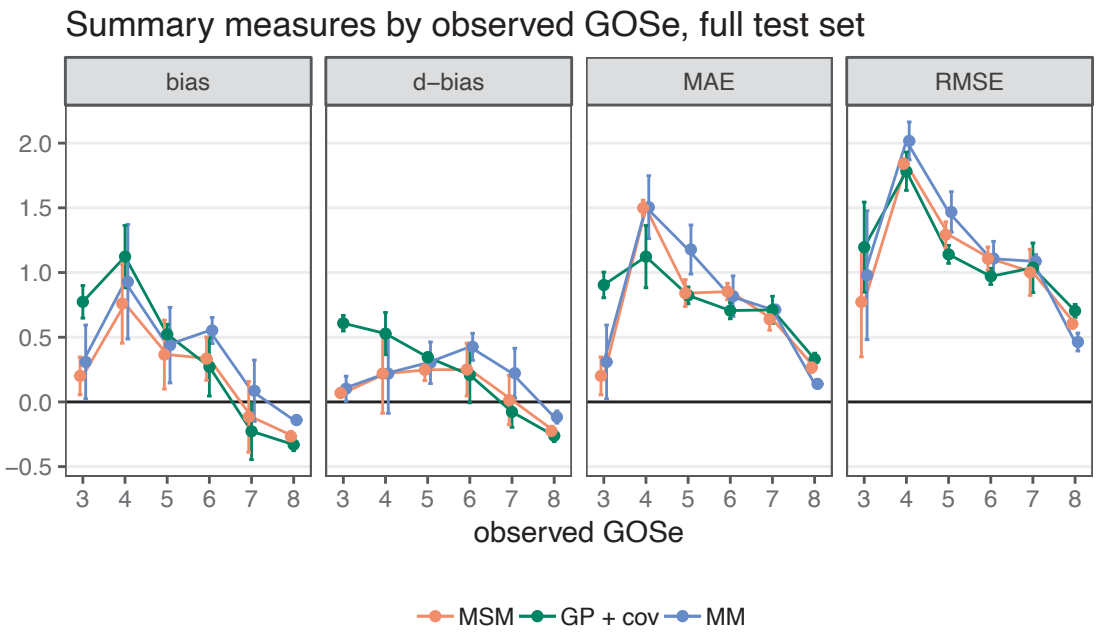

Figure 6: Performance measures by observed GOSe; LOCF subset only.

Discussion

Handling missing data post-hoc to mitigate biases in analyses often requires great effort. It is thus of the utmost importance to implement measures for avoiding missing data in the first place. Nevertheless, in practice, missing values due to loss-to-follow-up will always occur and should be addressed effectively.[3] There is a wide consensus that statistically sound imputation of missing values is beneficial for both the reduction of bias and for increasing statistical power. The current gold-standard for imputing missing values is multiple imputation on a per-analysis basis, including analysis-specific covariates to further reduce bias and to preserve the imputation uncertainty in the downstream analysis. In practice, however, there are good reasons for providing a set of single-imputed default values in large observational studies such as CENTER-TBI. CENTER-TBI is committed to providing a curated database to facilitate multiple subsequent analyses. Since one of the primary endpoints in CENTER-

TBI is functional outcome at 6 months, a single default imputed value for as many study participants as possible is desirable. Consortia are increasingly committed to making their databases available to a range of researchers. In fact, more liberal data-sharing policies are becoming a core requirement for funding bodies (cf. https://www.openaire.eu/). In this context, it might not be possible to ensure that every analysis team has the necessary statistical expertise to properly conduct a per-analysis multiple imputation in the future. Furthermore, the imputed values of a multiple-imputation procedure are inherently random, and it is thus difficult to ensure consistency across different analysis teams if the values themselves cannot be stored directly in a database. For this reason, as a practical way forward, we suggest providing a default single-imputation together with a predictive distribution (value probabilities) for key outcomes in the published data base itself. This mitigates problems with complete-case analyses and provides a principled and consistent default approach to handling missing values. Given the strong case for employing model-based approaches to imputation, it makes good sense to provide the predicted probabilities for each GOSe outcome in the core database alongside single imputed values as a transparent method for quantifying confidence in the imputation prediction. Based on these probabilities, it is easy to draw samples for a multiple imputation analysis if needed. Since we did not find any of the common predictors of GOSe to have substantial effect on the imputed values in the presence of at least one observed GOSe value (at another timepoint than 6 months), the imputed values can be used in a wide range of subsequent analyses.
Wherever necessary and practical, a custom, analysis-specific multiple imputation approach might still be employed. In these cases, the model providing the single-imputed values may be used as a starting point.

Several reasons disqualify LOCF as method of choice. Not only is it inherently biased, but it is also inefficient in that it fails to properly account for the category imbalance of GOSe in the respective target population. Albeit simple to implement, LOCF - by definition - is not capable of exploiting longitudinal information obtained after the target time point. This results in a

smaller subset of individuals for which imputed values can be provided in the first place. LOCF also lacks flexibility to adjust for further covariates which might be necessary in some cases to further reduce bias under a missing at random assumption. Finally, LOCF cannot produce an adequate measure of imputation uncertainty, since it is not model based.
Given this context, we draw two main conclusions from our comparison of three alternative, model-based approaches.

First and despite its theoretical drawbacks, LOCF is hard to beat in terms of accuracy (both MAE and RMSE). The main advantages of a model-based approach are thus the ability to impute values for the entire study population, the reduction of bias, the ability to provide a measure of uncertainty (value probabilities) together with the imputed values (or to use the same very same model to draw multiple imputations), as well as the possibility of including further analysis-specific covariates.

Second, we found that the inclusion of established baseline predictors had little effect on the imputation quality. Note that this does not refute their predictive value, and the IMPACT covariates may be more relevant in studies confined to moderate and severe injuries. However, the current study suggests that there is little added benefit once at least one GOSe value is known. Differences between the various model-based approaches are rather nuanced. The more complex longitudinal models (GP, MM) that where fitted using Bayesian methods did not perform better and due to the inherent complexities of a Bayesian analyses (e.g., convergence assessment) we discourage their use for the specific application at hand. The more standard MI-based approached achieved similar performance to the MSM approach in terms of precision (MAE or RMSE) but did not remove bias completely. We thus favour the multi-state model (MSM) for several reasons. It is well-interpretable in terms of transition intensities, and an efficient implementation is available[21] in standard statistical software.[24] Finally, it is the only method considered here that succeeds in eliminating the bias observed with LOCF. As all other model-based imputation methods, MSM is able to provide

imputed values for the entire population and to provide a probabilistic output to quantify imputation uncertainty.

Funding sources statement:

Data used in preparation of this manuscript were obtained in the context of CENTER-TBI, a large collaborative project with the support of the European Union 7th Framework program (EC grant 602150). Additional funding was obtained from the Hannelore Kohl Stiftung (Germany), from OneMind (USA) and from Integra LifeSciences Corporation (USA).

Acknowledgements:

The authors would like to express their gratitude towards Abhishek Dixit and Visakh Muraleedharan for their tireless efforts in setting up and maintaining the CENTER-TBI database.

Appendix / Supplemental Material

Ethical approval statement

The CENTER-TBI study (EC grant 602150) has been conducted in accordance with all relevant laws of the EU if directly applicable or of direct effect and all relevant laws of the country where the Recruiting sites were located, including but not limited to, the relevant privacy and data protection laws and regulations (the “Privacy Law”), the relevant laws and regulations on the use of human materials, and all relevant guidance relating to clinical studies from time to time in force including, but not limited to, the ICH Harmonised Tripartite Guideline for Good Clinical Practice (CPMP/ICH/135/95) (“ICH GCP”) and the World Medical Association Declaration of Helsinki entitled “Ethical Principles for Medical Research Involving Human Subjects”. Informed Consent by the patients and/or the legal representative/next of kin was obtained, accordingly to the local legislation, for all patients recruited in the Core Dataset of CENTER-TBI and documented in the e-CRF. Ethical approval was obtained for each recruiting site. The list of sites, Ethical Committees, approval numbers and approval dates can be found on the website: https://www.center-tbi.eu/project/ethical-approval.

Distribution of GOSe in validation folds

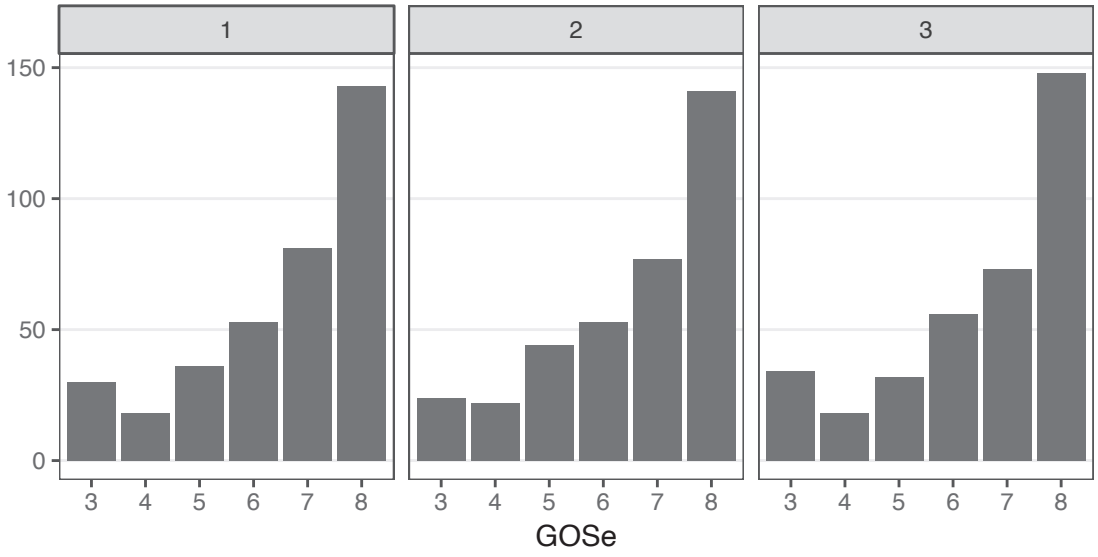

Figure A.1: Marginal distribution of GOSe over the three cross validation folds.

Models

MI approach

Multiple imputation via chained equations (MICE) is a standard approach to multiple imputation.[17] Instead of specifying a full (longitudinal) model for GOSe the MI approach uses a so called "fully conditional approach" to specify a model for each variable in a dataset given all other variables. By sequentially refitting these conditional models and then resampling missing values a set of imputed datasets can be generated. For the GOSe prediction, we treated the GOSe at the nominal query points of two weeks, 3 months, 6 months, and 12 months as separate variables and imputed them using the mice package for the R programming language.[24] We used proportional odds ordinal regression [25] as conditional models for the ordinal outcome GOSe at the four different time points and ran the algorithm with 100 chains up to convergence. This means that 100 potential GOSe outcomes per individual with missing GOSe at 6 months are available upon convergence of the procedure. The final probabilistic

predictions where then taken as the relative frequencies of the imputed outcomes and the point prediction as the most frequently imputed value.

Mixed-effects model

Mixed effects models are a widely used approach in longitudinal data analysis and model individual deviations from a population mean trajectory.[18] To account for the fact that the GOSe outcome is an ordered factor, we employ a cumulative link function model with flexible intercepts.[25] The population mean is modeled as a cubic spline function to allow a non-linear population mean trajectory. Patient-individual deviations from this population mean are modeled as quadratic polynomials to allow sufficient flexibility (random effects). Baseline covariates are added as linear fixed effects to the population mean. The model was fitted using Bayesian statistics via the BRMS package[26,27] for the R environment for statistical computing[24] and the Stan modelling language for Markov Chain Monte Carlo sampling.[28] The required burn-in length to reach a steady state for the Markov Chains was determined by inspecting the trace plot of the model on the complete data set. During the cross validation fits the same burn-in length was used and convergence was assessed via the potential scale reduction factor (PSRF) proposed by Gelman and Rubin.[29] A Bayesian approach was necessary since the quadratic random effect is not identifiable in individuals where only one or two GOSe values are available. The model with cubic spline fixed effect for time and quadratic random effects for time per individual was selected based on the highest expected log predictive density on the full data set (computed via the loo package[30], data not shown). Non-informative priors were used for the model parameters. These are sufficient to make all model parameter identifiable and effectively shrink the quadratic random effect in individuals with only two observed GOSe values to zero. A potential drawback of the proposed longitudinal mixed effects model is the fact that the individual deviations from the population mean are modeled globally using polynomials. Since linear and quadratic terms are not identifiable

for patients with only one observed GOSE value, this implies large uncertainty over the patient-specific effects for individuals with one or two observations only by falling back on the non-informative priors on these model parameters. Thus, the overall uncertainty associated with model-based imputations at exactly 180 days may become relatively large for these individuals. A more flexible regression model might avoid this particular pathology which is why we also implement a Gaussian process regression model (see below).

Gaussian process model

Gaussian process regression allows flexible modelling of both the individual GOSE trajectories as well as the population mean in a Bayesian non-parametric way.[19] This non-parametric paradigm leads to low model-uncertainty in the vicinity of actually observed GOSe outcomes. To account for the discreteness of the GOSe outcome, the continuous output of the Gaussian process model is rounded to the nearest integer in 1 to 8 (GOSe categories). The squared exponential covariance function with shared length scale for all individuals is used to model intra-individual dependency of GOSe outcomes. The population mean trajectory of the Gaussian process is modeled as mean of an independent Gaussian process with pseudo observations at 45, 90, 180, 270, 360 days post-injury (also using a squared exponential covariance function). This approach maintains flexibility of the population mean function similar to the spline-based approach for the mixed effects model while avoiding the computational complexity of a fully hierarchical Gaussian process model. Again, the impact of baseline covariates is modeled via linear effects on the population mean of the Gaussian process. All parameters are estimated in a fully Bayesian fashion using the Stan modelling language[28] and non-informative priors except for the length scale of the squared exponential kernel. Due to the sparseness of the data, the estimated length scale will naturally tend towards extremely large values implying unrealistically long-range dependency between observations. We therefore limit the length scale to a maximum of 120 days (4 months) and impose a Gaussian prior with a mean of 60 days post injury and a standard deviation of 14.

Multi-state model

Both the mixed effects model as well as the Gaussian process regression model are essentially non-linear regression techniques for longitudinal data. While they are both powerful tools to model longitudinal trajectories, they do not explicitly model the probability of transitions between GOSe states. Since the number of observations per individual is limited in our data set (1 to 4 GOSe observations per individual), an approach explicitly modelling transition probabilities might be more suitable to capture the dynamics of the GOSe trajectories. To explore this further, a Markov multi-state model is considered.[20] This model class assumes that the transitions between adjacent GOSe states can be modeled as a Markov process and the transition intensities between adjacent states are fitted to the observed data.

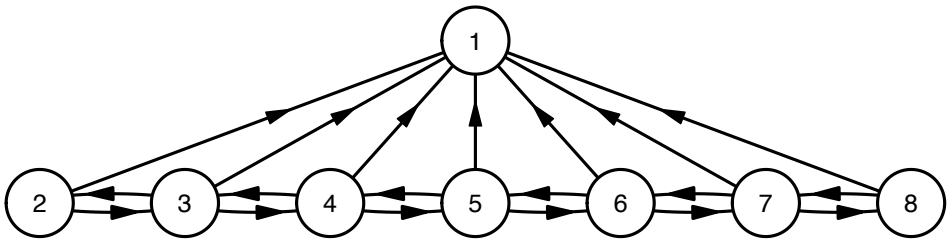

Figure A.2: Structural diagram of allowed transitions between GOSe states for the proposed multi-state model.

To account for the fact that state-transitions might be more frequent in the early post-injury phase, piecewise constant transition intensities were fitted to the intervals [0, 90), [90, 270), and 270+ days post-injury. The model was fit using the msm package[21] for the R environment for statistical computing.[24] Due to the relatively large number of 19 transition intensities in the proposed model (cf. arrows in Figure A.1, structure of transition graph), inclusion of all baseline covariates turned out to be numerically unstable. For the MSM model, instead of

including all covariates, only a model adjusting for age at injury via a proportional hazard approach was fit.

Reproducible Research Strategy

CENTER-TBI is committed to reproducible research. To this end, the entire source code to run the analyses is publicly available at https://git.center-tbi.eu/kunzmann/gose-6mo-imputation. Scripts for automatically downloading the required data from the central access restricted 'Neurobot' (https://neurobot.incf.org/) database at https://center-tbi.incf.org/ are provided. The analysis is completely automated using the workflow management tool 'snakemake[31] and a singularity[32] container image containing all required dependencies is publicly available from zenodo.org (DOI: 10.5281/zenodo.2600385). Detailed step-by-step instructions on how to reproduce the analysis are provided in the README.md file of the GitLab repository.

Author Disclosure Statement

No competing financial interests exist.

References

1. Horton, L., Rhodes, J., and Wilson, L. (2018). Randomized controlled trials in adult traumatic brain injury: a systematic review on the use and reporting of clinical outcome assessments. J. Neurotrauma 35, 2005–2014.

2. McMillan, T., Wilson, L., Ponsford, J., Levin, H., Teasdale, G., and Bond, M. (2016). The Glasgow Outcome Scale - 40 years of application and refinement. Nat. Rev. Neurol. 12, 477–485.

3. Richter, S., Stevenson, S., Newman, T., Wilson, L., Menon, D.K., Maas, A.I., Nieboer, D., Lingsma, H., Steyerberg, E.W., and Newcombe, V.F. (2019). Handling of missing outcome data in traumatic brain injury research: a systematic review. J. Neurotrauma 36, 2743–2752.

4. White, I.R., and Carlin, J.B. (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat. Med. 29, 2920–2931.

5. Yeatts, S.D., Palesch, Y.Y., and Temkin, N. (2018). Biostatistical issues in TBI clinical trials., in: *Handbook of Neuroemergency Clinical Trials*. Elsevier, pps. 167–185.

6. Skolnick, B.E., Maas, A.I., Narayan, R.K., Van Der Hoop, R.G., MacAllister, T., Ward, J.D., Nelson, N.R., and Stocchetti, N. (2014). A clinical trial of progesterone for severe traumatic brain injury. N. Engl. J. Med. 371, 2467–2476.

7. Steyerberg, E.W., Mushkudiani, N., Perel, P., Butcher, I., Lu, J., McHugh, G.S., Murray, G.D., Marmarou, A., Roberts, I., and Habbema, J.D.F. (2008). Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 5, e165.

8. Clifton, G.L., Valadka, A., Zygun, D., Coffey, C.S., Drever, P., Fourwinds, S., Janis, L.S., Wilde, E., Taylor, P., and Harshman, K. (2011). Very early hypothermia induction in patients with severe brain injury (the National Acute Brain Injury Study: Hypothermia II): a randomised trial. Lancet Neurol. 10, 131–139.

9. Silver, J.M., Koumaras, B., Chen, M., Mirski, D., Potkin, S.G., Reyes, P., Warden, D., Harvey, P.D., Arciniegas, D., and Katz, D.I. (2006). Effects of rivastigmine on cognitive function in patients with traumatic brain injury. Neurology 67, 748–755.

10. Bulger, E.M., May, S., Brasel, K.J., Schreiber, M., Kerby, J.D., Tisherman, S.A., Newgard, C., Slutsky, A., Coimbra, R., and Emerson, S. (2010). Out-of-hospital hypertonic resuscitation following severe traumatic brain injury: a randomized controlled trial. Jama 304, 1455–1464.

11. Kirkness, C.J., Burr, R.L., Cain, K.C., Newell, D.W., and Mitchell, P.H. (2006). Effect of continuous display of cerebral perfusion pressure on outcomes in patients with traumatic brain injury. Am. J. Crit. Care 15, 600–609.

12. Wright, D.W., Yeatts, S.D., Silbergleit, R., Palesch, Y.Y., Hertzberg, V.S., Frankel, M., Goldstein, F.C., Caveney, A.F., Howlett-Smith, H., and Bengelink, E.M. (2014). Very early administration of progesterone for acute traumatic brain injury. N. Engl. J. Med. 371, 2457–2466.

13. Robertson, C.S., Hannay, H.J., Yamal, J.-M., Gopinath, S., Goodman, J.C., Tilley, B.C., Baldwin, A., Lara, L.R., Saucedo-Crespo, H., and Ahmed, O. (2014). Effect of erythropoietin and transfusion threshold on neurological recovery after traumatic brain injury: a randomized clinical trial. Jama 312, 36–47.

14. Maas, A.I., Menon, D.K., Steyerberg, E.W., Citerio, G., Lecky, F., Manley, G.T., Hill, S., Legrand, V., and Sorgner, A. (2015). Collaborative European NeuroTrauma effectiveness research in traumatic brain injury (CENTER-TBI) a prospective longitudinal observational study. Neurosurgery 76, 67–80.

15. Wilson, J.L., Pettigrew, L.E., and Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J. Neurotrauma 15, 573–585.

16. Wilson, J.T.L., Edwards, P., Fiddes, H., Stewart, E., and Teasdale, G.M. (2002). Reliability of postal questionnaires for the Glasgow Outcome Scale. J. Neurotrauma 19, 999–1005.

17. Buuren, S. van, and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. J. Stat. Softw. 45, 1–68.

18. Verbeke, G., and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

19. Williams, C.K., and Rasmussen, C.E. (2006). Gaussian processes for machine learning. MIT press Cambridge, MA.

20. Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P.K. (2009). Multi-state models for the analysis of time-to-event data. Stat. Methods Med. Res. 18, 195–222.

21. Jackson, C.H. (2011). Multi-state models for panel data: the msm package for R. J. Stat. Softw. 38, 1–29.

22. Devroye, L. (1986). Discrete Random Variates., in: *Non-Uniform Random Variate Generation*. New York, NY: Springer, pps. 83–117.

23. Rubin, D.B. (1976). Inference and missing data. Biometrika 63, 581–592.

24. R Core Team. (2020). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

25. Agresti, A. (2003). Categorical data analysis. John Wiley & Sons.

26. Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. J. Stat. Softw. 80, 1–28.

27. Bürkner, P.-C. (2017). Advanced Bayesian multilevel modeling with the R package brms. ArXiv Prepr. ArXiv170511123 .

28. Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. J. Stat. Softw. 76.

29. Brooks, S.P., and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. 7, 434–455.

30. Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432.

31. Köster, J., and Rahmann, S. (2012). Snakemake - a scalable bioinformatics workflow engine. Bioinformatics 28, 2520–2522.

32. Kurtzer, G.M., Sochat, V., and Bauer, M.W. (2017). Singularity: Scientific containers for mobility of compute. PloS One 12, e0177459.

Table 1: Baseline descriptive variables stratified by applicability of LOCF; LOCF is not applicable when no GOSe observation prior to 180 days is available. N is the number of non-missing values. P values are based on the Chi-squared test for binary variables and on the Wilcoxon test for continuous variables.

	N	no LOCF possible (N=118)	LOCF possible (N=3225)	
Age	3343			P=0.46
Median (interquartile range)		47.0 (29.0—61.1)	49.0 (29.0—64.0)	
Range		7.0—90.0	0.0—95.0	
Sex: Male	3343	65/118 (55.085)	2145/3225 (66.512)	P=0.01
Stratum	3343			P=0.87
Emergency Room		22/118 (18.644)	664/3225 (20.589)	
Admission to Hospital		43/118 (36.441)	1155/3225 (35.814)	
Intensive Care Unit		53/118 (44.915)	1406/3225 (43.597)	
Cause of Injury	3334			P=0.11
Road traffic incident		57/118 (48.305)	1250/3216 (38.868)	
Incidental fall		38/118 (32.203)	1440/3216 (44.776)	
Other		14/118 (11.864)	316/3216 (9.826)	
Violence/assault		7/118 (5.932)	146/3216 (4.540)	
Unknown		2/118 (1.695)	64/3216 (1.990)	
ISS, total	3305			P=0.39
Median (interquartile range)		16 (9—27)	16 (9—26)	
Range		1—75	1—75	
GCS	3236			P=0.62
Mild		82/115 (71.304)	2297/3121 (73.598)	
Moderate		8/115 (6.957)	252/3121 (8.074)	
Severe		25/115 (21.739)	572/3121 (18.327)	
Marshall CT	3030			P=0.87
1		49/106 (46.2264)	1206/2924 (41.2449)	
2		40/106 (37.7358)	1230/2924 (42.0657)	
3		3/106 (2.8302)	82/2924 (2.8044)	
4		0/106 (0.0000)	16/2924 (0.5472)	
5		0/106 (0.0000)	6/2924 (0.2052)	
6		14/106 (13.2075)	384/2924 (13.1327)	
Subarachnoid Hematoma: yes	3262	41/115 (35.652)	1131/3147 (35.939)	P=0.95
Extradural Hematoma: yes	3243	7/113 (6.1947)	356/3130 (11.3738)	P=0.09
Hypoxia: yes	3167	7/109 (6.4220)	170/3058 (5.5592)	P=0.70
Hypotension: yes	3193	6/110 (5.4545)	178/3083 (5.7736)	P=0.89
Glucose [mmol/L]	2548			P=0.90
Median (interquartile range)		6.8 (5.9—8.3)	6.9 (5.9—8.2)	
Range		3.7—15.7	1.9—33.5	
Hemoglobin [g/dL]	2802			P=0.67
Median (interquartile range)		13.6 (12.4—14.6)	13.5 (12.0—14.6)	
Range		8.1—17.1	1.3—23.4	