## Title

Informatics tools to assess the success of procedural harmonization in preclinical multicenter biomarker discovery study on post-traumatic epileptogenesis



## Authors

Ciszek, Robert
Ndode-Ekane, Xavier Ekolle
Gomez, Cesar Santana
et al.

## Publication Date

2019-02-01

## DOI

10.1016/j.eplepsyres.2018.12.010

Peer reviewed

Published in final edited form as:

*Epilepsy Res*. 2019 February ; 150: 17–26. doi:10.1016/j.eplepsyres.2018.12.010.

# Informatics tools to assess the success of procedural harmonization in preclinical multicenter biomarker discovery study on post-traumatic epileptogenesis⋆

**Robert Ciszek**[a,*], **Xavier Ekolle Ndode-Ekane**[a], **Cesar Santana Gomez**[b], **Pablo M. Casillas-Espinosa**[c,d], **Idrish Ali**[c,d], **Gregory Smith**[b], **Noora Puhakka**[a], **Niina Lapinlampi**[a], **Pedro Andrade**[a], **Alaa Kamnaksh**[e], **Riikka Immonen**[a], **Tomi Paananen**[a], **Matthew R. Hudson**[c,d], **Rhys D. Brady**[c,d], **Sandy R. Shultz**[c], **Terence J. O'Brien**[b,d,f,g], **Richard J. Staba**[b], **Jussi Tohka**[a], **Asla Pitkänen**[a]

[a]A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Kuopio, Finland [b]Department of Neurology, David Geffen School of Medicine at UCLA, Los Angeles, CA, USA [c]The Department of Neuroscience, Central Clinical School, Monash University, Melbourne, Australia [d]Department of Medicine, The Royal Melbourne Hospital, The University of Melbourne, Victoria, 3052, Australia [e]Department of Anatomy, Physiology and Genetics, Uniformed Services University, MD, USA [f]Department of Neurology, The Alfred Hospital, Commercial Road, Melbourne, Victoria, 3004, Australia [g]Department of Neurology, The Royal Melbourne Hospital, Grattan Street, Parkville, Victoria, 3050, Australia

## Abstract

The Epilepsy Bioinformatics Study for Antiepileptogenic Therapy (EpiBioS4Rx) is a National Institutes for Neurological Diseases and Stoke funded Centers-Without-Walls international multidisciplinary study aimed at preventing epileptogenesis. The preclinical biomarker discovery in EpiBios4Rx applies a multicenter study design to allow the number of animals that are required for adequate statistical power for the analysis to be studied in an efficient manner. Further, the use of multiple centers mimics the clinical trial situation, and therefore potentially the chance of successful clinical translation of the outcomes of the study. Its successful implementation requires harmonization of procedures and data analyses between the three contributing centers in Finland, Australia, and USA. The objective of the present analysis was to develop metrics for analysis of the success of harmonization of procedures to guide further data analyses and plan the future multicenter preclinical studies. The *interim* analysis of data is based on the analysis of data from 212 rats with lateral fluid-percussion injury or sham-operation included in the biomarker discovery by April 30, 2018. The details of protocols, including production of injury, post-injury follow-up, blood sampling, electroencephalogram recording, and magnetic resonance imaging have been

⋆This article is part of a special issue 'Discovery of diagnostic biomarkers for post-traumatic epileptogenesis – an interim analysis of procedures in preclinical multicenter trial EpiBios4Rx'

*Corresponding author at: A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, P.O. Box 1627 (Neulaniementie 2), FI-70211, Kuopio, Finland. ciszek@uef.fi (R. Ciszek).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.eplepsyres.2018.12.010.

presented in the accompanying manuscripts in this Supplement. Implementation of protocols in EpiBios4Rx project participant centers was visualized in 2D using t-distributed stochastic neighborhood embedding (t-SNE). The protocols applied to each rat were presented as feature vectors of procedure related variables (*e.g.*, impact pressure, anesthesia time). The total number of protocol features linked to each rat was 112. The missing data was accounted in visualization by utilizing imputation and adding the number of missing values as a third dimension to 2D t-SNE plot, resulting in a 3D overview of protocol data. Intraclass correlation coefficient (ICC) using Euclidean distances and area under receiver operating characteristic curve (AUC) of k-nearest neighbor classifier (KNN) were utilized to quantify the degree of clustering by center. Both subsets of data with incomplete protocol vectors omitted and missing protocol data imputed were assessed. Our data show that a visible clustering by center was observed in all t-SNE plots, except for day 7 neuroscores. Both ICC and AUC indicated clustering by center in all protocol variable subsets, excluding unimputed day 7 neuroscores (ICC 0.04 and AUC 0.6). ICC for imputed set of all protocol related variables was 0.1 and KNN AUC 0.92. In conclusion, both ICC and AUC indicated differences in protocol between EpiBios4Rx participating centers, which needs to be taken into account in data analysis. Importantly, the majority of observed differences are recoverable as they relate to insufficient updates in record keeping. While AUC score of KNN is a more sensitive measure for protocol harmonization than ICC for data that displays complex splintered clustering, ICC and AUC provide complementary measures to assess the degree of procedural harmonization. This experience should be helpful for other groups planning such multicenter post-traumatic epileptogenesis studies in the future.

### Keywords

Classification; Common data element; Dimensionality reduction; Intraclass correlation; k-nearest neighbor; Lateral fluid-percussion; Machine learning; Traumatic brain injury

## 1. Background

Preclinical studies have been criticized for lack of statistical power and reproducibility (Landis et al., 2012). This is also a concern in studies on epileptogenesis in rodent models of traumatic brain injury (TBI) as epileptogenesis is slow and frequency of spontaneous seizures in rats with post-traumatic epilepsy (PTE) is low (Pitkänen et al., 2017). Consequently, post-traumatic epileptogenesis studies require a long-lasting and laborious follow-up with prolonged period of video-EEG recording in order to conduct a statistically powered biomarker or treatment discovery study. Such studies are possible but challenging in a single-center study design, typically the video-EEG monitoring capacity being the bottleneck limiting the cohort size (Dongjun Guo et al., 2013; Liu et al., 2016; Nissinen et al., 2017).

The Epilepsy Bioinformatics Study for Antiepileptogenic Therapy (EpiBioS4Rx, 2018) is an international, multicenter, multidisciplinary study aimed at preventing epileptogenesis after TBI (https://epibios.loni.usc.edu/). One of the major objectives of the EpiBioS4Rx project is to find diagnostic biomarkers for post-traumatic epileptogenesis, which would also serve as predictive biomarkers in antiepileptogenesis treatment trials (Pitkänen et al., 2018). To

achieve sufficient statistical power in biomarker discovery for PTE, EpiBios4Rx is performing a preclinical multicenter study using the same rat model, lateral fluid-percussion injury (FPI), in three participating centers: University of Eastern Finland (Kuopio, Finland), Monash University (Melbourne, Australia) and University of Southern California in Los Angeles (UCLA). This approach is also expected to provide variability within the animal cohort similar to the heterogeneity of severe human closed-head TBI that can lead to PTE (Maas et al., 2017), and to more closely mimic the situation of a clinical trial – which would be multicenter – and therefore potentially increase the chance of positive translation of the outcomes. Further, multi-center trial design provides larger video-EEG monitoring capacity to achieve a high number of phenotyped animals within a feasible time window to facilitate the progress in biomarker discovery process. Moreover, multicenter study has additional advantages by reducing the risks of single-center related complications like unexpected health issues in animal colony or equipment-related technical issues.

To date, there have been no preclinical multicenter studies in epilepsy. In TBI field, Operation Brain Trauma Therapy (OBTT) has conducted a hypothesis-driven assessment of the efficacy of various therapies on post-TBI recovery in three centers, which also included biomarker analysis (Kochanek et al., 2011, 2016; Yang et al., 2018). However, OBTT applied three different models to reflect the heterogeneity of TBI, and thus, the model harmonization was not an issue. In parallel to this pioneering activity, both the TBI and epilepsy communities have been developing methodologies for harmonizing the methodologies and data collection to improve accuracy of reporting and reproducibility of experiments. The most concrete outcome being the generation of common data elements (CDEs) and case report forms (CRFs) for systematic data collection in preclinical TBI and epilepsy models (Smith et al., 2015; Lapinlampi et al., 2017; Harte-Hargrove et al., 2017).

EpiBios4Rx biomarker discovery study was started on January 15, 2017. All experimental procedures were harmonized between the three centers and procedural information was collected using project-tailored CDEs and CRFs. In the current study, we assessed the extent of harmonization for the induction of lateral FPI model production and post-injury follow-up (Ekolle Ndode-Ekane et al., 2019), blood sampling (Kamnaksh et al., 2018), seizure detection in electroencephalogram (EEG)(Casillas-Espinosa et al., 2019), detection of high-frequency oscillations (HFOs)(Santana Gomez et al., 2019), and magnetic resonance imaging (MRI)(Immonen et al., 2019). Each of these procedures is also the subject of separate articles in this special Supplementary Issue. Here we present an approach using multivariate intraclass correlation and machine learning to develop quantitative metrics for analyzing the quality of inter-center harmonization to guide data analysis. The analysis results will have implications for the current EpiBioS4Rx studies as well as planning future preclinical multicenter trials. In particular, when considering the pre—project activities needed to train the personnel to achieve sufficient procedural harmonization and for collection of preliminary data to predict inter-center procedural variability to be included in power calculations.

## 2. Materials and methods

The detailed methodologies have been presented in accompanying articles by Ekolle Ndode-Ekane et al., (2019) for injury production and post-impact follow-up by Kamnaksh et al. (2018) for blood sampling, Casillas-Espinosa et al. (2019) for EEG analysis, and Santana Gomez et al. (2019) for HFO analysis. The outline of the study design in terms of variables included in this harmonization assessment is presented in Fig. 1. All protocols were approved by institutional and/or national animal research committees.

### 2.1. The EpiBioS4Rx dataset

EpiBios4Rx protocol dataset included 189 protocol related variables collected from 212 rats at the University of Eastern Finland (UEF), Monash University in Melbourne (Melbourne), and University of Southern California in Los Angeles (UCLA). The extent of protocol harmonization was assessed in terms of variables related to TBI induction, neuroscore measurements, blood sampling, animal weight and the timing of procedures (for a complete variable list, see Supplementary Table 1). TBI induction related variables included impact pressure, impact angle, apnea duration, time to self-righting and anesthesia time. Neuroscore measurements included test results for left and right sides of the animal for contraflexion, hind limb flexion, lateral pulsion, angle board, and total neuroscore of an animal. Neuroscores were included from baseline and on days 2, 7, 14, and 28 after TBI. The set of blood absorbance features consisted of absorbance values for A and B samples. The blood absorbance variables were included from baseline, days 2 and 9 after the impact, and one month after the impact. Body weight values were measured at baseline, days 2 to 9, day 14 and day 30. Timing related features included different time points for blood sampling, anesthesia initiation and anesthesia termination. For use in Euclidean distance calculations between protocol vectors, times of the day were converted to seconds from midnight (00:00 AM).

Additionally, distances in time between selected points in the protocol were calculated. The distances were measured from start of the quarantine to the induction of injury, from the start of quarantine to baseline blood sampling, from the baseline blood sampling to the injury, from injury to day 2, day 9 and 1 month blood sampling, and from injury to MRI at day 9 and one month time points. Finally, variable denoting the mortality at the level of the cohort of each animal was added. This resulted in a feature vector consisting of 112 variables.

As none of the feature vectors contained data for all protocol related variables, separate subdatasets containing subsets of features with representative number of animals without missing values available from at least two centers were constructed. Six separate feature subsets were constructed for this purpose: neuroscore measurements at different time points, concatenation of all neuroscore measurements, variables related to blood sampling, variables related to timing, and temporal distances.

Animals that died before the latest measurement time points included in a feature subset were excluded from respective subdataset. The number of missing variables for each animal was counted. Non-numeric entries, *e.g.* “between 8–9 AM” or “midday” for time of day,

were treated as missing variables. However, during imputation described in Section 2.4, the variables were separately imputed using sensible heuristics, e.g. imputing “between 8–9 AM” with 8:30, instead of utilizing imputation algorithms. Fig. 4. illustrates the missing data profiles for the three centers.

Analysis were conducted using Python 3.6.6, scikit-learn 1.9.2, scipy 1.1.0 and R 3.5.1 on Centos 7. t-SNE (van der Maatenand Hinton, 2008) to two dimensions was performed on each of the eight EPIBIOS protocol data subset using perplexity of 50. Protocol variables subsets and the complete protocol data were additionally imputed using MICE (Van Buuren, 2007) and reduced with t-SNE to provide an overview of the level of harmonization. The analysis scripts are available from a public git repository (https://github.com/UEFepilepsyAIVI/epibios4rx-harmonization/)

Multivariate ICC using unweighted Euclidean distance was calculated for both imputed and unimputed protocol vectors. In addition, univariate ICC scores were calculated using the ANOVA approach for all protocol variables. KNN classifier with a k of 10 was used to produce AUC score for imputed and unimputed protocol vector subsets. The 95% confidence intervals for ICC and KNN AUC were estimated by bootstrapping (Ukoumunne et al., 2003) using 1000 iterations.

Mann Whitney-U test with Bonferroni correction was performed for each pair of protocol variable and pair of centers to support the observed differences in protocols. Fifty-two out of 112 protocol variables displayed statistically significant differences between centers ($p < 0.05$).

## 2.2. Visualizing protocol implementation

Given a dataset consisting $d$ protocol-related variables from $n$ animals, each entry in the dataset can be treated as a feature vector with dimensionality $d$. Such feature vector, referred as protocol vector for the rest of the article, represents the protocol implemented for a single animal. The similarities among protocol vectors can be measured using a distance metric, for example Euclidean distance. For two animals subjected to similar protocol, the distance will be near zero with differences between animals resulting only from biological variability. As the dissimilarity between the implementations of the protocol increases, the distance between vectors increases.

Visualizing the similarities among protocol vectors provides an overview of the overall level of harmonization. Similarities of protocol vectors with more than three dimensions can presented in an intuitively interpretable form by embedding the vectors into a low dimensional space. In embedding, a low dimensional presentation of the data is constructed ensuring that vectors similar to each other in the original high dimensional space are located near each other in the low dimensional embedding, while dissimilar vectors are placed far apart. Multiple embedding methods have been presented, for example, Isomap (Tenenbaum et al., 2000) and Laplacian eigenmaps (a.k.a. spectral embedding) (Belkin and Niyogi, 2003). The most suitable embedding method for strictly visualization purposes is t-distributed stochastic neighborhood embedding (t-SNE) (van der Maaten and Hinton, 2008). t-SNE models the similarity of samples in high dimensional space as conditional

probabilities using normal distributions centered on each sample. Similar samples are given high probabilities, and the probabilities decrease quickly due to the shape of the distribution as the dissimilarity between data points increases. The similarities of data points in the low dimensional space are modelled using Student's t-distribution, and the embedding is performed by minimizing the Kull-back-Leibler divergence between the two distributions. t-SNE fills efficiently the low dimensional space, that is, it minimizes the empty space on a plot, for presenting the similarities of points in low dimensional space and retains the clustering among samples during embedding.

The low dimensional representation may not be able to faithfully present the all relationships between the vectors of the original high-dimensional data, as compromises in terms of distances must be made when the data is squeezed from high to low dimensional space. For visualization purposes t-SNE emphasizes the clustering present in the data and with low values of t-SNEs perplexity parameter, which controls the size of the fitted distributions, the algorithm can result in a reduction displaying exaggerated granularity. This would cause points within each cluster to additionally aggregate into small adjacent granules in the visualization. Additionally, an artificially low perplexity may result in a misleadingly heterogenous reduction by scattering the points from centers with a very little variance in protocol. In the analysis conducted, perplexity of 50 was utilized in all reductions. It should be also noted that if vectors belonging to the same high-dimensional manifold cannot be presented as a continuous shape in the low dimensional embedding, a cluster of points may be broken to two or more separate clusters in the embedding. Nevertheless, as the neighborhoods are retained in the embedding, the proximity of points in the embedding implies proximity in the original high dimensional space. By plotting the embedded feature vectors and by coloring the points by center the level of harmonization can be quickly assessed. In an ideal case (Fig. 2A), no clustering by site should be visible in the plot.

### 2.3. Quantifying center similarity

Humans have an aptitude for detecting patterns and shapes, whether they exist or not. To quantify the degree of clustering observed visually in the reductions, numerical measures for the similarity among clusters can be calculated. Two intuitive measures exist for the clustering by center. First, protocol vectors are known to be clustered by center if the between center variance is higher than the within center variance (Fig. 2B). Secondly, from clustering by center it follows that each protocol vector's adjacent vectors belong to the same center. The former can be quantified using intraclass correlation coefficient and the latter by center-wise classification performance utilizing machine learning.

Intraclass correlation (ICC), first introduced by Fisher (1950) as a modification of Pearson correlation coefficient, measures the ratio by which the variability in the data is explained by clustering. Assume a dataset containing $n$ samples from $k$ centers, with $n_k$ animals from each center. Let $\acute{x}$ denote the mean of a variable $x$ over the while data and let $\bar{x}_k$ denote the mean of $x$ for samples from center $k$. Using ANOVA framework, ICC is defined as (Wu et al., 2012):

$$\rho = \frac{MSB - MSW}{MSB + MSW(nA - 1)} \quad (1)$$

where $MSW = \frac{\sum_i^k \left( \sum_j^{n_k} x_j - \overline{x_i} \right)^2}{n - 1}$, $MSB = \frac{\sum_i^k \left( \bar{x} - \bar{x}_i \right)^2}{k - 1}$ and $nA = \frac{1}{k - 1}$

ICC is a univariate measure for correlation of vectors within each center. The idea of ICC can however be applied in the context of multivariate feature vectors by utilizing distance metrics. Within group variability is then measured by the mean distance to the cluster center in the high dimensional feature space, and between group similarity by the mean distance of group centroids to the center of the feature space. This approach results in a definition of ICC as

$$\rho_d = \frac{d_b - d_w}{d_b + d_w(nA - 1)}$$

where $d_w = \sum_i^k \sum_j^{n_k} d\left(x_i, c_j\right)^2$, $d_b = \sum_i^k d\left(x_i, \bar{x}\right)^2$ and d (*a*, *b*) is some distance metric (*e.g.* Euclidean distance).

As with univariate ICC, 0 indicates similarity among centers, *i.e.* perfect harmonization, and values near 1 clear separation into different clusters. Weighted Euclidean distance can be used to give different weight for each protocol related variables.

K-nearest neighbor classification (KNN) is a non-parametric machine learning algorithm that assigns a class to each vector based on the classes of the vector's *k* "nearest" – *e.g.* the most similar, vectors (Duda et al., 2016). To assess the degree of aggregation by center, leave-one-out cross-validation is utilized. In leave-one-out cross-validation, a single feature vector is held out from the data and rest of the data is used to assign a center to the held-out vector by calculating the most common center among the k-nearest protocol vectors. This process is repeated on for all protocol vectors. By comparing the resulting classification to known real centers, multiclass micro-averaged area under curve (AUC)(Flach and Ferri, 2011) score can be calculated. Micro-averaged AUC is calculated by aggregating the classification performances scores (number of true positives, true negatives *etc.*) from each class and calculating the AUC using the combined scores. AUC score near 0.5 will indicate dispersion of the points in the feature space (proper harmonization) and value near 1 perfect separation of points by center. As with ICC, weighted Euclidean distances can be utilized with KNN to weight protocol variables.

ICC is a global measure of protocol vector aggregation. It denotes solely the level correlation within centers, and in special cases, low ICC can be reached with data with significant center-wise clustering. Example of such scenario is splintering of data from each center into intermingled sub-clusters. If such splintered clusters are suitably located in the feature space, the overall ICC of the data may remain low despite clear clustering by center. Conversely, KNN AUC is a local measure of aggregation by center. It is sensitive to

splintering of vectors to small granules and to small differences in the centers. Example of the latter behavior is shown in Fig. 3, which illustrates t-SNE reductions of a synthetic data with 128-dimensional protocol vectors and three centers. All protocol variables were drawn from a normal distribution. A near ideal scenario is displayed in Fig. 3A, where all centers have equal means and differ in variance. Both the ICC and AUC indicate perfect harmonization. In Fig. 3B, centers have equal variances, but the means are placed a variance apart. ICC reports very little correlation within centers, which is visually explained by the number of points from each center mingled among points of other centers. The KNN AUC however is high, implying separation in terms of protocol. Fig. 3C displays a scenario with center means unit variance apart. As expected both AUC and ICC reach high values. In the last synthetic scenario (Fig. 3D), both the means and variances of each center differ, but only two of the centers (center 2 and 3) are partially mixed together. As a result, the micro-averaged AUC is dominated by the separability of the less overlapped centers (centers 1 and 2) and reports no clustering, while ICC indicates clear clustering by center.

### 2.4. Handling missing data

To be able to measure the similarity of two protocol vectors, no variables can be missing from neither of the vectors. Missing variables in the protocol data can be dealt either by sub-setting the data into small sets of features within which a representative set of animals with intact data is available from all centers. Alternatively, the missing values for a vector can be imputed by filling the values based on the existing observations in the vector. In multiple imputation by chained equations (MICE), regression models are iteratively inferred based on the existing values and missing data is imputed using the model predictions (Van Buuren, 2007). This provides a reasonable guess for the possible values and allows generation of an overview of the data in cases where omitting incomplete feature vectors would shrink the data to unrepresentative size.

### 2.5. Visualizing missing data

The number of missing values in an imputed reduction can be presented in a visualization by, for example, reducing the protocol data to two dimensions and adding a third dimension to represent the missing data count (Figs. 5 and 6). This enables quick assessment of both clustering by center, amount of missing data and the reliability of the assignment of points to the vicinity of each other.

## 3. Results

2D t-SNE reductions for intact subsets of protocol variables are presented in Fig. 5. No clear clustering by center was evident from the reduction of lateral FPI impact-related variables, but the points from the same cluster appeared to form small string-like groups (Fig. 5A). In blood related variables similarity among animals from UCLA and some mixing between UEF and Melbourne can be observed (Fig. 5B). Clear separation of UCLA from UEF and Melbourne can be seen in day 14 neuroscores (Fig. 5C).

In Fig. 6A–F, 3D plots of t-SNE reductions protocol vectors with missing data imputed are presented. The vertical axis in the 3D figures presents the number of imputed missing

variables. The number of missing variables contributes significantly to the clustering evident from the reduction of complete set of protocol variables and the subsets (Fig. 6A). The 3D plot with imputed blood sampling variables resembles the 2D t-SNE without imputation, with a single center appearing separated due to missing data (Fig. 6B). Clustering by vector similarity can be seen *e.g.* in combined neuroscores, where UCLA and UEF are clearly separated (Fig. 6C). In terms distance between procedure time points (Fig. 6D), Melbourne forms a cluster of its own. Majority of UEF protocol vectors are distributed to small clusters separate from Melbourne and UCLA. Lateral FPI (Fig. 6E) divides protocol data into two clusters. The smaller compact cluster appears mixed, while within the larger cluster some clustering by center is visible. In terms of weight, the centers appear to form a single mixed elongated cluster (Fig. 6F). As a large portion of the data visualized in Fig. 6F consisted of imputed variables, the apparent similarity of UEF, UCLA and Melbourne in terms of weight is questionable. The clustering observed in Fig.6 is most evident in the supplementary interactive 3D plots, which can be freely rotated (Supplementary Fig 1–13).

Tables 1 and 2 present the ICC values for protocol variable subsets with and without imputation, respectively. The overall protocol ICC calculated using imputed set of all protocol variables data was 0.08, which implies correlation among samples within the same center (Table 1). Unimputed baseline and day 14, 21 and 28 neuroscores show ICC above 0.3, which likewise implies correlation within center (Table 2). KNN AUC scores for both imputed (Table 3) and unimputed (Table 4) data similarly indicate clustering by center. AUC score of 0.75 or higher was reached on every protocol variable subset, except for day 7 neuroscores. While ICC for the overall protocol in the imputed dataset was only 0.08 (Table 1), the respective KNN AUC was 0.93 (Table 3), implying strong tendency to form clusters by center. This conclusion is supported by the tSNE presented in Fig. 6. Although the ICC score 0.01 for timing (Table 1) appears near ideal, the AUC score 0.87 for the same subset (Table 3) is high. This can be explained by the positioning of the vectors in the feature space (Fig. 7), resulting from the imputed feature vectors for Melbourne and UCLA consisting of near identical values.

## 4. Discussion

We presented methods for visualization and quantification the level of procedural harmonization in preclinical trials. We visualized protocol similarity using t-SNE, and data integrity in a 3D plot by adding a third axis to a 2D protocol vector embedding to denote the amount of missing data.

ICC is an intuitive measure of the amount of procedural variance resulting from between center variance. However, for non-normally distributed data and in the presence of significant outliers, ICC may produce misleadingly low values. AUC score of KNN on the other hand is insensitive to the shape of distribution but requires the parameter $k$ for the number of neighbors to be set on case-by-case basis. As global and local measures of center-wise clustering ICC and AUC can be considered complementary. A high ICC combined with a high AUC implies clear separation between centers, whereas low to moderate ICC and moderate to high AUC implies center-wise clustering with some mixing between centers. Further, the distribution of the data may obscure the center-wise clustering in terms of a

single measure, but by combining both measures clustering can still be detected. Nevertheless, of the two measures the results presented indicate KNN AUC to be more straightforwardly interpretable measure for the success of procedural harmonization. In contrast to ICC, KNN AUC was shown to more often produce scores in line with the separation visible from t-SNE plots. Any machine learning algorithm could be applied instead of KNN in harmonization evaluation, as the purpose of machine learning in this context is merely to assess the center-wise clustering through the difficulty of assigning protocols to centers.

EpiBios4Rx protocol data displayed clear clustering by center both in the case of unimputed subdatasets and imputed complete protocol data. This suggests that assessment made using the imputed complete protocol data is not significantly distorted by imputation. Differences in protocol manifesting as significant center-wise clustering must be taken account during subsequent analysis. Majority of statistical analyses and sample size calculations assume statistical independence across subjects. Assumption of independence is violated when measurements are clustered by center, which can result in incorrect p-values and confidence intervals (Localio et al., 2001). When center-wise clustering is observed, suitable models accounting for the center-effect (*e.g.*, mixed linear models or generalized estimating equations) should be utilized (Kahan, 2014).

About 50% of EpiBioS4Rx data consisted of missing variables, such as notes related to blood sampling or neuroscores, even though the tests had been performed. In the context of protocol harmonization, missing data forms one dimension and must be considered separately. Protocol containing a large ratio of inconsistently missing data cannot be considered harmonized as differing patterns of missing data indicate lack of consistency in recording of practices between the centers. However, missing data affects ICC or AUC only indirectly as entries containing missing values are either omitted or the missing values are imputed prior the calculations. In both cases, the steps taken to account for missing data can distort the indexes towards either failure or success in harmonization.

Observed differences in protocols can result from recoverable or unrecoverable deviations. For example, missing data caused by inadequate record keeping is recoverable and can be fixed by increasing the efficiency in documentation even retrospectively. In contrast, omitted procedures or differences in conducted procedures are un-recoverable. Majority of missing EpiBioS4Rx data falls into the recoverable category.

Use of electrical data capture systems - such as REDCap (Harris et al., 2009), can largely eliminate the discrepancy between protocols stemming from missing or malformed data. REDCap will be utilized in the next stages of the EpiBioS4Rx, and the centers are expected to converge in terms of protocol harmonization in the near future.

Our *interim* assessment of EpiBioS4Rx protocols emphasizes the importance of early analysis of harmonization success in order to specify the parameters, which contribute to statistical indicators of incomplete harmonization. Using this information, it will be possible to find and apply target-specific remedy to alleviate or cure the incomplete harmonization,

for example, by increasing the monitoring of consistent book-keeping, by performing continuous training of investigators, and by facilitating communication between the centers.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgements

This work was supported by the National Institute of Neurological Disorders and Stroke (NINDS) Centers without Walls [grant number U54 NS100064].

## References

Belkin Mikhail, Niyogi Partha, 2003 Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15, 1373–1396.

Casillas-Espinosa PM, de Abreu PA, Santana-Gomez C, Paananen T, Smith G, Peruca P, Ali I, Ciszek R, Ndode-Ekane XE, Immonen R, Puhakka N, Staba RJ, Pitkänen A, O'Brien TJ, 2019 Harmonization of pipeline for automated seizure detection for phenotyping of post-traumatic epilepsy in a preclinical multicenter study on post-traumatic epileptogenesis. Epilepsy Res.

Duda RO, Hart PE, Stork DG, 2016 Pattern Classification, 3. revised edition ed. Wiley, New York [u.a.].

Ekolle Ndode-Ekane X, Gomez CS, Casillas-Espinosa PM, Ali I, Smith G, de Abreu PA, Immonen R, Puhakka N, Hudson MR, Brady RD, Shultz SR, Staba RJ, O'Brien TJ, Pitkänen A, 2019 Harmonization of lateral fluid-percussion injury model production and post-injury monitoring in a preclinical multicenter biomarker discovery study on post-traumatic epileptogenesis. Epilepsy Res.

EpiBioS4Rx, https://epibios.loni.usc.edu/, Accessed 2 November 2018.

Fisher RA, 1950 Statistical Methods for Research Workers, 11. ed., rev. ed. Oliver and Boyd, Edinburgh [u.a.].

Flach P, Ferri C, 2011 A Coherent Interpretation of AUC as a Measure of Aggregated Classification Performance.

Guo Dongjun, Zeng Linghui, Brody David L., Wong Michael, 2013 Rapamycin attenuates the development of posttraumatic epilepsy in a mouse model of traumatic brain injury. PLoS One 8, e64078. [PubMed: 23691153]

Harte-Hargrove LC, French JA, Pitkänen A, Galanopoulou AS, Whittemore V, Scharfman HE, 2017 Common data elements for preclinical epilepsy research: standards for data collection and reporting. A TASK3 report of the AES/ILAE Translational Task Force of the ILAE. Epilepsia 58, 78–86. [PubMed: 29105074]

Immonen R, Smith G, Brady RD, Wright D, Johnston L, Harris NG, Manninen E, Branch C, Ndode-Ekane XE, Santana Gomez C, Casillas-Espinosa PM, Shultz SR, Ali I, Jones NC, de Abreu PA, Puhakka N, Cabeen R, Duncan D, Staba RJ, O'Brien T, Toga AW, Pitkänen A, Gröhn O, 2019 Harmonization of pipeline for preclinical multicenter MRI biomarker discovery in a rat model of post-traumatic epileptogenesis. Epilepsy Res.

Kahan BC, 2014 Accounting for centre-effects in multicentre trials with a binary outcome - when, why, and how? BMC Med. Res. Methodol 14, 20. [PubMed: 24512175]

Kamnaksh A, Puhakka N, Ali I, Smith G, Aniceto R, McCullough J, Ndode-Ekane XE, Brady R, Casillas-Espinosa P, Gomez CS, Immonen R, de Abreu PA, Jones N, Shultz S, Staba RJ, O'Brien TJ, Agoston D, Pitkänen A, 2018 Harmonization of pipeline for preclinical multicenter plasma protein and miRNA biomarker discovery in a rat model of post-traumatic epileptogenesis. Epilepsy Res.

Kochanek PM, Bramlett H, Dietrich WD, Dixon CE, Hayes RL, Povlishock J, Tortella FC, Wang KKW, 2011 A novel multicenter preclinical drug screening and biomarker consortium for experimental traumatic brain injury: operation brain trauma therapy. J. Trauma Inj. Infect. Crit. Care 71, S24.

Kochanek PM, Bramlett HM, Dixon CE, Shear DA, Dietrich WD, Schmid KE, Mondello S, Wang KKW, Hayes RL, Povlishock JT, Tortella FC, 2016 Approach to modeling, therapy evaluation, drug selection, and biomarker assessments for a multicenter pre-clinical drug screening consortium for acute therapies in severe traumatic brain injury: operation brain trauma therapy. J. Neurotrauma 33, 513–522. [PubMed: 26439468]

Landis Story C., Amara Susan G., Asadullah Khusru, Austin Chris P., Blumenstein Robi, Bradley Eileen W., Crystal Ronald G., Darnell Robert B., Ferrante Robert J., Fillit Howard, Finkelstein Robert, Fisher Marc, Gendelman Howard E., Golub Robert M., Goudreau John L., Gross Robert A., Gubitz Amelie K., Hesterlee Sharon E., Howells David W., Huguenard John, Kelner Katrina, Koroshetz Walter, Krainc Dimitri, Lazic Stanley E., Levine Michael S., Macleod Malcolm R., McCall John M., Moxley III Richard T., Narasimhan Kalyani, Noble Linda J., Perrin Steve, Porter John D., Steward Oswald, Unger Ellis, Utz Ursula, Silberberg Shai D., 2012 A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490, 187–191. [PubMed: 23060188]

Lapinlampi N, Gorter JA, Bernard C, Löscher W, Kokaia M, Lukasiuk K, Roncon P, Aronica E, Bankstahl JP, Melin E, Simonato M, Ravizza T, Pitkänen A, Gröhn O, Paananen J, Vezzani A, Lipsanen A, Becker A, 2017 Common data elements and data management: remedy to cure underpowered preclinical studies. Epilepsy Res. 129, 87–90. [PubMed: 28038337]

Liu S, Zheng P, Wright DK, Dezsi G, Braine E, Nguyen T, Corcoran NM, Johnston LA, Hovens CM, Mayo JN, Hudson M, Shultz SR, Jones NC, O'Brien TJ, 2016 Sodium selenate retards epileptogenesis in acquired epilepsy models reversing changes in protein phosphatase 2A and hyperphosphorylated tau. Brain 139, 1919–1938. [PubMed: 27289302]

Localio A. Russell, Berlin Jesse A., Ten Have Thomas R., Kimmel Stephen E., 2001 Adjustments for center in multicenter studies: an overview. Ann. Intern. Med 135, 112. [PubMed: 11453711]

Maas AIR, Menon DK, Adelson PD, Andelic N, Bell MJ, Belli A, Bragge P, Brazinova A, Büki A, Chesnut RM, Citerio G, Coburn M, Cooper DJ, Crowder AT, Czeiter E, Czosnyka M, Diaz-Arrastia R, Dreier JP, Duhaime A, Ercole A, van Essen TA, Feigin VL, Gao G, Giacino J, Gonzalez-Lara LE, Gruen RL, Gupta D, Hartings JA, Hill S, Jiang J, Ketharanathan N, Kompanje EJO, Lanyon L, Laureys S, Lecky F, Levin H, Lingsma HF, Maegele M, Majdan M, Manley G, Marsteller J, Mascia L, McFadyen B, McFadyen C, Mondello S, Newcombe V, Palotie A, Parizel PM, Peul W, Piercy J, Polinder S, Puybasset L, Rasmussen TE, Rossaint R, Smielewski P, Söderberg J, Stanworth SJ, Stein MB, von Steinbüchel N, Stewart W, Steyerberg EW, Stocchetti N, Synnot A, Te Ao B, Tenovuo O, Theadom A, Tibboel D, Videtta W, Wang KKW, Williams WH, Williams G, Wilson L, Yaffe K, Adams H, Agnoletti V, Allanson J, Amrein K, Andaluz N, Anke A, Antoni A, van As AB, Audibert G, Azaševac A, Azouvi P, Azzolini ML, Baciu C, Badenes R, Barlow KM, Bartels R, Bauerfeind U, Beauchamp M, Beer D, Beer R, Belda FJ, Bellander B, Bellier R, Benali H, Benard T, Beqiri V, Beretta L, et al., 2017 Traumatic brain injury: integrated approaches to improve prevention, clinical care, and research. Lancet Neurol. 16, 987–1048. [PubMed: 29122524]

Nissinen J, Andrade P, Natunen T, Hiltunen M, Malm T, Kanninen K, Soares JI, Shatillo O, Sallinen J, Ndode-Ekane XE, Pitkänen A, 2017 Disease-modifying effect of atipamezole in a model of post-traumatic epilepsy. Epilepsy Res. 136, 18–34. [PubMed: 28753497]

Pitkänen A, Kyyriäinen J, Andrade P, Pasanen L, Ndode-Ekane XE, 2017 Chapter 45 - epilepsy after traumatic brain injury Models of Seizures and Epilepsy, second edition pp. 661–681.

Pitkänen A, Ekolle Ndode-Ekane X, Lapinlampi N, Puhakka N, 2018 Epilepsy biomarkers – toward etiology and pathology specificity. Neurobiol. Dis in press.

Santana Gomez C, Andrade P, Hudson M, Paananen T, Ciszek R, Smith G, Ali I, Ndode-Ekane XE, Casillas-Espinosa PM, Immonen R, Puhakka N, Jones N, Perucca P, Pitkänen A, O'Brien TJ, Staba R, 2019 Harmonization of pipeline for detection of HFOs in a rat model of post-traumatic epilepsy in preclinical multicenter study on post-traumatic epileptogenesis. Epilepsy Res.

Smith DH, Hicks RR, Johnson VE, Bergstrom DA, Cummings DM, Noble LJ, Hovda D, Whalen M, Ahlers ST, LaPlaca M, Tortella FC, Duhaime A, Dixon CE, 2015 Pre-clinical traumatic brain injury common data elements: toward a common language across laboratories. J. Neurotrauma 32, 1725–1735. [PubMed: 26058402]

Tenenbaum Joshua B., de Silva Vin, Langford John C., 2000 A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323. [PubMed: 11125149]

Ukoumunne OC, Davison AC, Gulliford MC, Chinn S, 2003 Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. Stat. Med 22, 3805–3821. [PubMed: 14673940]

Van Buuren S, 2007 Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res 16, 219–242. [PubMed: 17621469]

van der Maaten LJP, Hinton GE, 2008 Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res 9, 2579–2605.

Wu S, Crespi CM, Wong WK, 2012 Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials. Contemp. Clin. Trials 33, 869–880. [PubMed: 22627076]

Yang Z, Zhu T, Mondello S, Akel M, Wong AT, Kothari IM, Lin F, Shear DA, Gilsdorf JS, Leung LY, Bramlett HM, Dixon CE, Dietrich WD, Hayes RL, Povlishock JT, Tortella FC, Kochanek PM, Wang KKW, 2018 Serum-based phospho-neurofilament-heavy protein as theranostic biomarker in three models of traumatic brain injury: an operation brain trauma therapy study. J. Neurotrauma 36 (19).

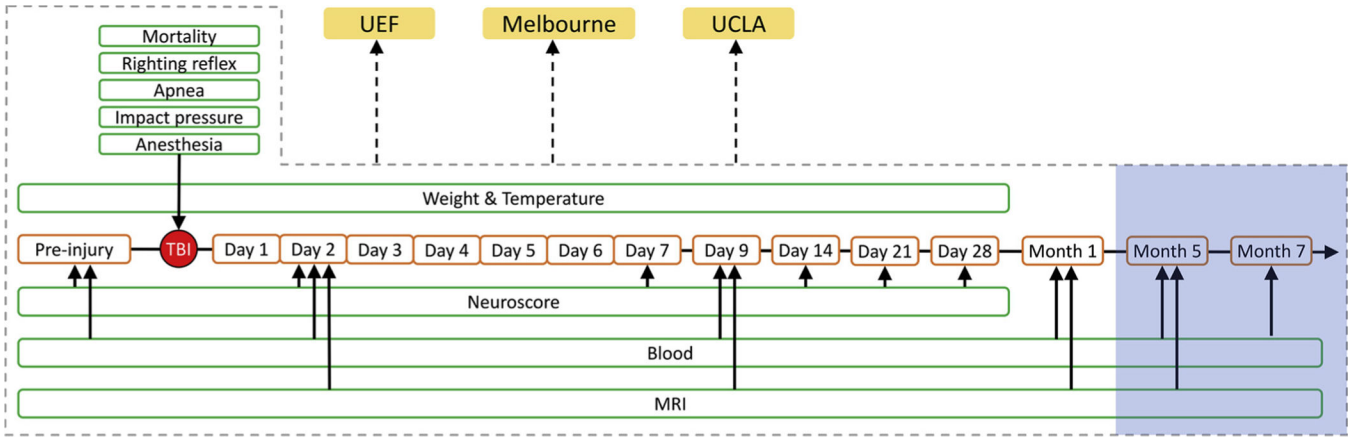


**Fig. 1.**

Outline of the EpiBioS4rx study design and sampling time of parameters included in the present analysis. Weight and temperature were measured at baseline (pre-injury), and thereafter, on days (d) 1 to d7, d9, d14, d21 and d28. Neuroscore was measured at baseline, and on d1–d7, d14, d21 and d28. Blood sampling was performed at baseline, and thereafter, on d2, d9, and at 1, 5 and 7 months post-injury. Note that data collected after the 1st post-injury month (shaded area) was not included in the present analysis.

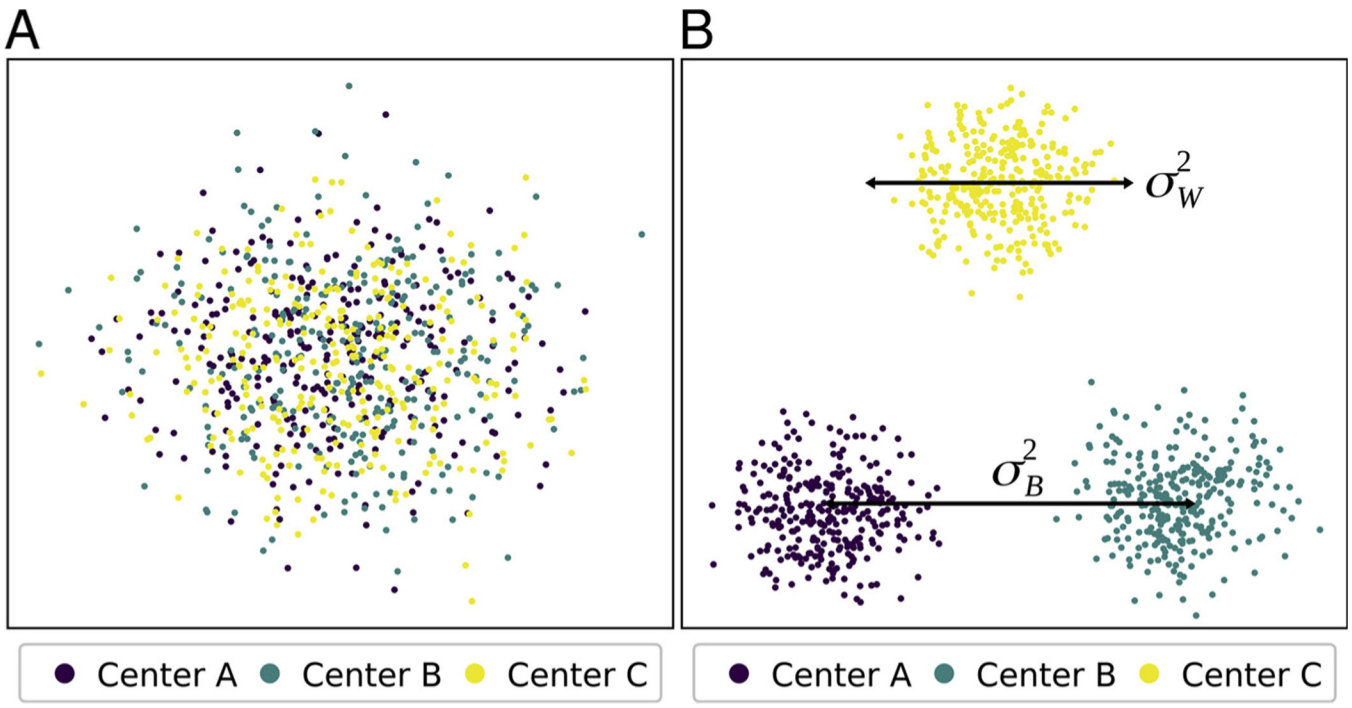


**Fig. 2.**

Ideal harmonization and illustration of between and within center distances. (**A**) A scatter plot illustrating an ideal situation where protocol vectors from all centers are dispersed with no clustering by center. (**B**) The ratio of between center variance $\sigma_b^2$ to within center variance $\sigma_w^2$ is an intuitive measure for clustering by center. When the between center variance decreases, the scattering of points in panel B approaches that of panel A.

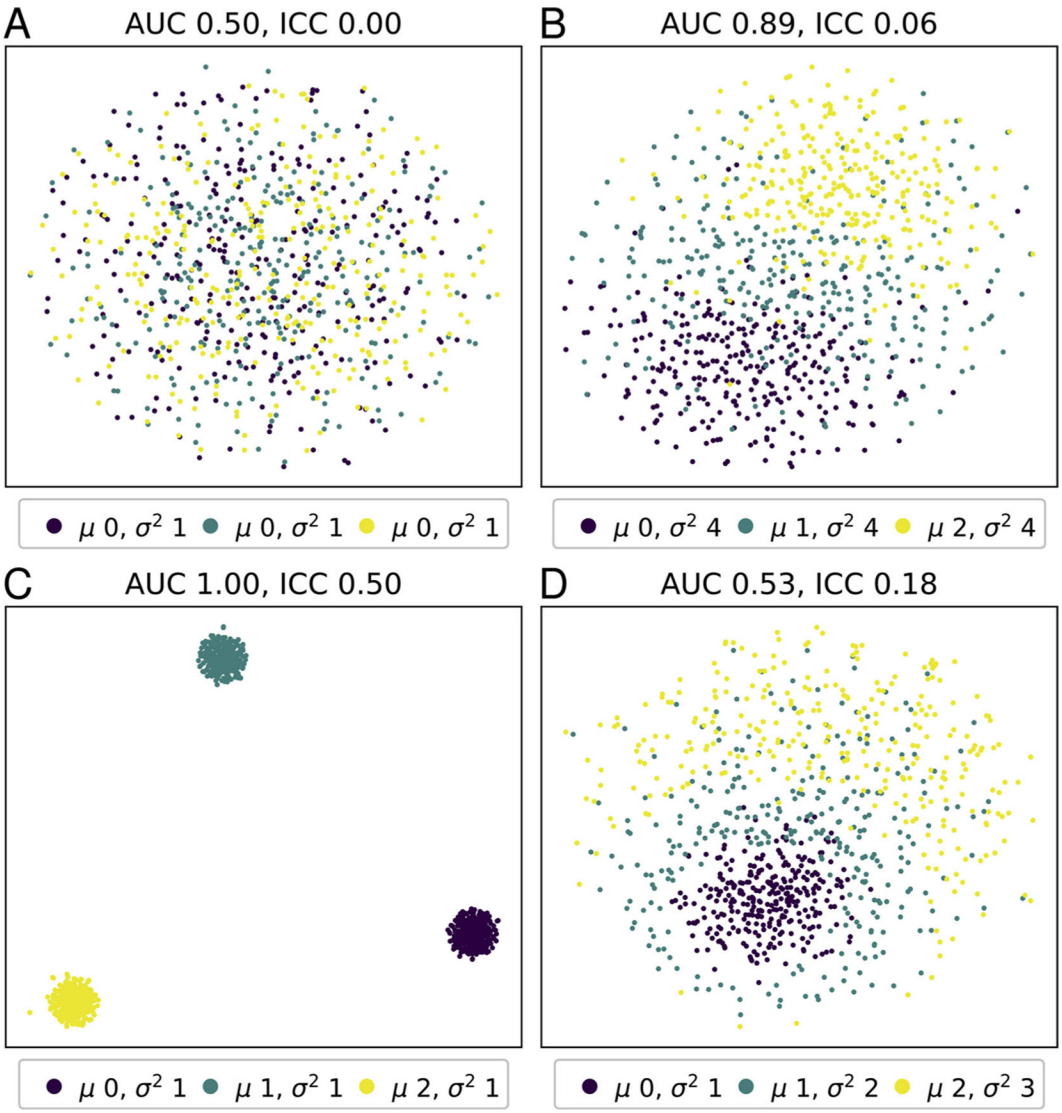


**Fig. 3.**

t-SNE reductions of synthetic data consisting of 128-dimensional protocol vectors drawn from normal distribution. (**A**) Near-ideal situation with centers having equal means and equal variances. (**B**) Slight difference in means results in high AUC score but only a small increase in ICC. (**C**) Clear difference in means results in both high AUC and ICC scores. Note that t-SNE exaggerates the distances between clusters but preserves the neighborhood relationships between similar protocol vectors. (**D**) An opposite case as presented in panel B. With suitably chosen different means and variances for each center, AUC reports poor separation into clusters while and correctly ICC implies tendency to form center-wise clusters.

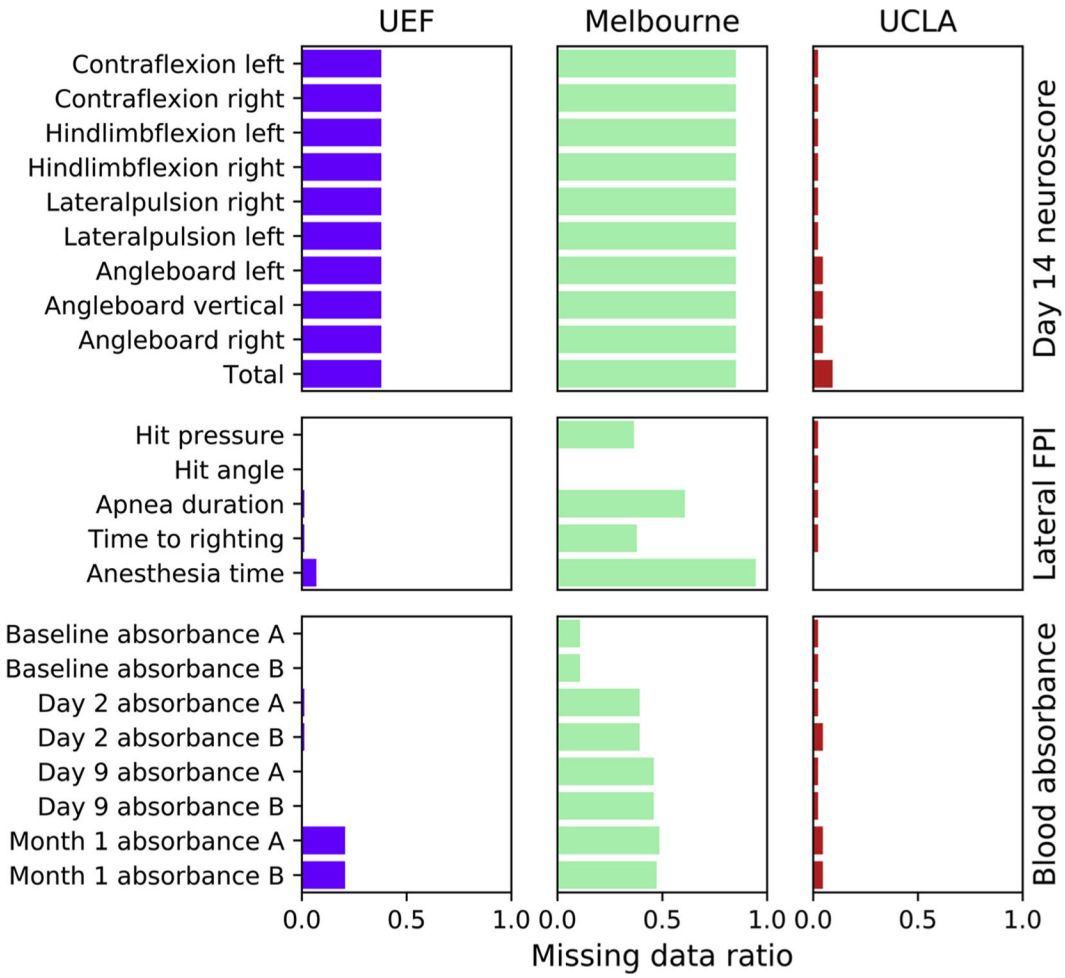


**Fig. 4.**

The ratio of missing entries to total entries in feature sets containing variables related to day 14 neuroscores (top), injury induction (middle), and blood sampling (bottom). Majority of day 14 neuroscore measurements from Melbourne are missing (*i.e.*, ratios near 1.0), while neuroscore dataset from UCLA is mostly intact (ratios near 0). Similarly, ratios for injury induction and blood sampling-related variables from UCLA and UEF indicate very few missing entries, while larger ratios from Melbourne indicate several missing entries.

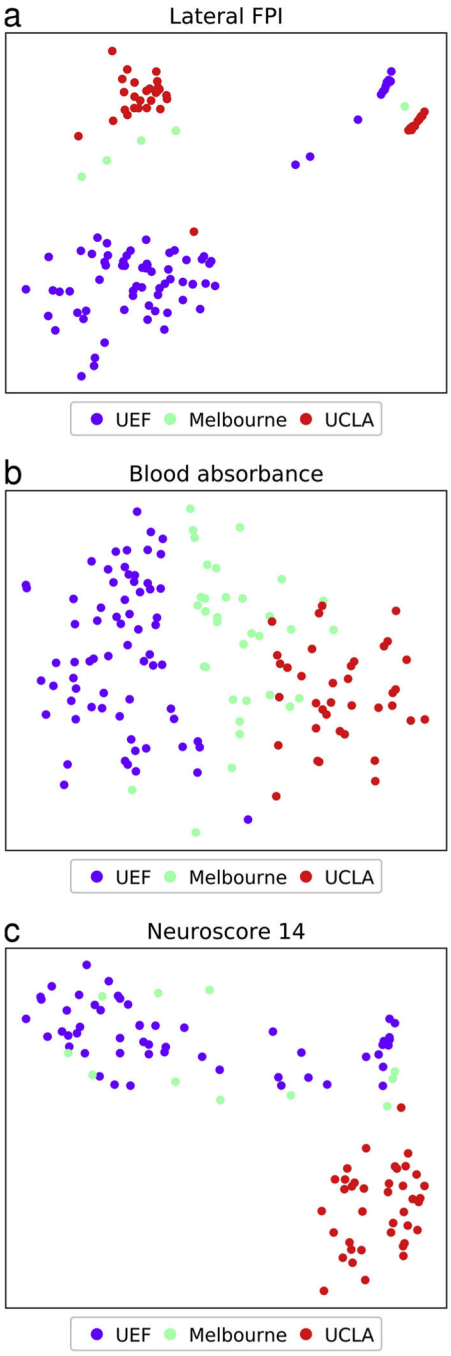


**Fig. 5.**

t-SNE reductions of subsets of protocol variables containing intact protocol vectors. (**A**) Lateral fluid-percussion injury (FPI) -related variables show clear clustering by center. **B**) Blood absorbance measurements from Melbourne and UCLA appear somewhat similar. Measurements from UEF aggregate into a group of their own. The three centers form layers on a single large cluster, indicating clear similarity within centers but nevertheless small distance between centers. (**C**) In day 14 neuroscore measurents UCLA clearly differs from UEF and Melbourne.

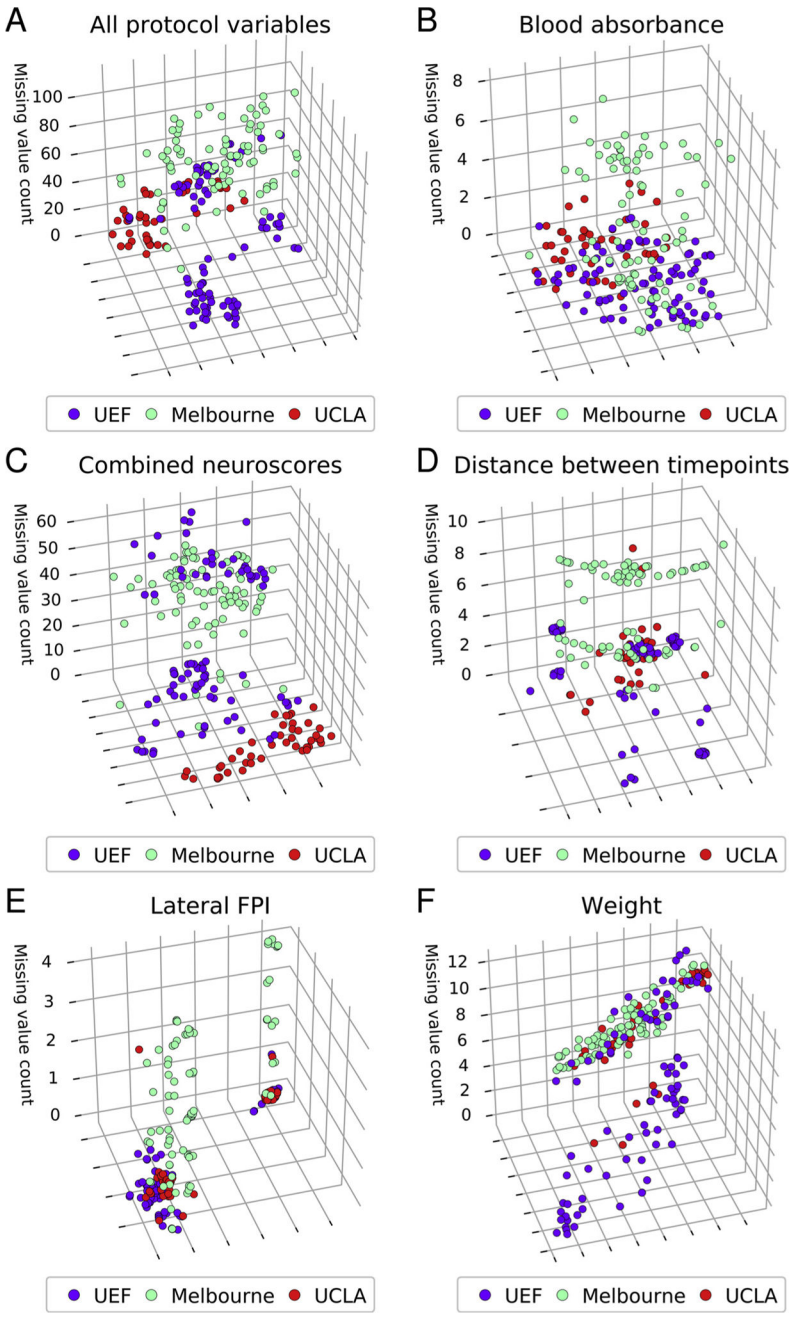


**Fig. 6.**

t-SNE reductions of imputed protocol vectors. The similarity among protocol implementations is presented in plane formed by width and depth axis. The axis perpendicular to the similarity plane denotes the number of missing measurements for each protocol vector. Overall in terms of all protocol variables and variable subsets, Melbourne is missing the largest number of protocol data points. Conversely, data from UEF is mostly complete. (**A**) The centers appear separated both in terms of protocol implementation and the amount of missing data. (**B**) While more mixing appears in the impute blood absorbance data, measurements from different centers are grouped together and measurements from

Melbourne shows high number of missing values. (**C**) Clustering by vector similarity can be seen in the combined neuroscores, where UCLA and UEF are clearly separated. (**D**) In terms distance between procedure time points Melbourne forms a cluster of its own with high missing data count. Majority of UEF protocol vectors are distributed to small clusters separate from Melbourne and UCLA. (**E**) Lateral FPI protocol data is divided into two small clusters, with missing data clearly separating Melbourne from the rest. (**F**) While the centers appear to form a single mixed elongated cluster in terms of weight, data from UCLA and Melbourne vectors contains a significant number of missing values. This results from UCLA and Melbourne data containing only baseline weight. As majority of the visualized vectors from the two centers consisted of imputed variables, the similarity of points in the upper cluster is questionable.

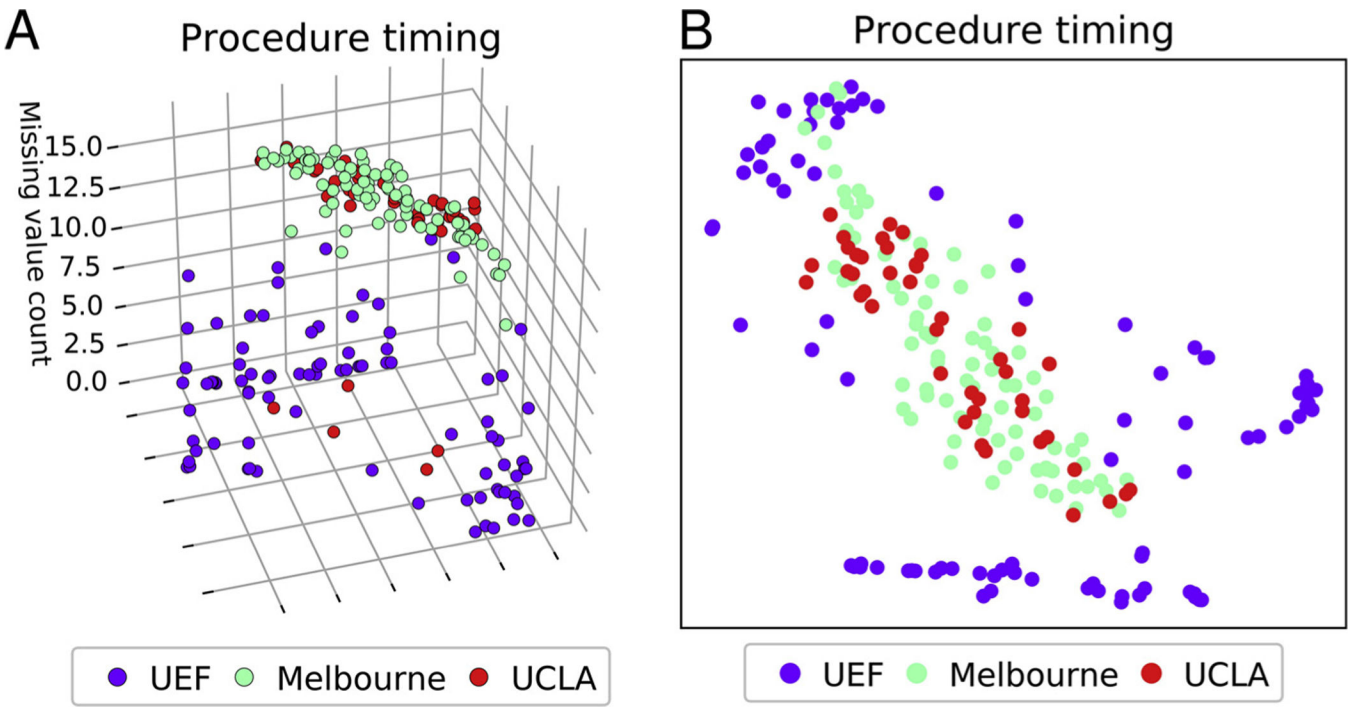


**Fig. 7.**

3D tSNE presenting the protocol vectors for variables related to the timing of experiments. The positioning of the points in the feature space results in a low ICC but nevertheless high AUC. UCLA contains mostly “midday” and Melbourne “between 8–9” values. As explained in the text, this is counted as missing data but imputed using 12:00 and 8:30 instead of using MICE.

## Table 1

ICC values for the EpiBioS4Rx protocol data with imputation. LCI and UCI present upper and lower 95% confidence intervals, respectively. All feature sets display lower within center correlation in comparison to unimputed data. Weight, lateral FPI, procedure timing and combination of all protocol variables displayed relatively low ICC (bolded and underlined values). The ICC values nevertheless indicate clustering by center – especially in terms of baseline neuroscores (bolded).

| Protocol variable | TBI | | | SHAM | | | TBI + SHAM | | |
|---|---|---|---|---|---|---|---|---|---|
| | ICC | LCI | UCI | ICC | LCI | UCI | ICC | LCI | UCI |
| Combined neuroscore | 0.37 | 0.32 | 0.4 | 0.04 | 0 | 0.13 | 0.28 | 0.23 | 0.33 |
| Neuroscore baseline | 0.40 | 0.33 | 0.48 | 0.41 | 0.31 | 0.5 | **0.42** | 0.37 | 0.47 |
| Neuroscore 2 | 0.14 | 0.07 | 0.22 | 0.26 | 0.13 | 0.42 | 0.11 | 0.05 | 0.15 |
| Neuroscore 7 | 0.33 | 0.24 | 0.44 | 0.4 | 0.29 | 0.52 | 0.21 | 0.15 | 0.29 |
| Neuroscore 14 | 0.48 | 0.41 | 0.56 | 0.36 | 0.25 | 0.50 | 0.31 | 0.26 | 0.36 |
| Neuroscore 21 | 0.41 | 0.31 | 0.5 | 0.37 | 0.27 | 0.51 | 0.28 | 0.22 | 0.35 |
| Neuroscore 28 | 0.38 | 0.27 | 0.47 | 0.39 | 0.27 | 0.52 | 0.28 | 0.23 | 0.35 |
| Lateral FPI | 0.46 | 0.34 | 0.6 | 0.1 | 0.02 | 0.19 | **0.07** | 0.03 | 0.17 |
| Blood sampling | 0.22 | 0.14 | 0.3 | 0.26 | 0.16 | 0.39 | 0.21 | 0.14 | 0.27 |
| Weight | 0.03 | 0 | 0.09 | 0.04 | 0 | 0.19 | **0.03** | 0 | 0.09 |
| Distance between timepoints | 0.13 | 0.08 | 0.18 | 0.1 | 0.03 | 0.18 | 0.12 | 0.08 | 0.18 |
| Procedure timing | 0.02 | 0 | 0.11 | 0.02 | 0 | 0.09 | **0.01** | 0 | 0.05 |
| All protocol variables | 0.11 | 0.04 | 0.27 | 0.17 | 0.06 | 0.24 | **0.08** | 0.03 | 0.15 |

Abbreviations: AUC, area under curve; ICC, intraclass correlation; FPI, fluid-percussion injury; LCI, lower confidence interval; UCI, upper confidence interval; TBI, traumatic brain injury.

## Table 2

ICC values for the EpiBioS4Rx protocol data without imputation. LCI and UCI present upper and lower 95% confidence intervals, respectively. Feature sets that did not contain data from at least two centers were excluded. Unimputed baseline and day 14, 21 and 28 neuroscores show clear correlation within center (bolded values). Conversely, day 7 neuroscore (bolded and underlined) displays smallest but existing within center correlation.

| | TBI | | | SHAM | | | TBI + SHAM | | |
|---|---|---|---|---|---|---|---|---|---|
| | ICC | LCI | UCI | ICC | LCI | UCI | ICC | LCI | UCI |
| Neuroscore baseline | 0.51 | 0.43 | 0.60 | 0.49 | 0.36 | 0.61 | **0.50** | 0.44 | 0.56 |
| Neuroscore 2 | 0.29 | 0.2 | 0.39 | 0.36 | 0.19 | 0.52 | 0.16 | 0.09 | 0.24 |
| Neuroscore 7 | 0.17 | 0 | 0.58 | 0.05 | 0 | 0.46 | **0.04** | 0 | 0.22 |
| Neuroscore 14 | 0.60 | 0.53 | 0.66 | 0.5 | 0.39 | 0.61 | **0.39** | 0.32 | 0.47 |
| Neuroscore 21 | 0.59 | 0.51 | 0.66 | 0.5 | 0.41 | 0.61 | **0.39** | 0.33 | 0.47 |
| Neuroscore 28 | 0.60 | 0.51 | 0.68 | 0.53 | 0.43 | 0.63 | **0.42** | 0.35 | 0.52 |
| Lateral FPI | 0.48 | 0.35 | 0.62 | 0.11 | 0 | 0.7 | 0.18 | 0.05 | 0.33 |
| Blood sampling | 0.24 | 0.15 | 0.34 | 0.32 | 0.21 | 0.44 | 0.24 | 0.17 | 0.31 |

Abbreviations: ICC, intraclass correlation; FPI, fluid-percussion injury; LCI, lower confidence interval; UCI, upper confidence interval; TBI, traumatic brain injury.

**Table 3**

KNN AUC values for imputed feature sets. LCI and UCI present upper and lower 95% confidence intervals, respectively. Moderate to high AUC scores indicate center-wise clustering in terms of all feature sets, excluding blood-sampling for sham animals (bolded and underlined). The overall protocol differs clearly between centers, with AUC of 0.93. The bolded values denote feature sets for which ICC was deemed relatively low (< 0.1, Table 1), but which are clearly aggregated by center. While ICC for procedure timing was only 0.01 (Table 1), centers can nevertheless be classified by procedure timing.

| | TBI | | | SHAM | | | TBI + SHAM | | |
|---|---|---|---|---|---|---|---|---|---|
| | AUC | LCI | UCI | AUC | LCI | UCI | AUC | LCI | UCI |
| Combined Neuroscore | 0.84 | 0.78 | 0.90 | 0.74 | 0.61 | 0.84 | **0.89** | 0.85 | 0.93 |
| Neuroscore baseline | 0.88 | 0.84 | 0.92 | 0.78 | 0.68 | 0.86 | 0.9 | 0.87 | 0.92 |
| Neuroscore 2 | 0.8 | 0.76 | 0.85 | 0.76 | 0.67 | 0.86 | 0.86 | 0.81 | 0.9 |
| Neuroscore 7 | 0.84 | 0.79 | 0.9 | 0.77 | 0.66 | 0.87 | 0.85 | 0.81 | 0.89 |
| Neuroscore 14 | 0.9 | 0.86 | 0.93 | 0.77 | 0.68 | 0.86 | 0.88 | 0.84 | 0.92 |
| Neuroscore 21 | 0.86 | 0.82 | 0.91 | 0.76 | 0.65 | 0.85 | 0.87 | 0.83 | 0.9 |
| Neuroscore 28 | 0.83 | 0.77 | 0.89 | 0.72 | 0.63 | 0.85 | 0.85 | 0.81 | 0.9 |
| Injury | 1 | 0.99 | 1 | 0.7 | 0.54 | 0.86 | 0.95 | 0.92 | 0.97 |
| Blood sampling | 0.74 | 0.68 | 0.81 | **0.56** | 0.42 | 0.68 | 0.77 | 0.72 | 0.82 |
| Weight | 0.83 | 0.76 | 0.89 | 0.74 | 0.62 | 0.83 | 0.83 | 0.8 | 0.88 |
| Distance between timepoints | 0.97 | 0.95 | 0.99 | 0.82 | 0.65 | 0.94 | 0.98 | 0.96 | 0.99 |
| Procedure timing | 0.86 | 0.83 | 0.9 | 0.68 | 0.54 | 0.78 | **0.87** | 0.85 | 0.89 |
| All protocol variables | 0.95 | 0.88 | 0.99 | 0.74 | 0.63 | 0.85 | **0.93** | 0.90 | 0.97 |

Abbreviations: AUC, area under curve; ICC, Intraclass correlation; KNN, k-nearest neighbor; LCI, lower confidence interval; UCI, upper confidence interval; TBI, traumatic brain injury.

**Table 4**

KNN AUC values for feature sets without imputation. LCI and UCI present upper and lower 95% confidence intervals, respectively. Feature sets that did not contain data from at least two centers were excluded. Except for day 7 neuroscores(bolded and underlined), KNN AUC indicates high classification performance and therefore clear center-wise clustering.

| | TBI | | | SHAM | | | TBI + SHAM | | |
|---|---|---|---|---|---|---|---|---|---|
| | **AUC** | **LCI** | **UCI** | **AUC** | **LCI** | **UCI** | **AUC** | **LCI** | **UCI** |
| Neuroscore baseline | 0.99 | 0.98 | 1 | 0.99 | 0.94 | 1 | 0.99 | 0.97 | 1 |
| Neuroscore 2 | 0.94 | 0.9 | 0.97 | 0.91 | 0.82 | 0.97 | 0.94 | 0.9 | 0.96 |
| Neuroscore 7 | 0.65 | 0.38 | 0.91 | 0.17 | 0 | 0.69 | **0.6** | 0.37 | 0.82 |
| Neuroscore 14 | 0.97 | 0.96 | 0.99 | 0.93 | 0.89 | 0.98 | 0.96 | 0.94 | 0.98 |
| Neuroscore 21 | 0.98 | 0.96 | 0.99 | 0.98 | 0.96 | 0.99 | 0.96 | 0.94 | 0.98 |
| Neuroscore 28 | 0.98 | 0.96 | 0.99 | 0.95 | 0.95 | 0.96 | 0.97 | 0.95 | 0.98 |
| Injury | 0.99 | 0.98 | 1 | 0.88 | 0.82 | 0.95 | 0.92 | 0.89 | 0.95 |
| Blood sampling | 0.84 | 0.78 | 0.89 | 0.82 | 0.73 | 0.9 | 0.82 | 0.76 | 0.87 |

Abbreviations: AUC, area under curve; ICC, Intraclass correlation; KNN, k-nearest neighbor; LCI, lower confidence interval; UCI, upper confidence interval; TBI, traumatic brain injury.