Data imputation and reconstruction of distributed Parkinson's disease clinical assessments: A comparative evaluation of two aggregation algorithms

Jonatan Reyes[1] *, Yiming Xiao[1,2], and Marta Kersten-Oertel[1,2]

[1] Department of Computer Science and Software Engineering, Gina Cody School of Engineering and Computer Science, Concordia University, Montréal, QC, Canada
[2] PERFORM Center, Concordia University, Montréal, QC, Canada

Abstract. Clinical assessments are an integral part of the care and management of Parkinson's disease, but full spectral assessments are difficult to obtain consistently, especially at follow-up visits. To better understand the etiology and pathogenesis of the disease and to offer accurate prognosis and tailored treatment plans, data-driven computational methods that rely on a large amount of quality clinical assessments have been proposed. However, major limitations, such as privacy and security issues have hindered the greater impact of these techniques. Motivated by the advantages of distributed and collaborative learning, we explore data imputation and reconstruction of clinical scores from the Parkinson Progression Marker Initiative (PPMI) in a multi-center distributed learning environment and we evaluate the reconstruction performance with two aggregation algorithms: Federated Averaging and Precision-weighted Federated Learning[3]. Our results suggest that while the first algorithm provides accurate reconstruction, the latter can better handle data heterogeneity between centers, reaching up to 19% lower reconstruction error.

Keywords: Parkinson's disease · Distributed learning · Imputation · Clinical assessments · Federated Averaging · Precision-weighted Federated Learning

1 Introduction

Parkinson's disease (PD) is the second most frequent neurodegenerative disorder worldwide, associated with both motor and non-motor symptoms. Due to dopamine depletion from the disease, classic motor symptoms involve tremor, rigidity and bradykinesia, severely affecting the patient's daily functions. In addition, psychiatric issues including compulsive behaviors and depression [22],

* Corresponding author: j_yes@encs.concordia.ca

[3] A US provisional patent application has been filed for protecting at least one part of the innovation disclosed in this article

cognitive decline, and sleep disorders can also affect PD patients [9]. As the complexity of the disease has become more evident than previously assumed, further research is crucial to better understand the disorder.

1.1 Clinical Assessments and Challenges

To assess the degree of impairment and clinical progression of PD, clinical assessments are crucial in the care and management of PD, and are conducted in the forms of questionnaires, interactive tests, and objective measurements to gather information regarding both motor and non-motor symptoms. With the aim to search for effective biomarkers for accurate diagnosis and prognosis of PD, various data-driven approaches have been proposed to link imaging or biosample data with these clinical evaluations through statistical or machine/deep learning methods [12].

The Parkinson Progression Marker Initiative (PPMI) [11], sponsored by Michael J. Fox Foundation for Parkinson Research, is a comprehensive public multi-center database (www.ppmi-info.org/data), which includes longitudinal imaging, genetic, biosample and clinical assessment data of large PD cohorts. The up-to-date information on the study can be found at www.ppmi-info.org. Due to missing paper records and the nature of clinical protocols, some tests cannot be performed on the patient at the time of the visits, thus it is not uncommon for there to be missing clinical scores, especially in follow-up visits. This is a common issue when pooling data from multiple centres for disease-related studies, and exclusion of subjects with incomplete records has been adopted in some studies. As PD is highly heterogeneous in disease progression among individuals [8], exclusion of subjects may introduce sample bias in the related statistical and machine/deep learning methods, resulting in unreliable insights. Often data imputation is a solution to the problem. In this technique missing values are replaced with estimations based on the interpretation of contextual information and population distribution [7]. Accurate data imputation of the clinical scores in PD will effectively ensure the reliability and accuracy of the related studies and the proposed machine learning algorithms.

Another major challenge in the application of clinical data-driven algorithms is associated to restrictions on sharing patient data between medical and research centers. Patient data is sensitive and cannot be disclosed as serious privacy issues may arise. For example, patient may be discriminated by employers, insurance companies, or peers based on their health condition, causing embarrassment, paranoia or mental pain [16]. Owing to this, the access to clinical data is limited to the amount of information available within the medical or research center. This may not be adequate for training optimal machine/deep learning algorithms for clinical tasks and biomarker discoveries as it often leads to biased analysis. This is especially relevant for rare diseases, where very few patients are seen at any single institution [20, 5]. To alleviate these issues, emerging data aggregation techniques have been developed to combine data from multiple sources with special attention to security and privacy.

1.2 Contributions

In this paper, we compare two types of aggregation algorithms based on distributed and collaborative learning: Federated Averaging and Precision-weighted Federated Learning. We explore the task of data imputation and reconstruction of clinical scores for PD using the PPMI database and investigate the effects of: (1) the performance of distributed machine learning aggregation algorithms, and (2) the imputation and reconstruction performance varying the number of missing values in the training dataset. To the best of our knowledge, we are the first to evaluate the performance of Federated Averaging and Precision-weighted Federated Learning aggregation algorithms in distributed learning environments for the task of data imputation and reconstruction of PD clinical assessments.

2 Related Work

Previous studies have addressed the problem of data imputation in biomedical data from different angles. The simplest case is deletion, where an entire patient record is removed from the database in the presence of missing values. However, it has been demonstrated that this method degrades the statistical power and yields bias estimates [1]. A more sophisticated statistical method for handling missing values is the single imputation technique. An example of this method is the last observation carried forward (LOCF) approach, where missing values in a longitudinal study are replaced by the last observation recorded [10].

Other methods require the analysis of multiple instances of the data, such as is the case of multiple imputation (MI). MI creates multiple copies of the plausible imputed data sets and estimates associations between the aggregated results [19]. Similarly, the multivariate imputation by chained equations (MICE) creates multiple imputation predictions for each missing value [2]. This method was particularly used to prepare the PPMI data used by Long-Short Term Memory networks to define PD subtypes and predict symptom progression [23]. Alternatively, machine learning-based techniques have been used to compressed clinical assessments and estimate missing scores. Peralta *et al.* investigated data imputation and reconstruction performance of clinical assessments with autoencoders [15]. With FCAEs, the encoder and decoder layers are organized in a fully-connected fashion, which rely on residual substructures called "Computational Blocks".

3 Methods

3.1 Data

We included 17 primary clinical assessments (and their sub-scores) and factors of the PPMI dataset (Table 1). These clinical assessments evaluate both motor and non-motor symptoms of the disease, including motor dysfunctions, psychiatric issues, cognitive functioning. Compared to the work of Peralta *et al.* [15], we utilize half the number of features for the training of machine learning models.

To prepare for the data ingestion process, we performed feature scaling by applying min-max normalization to transform the input data into the range [0,1]. Categorical data were mapped to ordinal values. Crucial information to the diagnosis and prognosis of the disease, such as demographic data (e.g., age, sex, and education) and patients' genotypes were collected at baseline visits and considered constant responses across visits (with the exception of age that changed according to the year of the next assessment). Similar to [15], we excluded the level of dopamine (LEDD) and SPECT imaging data, and focused our work in the primary clinical assessments. The final database contained baseline clinical assessments of 678 subjects and their follow-up visits over 3 years, containing 2466 rows and 102 columns. Table 1 demonstrates the percentages of missing values per assessment.

Table 1. Clinical assessments taken from PPMI and the percentage of rows with missing scores

Questionnaire	% missing value	Questionnaire	% missing value
Benton JLO Test	0.64	SCOPA	1.09
Epworth Sleepiness	0.32	Semantic Fluency	0.40
Geriatric Depression Scale	0.24	Schwab & England ADL	28.66
Hopkins VLT	0.68	STAI	0.36
LNS	0.40	UPDRS I	0.36
MoCA	0.77	UPDRS IP	0.32
QUIP	0.28	UPDRS II	41.20
REM Behavior Disorder	0.40	UPDRS II	41.20
Symbol Digit Modalities Score	0.48	UPDRS III	41.24

Additionally, to simulate the training of multi-institutional models in a distributed learning environment, we split patients randomly into four cohorts. With each cohort assigned to a site, patient information remains independent and is never shared between centers. Given this setup, we reserve one of these sites as the test set to measure the model's generalization performance with patient data never used during training based on the accuracy of reconstruction estimations for missing and non-missing clinical scores.

3.2 Model Setup

Fully-Connected Autoencoders (FCAEs) We use FCAEs introduced by Peralta *et al.* [15]. The architecture is available in a public repository [4]. Each FCAE is trained with a NAdam optimizer using an initial learning rate of 0.001. The NAdam algorithm is a variation of the Adam optimizer that implements Nesterov momentum that accelerates convergence. To control the amount of unnecessary computation per client, we apply a strategy to reduce the learning rate when the model reaches a plateau in learning and stop training when this plateau persists for more than 60 training passes. We use a Mean Squared Error

[4] https://github.com/m-prl/PatiNAE

(MSE) to minimize the loss function over the reconstruction estimations. FCAEs models are implemented with Keras 2.4.3 and Tensorflow 2.4.1.

For evaluation purposes, reconstruction estimations are evaluated on the test site based on the two accuracy measurements defined in [15]:

$$A_1 = \frac{1}{K} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{x}_j^i - x_j^i)^2 * M_j^i \tag{1}$$

$$A_2 = \frac{1}{U} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{x}_j^i - x_j^i)^2 * (1 - M_j^i) \tag{2}$$

where Equation 1 measures the reconstruction performance for non-missing clinical scores based on the total number of known scores (K) in the database, and a mask that identifies non-missing values M_j^i, and where x_j^i represents individual clinical scores and their respective reconstruction estimation $\hat{x}_j^i$. Similarly, Equation 2 quantifies the reconstruction performance of missing clinical scores given the total number of unknown clinical scores (U) and a mask that identifies missing values.

3.3 Aggregation Algorithms

Federated Averaging (FedAvg) Federated learning introduced by McMahan *et al.* [13] works in rounds of communication through a distributed batch of local devices to learn a shared global model. At the beginning of each round, a server sends the initial shared global model to every client. Then, every client uses the shared model to compute stochastic gradient descent (SGD) optimizations with the local data and the resulting update (e.g., network weights) is sent to the server for further processing. After receiving all individual local updates, the central server aggregates them via the FedAvg algorithm to update the parameters of the shared global model, such that:

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k. \tag{3}$$

where $w_t^k + 1$ denotes the model weights of client k at iteration t, n_k is the number of local training samples n_k and n is the total number of samples. The round of communication repeats and as more rounds of communications are performed with this setting, the model learns a better representation of the data distribution and thus performance of the shared global model is optimized. Furthermore, when a new client joins the round of communication, the global model contains enough information from other clients that there is no need to retrain the model as it can be used immediately on the new device.

Precision-weighted Federated Learning (PW) In [18], we proposed the PW algorithm as a variance-based aggregation scheme that averages the weights of distributed machine learning models. This algorithm differs from FedAvg in the way that individual local updates are aggregated. Instead of using the ratio of data samples as the multiplicative factor for weight update, PW takes into account local variance estimations, which are computed by the optimizer, and the update of the parameters of the shared global model is made in proportion to the inverse of this variance:

$$w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{\left(v_{t+1}^{k}\right)^{-1}}{\sum_{k=1}^{K} \left(v_{t+1}^{k}\right)^{-1}} w_{t+1}^{k} \tag{4}$$

where v_{t+1}^{k} denotes the estimated variance of a given weight w at iteration t for client k. This method has shown significant advantages when the data is highly-heterogeneous across clients for benchmark datasets (e.g., MNIST, Fashion-MNIST, and CIFAR-10). However, to the best of our knowledge, no prior studies have examined the performance of this algorithm with medical data.

4 Experimental Results

4.1 Effect of number of missing modalities during training

Given a *corruption ratio*, FCAEs implement a masking layer to remove an entire modality (i.e., a clinical test) from patient records at random. We vary the corruption ratio of the training data to show how the heterogeneity (in terms of the number of missing entries of clinical information) affects the performance of each aggregation method. To do so, we use a fixed batch size of 50 epochs to train each FCAE with corruption rates of 10%, 30% and 60% for 300 iterations. Figure 1 shows the performance evaluations between FedAvg and PW and the reconstruction estimations for non-missing scores (A1). We observe that FedAvg achieves up to 2.7% lower reconstruction error than PW with a 10% corruption ratio. However, PW reaches 16.4% and 5% lower errors than FedAvg with a 30% and 60% corruption ratio, respectively. This experiment suggests that a weighted average can be effective with subtle variations in training data across centers, but when variability in the input data is considered into the aggregation, more accurate imputations and reconstructions may be obtained with highly heterogeneous inputs.

As a point of comparison for model generalization with distributed models, we trained a central model with the pooled dataset (test set excluded). Table 2 summarizes the A1 (MSE) scores obtained on the hold-out test set. Interestingly, these results begin to demonstrate an advantage in using distributed algorithms as we observe lower A1 scores on the reconstruction of non-missing clinical scores compared to those values obtained with a centralized learning setting.

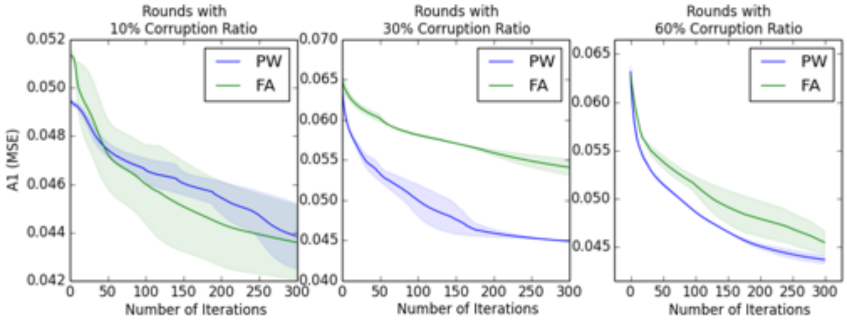

Fig. 1. Performance of FedAvg and PW aggregation algorithms based on the reconstruction error of known values (A1) with an increasing number of missing data in the training dataset.

Table 2. Summary of the performance of the central model and two aggregation algorithms based on known values (A1) on the test set

Corruption Ratio	FA	PW	Central
10%	0.044 ± 0.002	0.044 ± 0.001	0.192 ± 0.000
30%	0.054 ± 0.002	0.045 ± 0.004	0.183 ± 0.000
60%	0.045 ± 0.004	0.044 ± 0.004	0.166 ± 0.000

4.2 Effect of number of missing values during evaluation

We evaluate the reconstruction performance of each aggregation method with additional missing scores. We set the experiment with the same training hyperparameters as above. This time, we introduce additional missing scores into the test set with a 10% and 20% chances for a given score to be missing and compute its A2 error between the estimation and ground truth. Given this setup, it will allow us to explore the scenario when the imputation is implemented in new sites with different levels of missing patient records. Figure 2 shows the performance evaluations based on the A2 (MSE). As expected, the accuracy of the reconstruction A2 is affected by the number of missing scores injected. More specifically, FedAvg shows 1% and 4% lower error than PW with 10% and 20% missing ratio in test, respectively, and 10% corruption ratio. This suggests that FedAvg is a more robust method for the aggregation of homogeneous data. On the contrary, PW reached up to 19% lower reconstruction error than FedAvg with additional heterogeneity in either the training or test sets.

Further evaluations with a central model trained with the pooled dataset were conducted to measure the model generalization based on the reconstruction estimations for missing values. Table 3 summarizes the A2 (MSE) scores of the central model and each aggregation algorithm with 10% and 20% missing values in the test set. We obtained lower A2 scores on central models as more data is available during training for the estimation of missing scores. However, this effect was pronounced when we introduced 10% missing values into the test set.

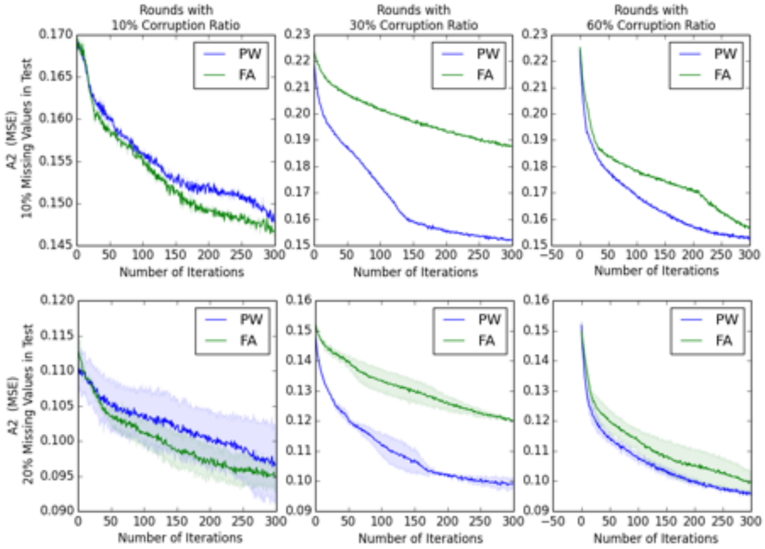

Fig. 2. Performance of FedAvg and PW aggregation algorithms based on the reconstruction error of missing values (A2) with an increasing number of missing values in the testing dataset.

Alternatively, distributed learning improves performance over sites with higher levels of missing patient scores (20% random missing values). These results are consistent with the findings of Tuladhar *et al.* in [20] and suggest that better generalization can be obtained with distributed models.

Table 3. Summary of the performance of the central model and two aggregation algorithms based on the missing values (A2) on the test set

Miss Ratio	Corruption Ratio	FA	PW	Central
10%	10%	0.146 ± 0.006	0.148 ± 0.005	0.117 ± 0.000
	30%	0.187 ± 0.008	0.152 ± 0.017	0.117 ± 0.001
	60%	0.156 ± 0.012	0.152 ± 0.013	0.118 ± 0.002
20%	10%	0.095 ± 0.004	0.096 ± 0.003	0.164 ± 0.000
	30%	0.120 ± 0.008	0.098 ± 0.011	0.167 ± 0.000
	60%	0.099 ± 0.010	0.095 ± 0.010	0.165 ± 0.001

5 Discussion and Conclusion

We compared two aggregation algorithms for distributed learning environments and demonstrated that medical and research centers can embrace collaborate learning to enrich estimations of statistical analysis for the severity and progression of Parkinson's disease. To the best of our knowledge, this is the first work that provides a comparative evaluation of the Federated Averaging and the Precision-weighted Federated Learning aggregation algorithms in the demonstrated domain.

The first point of discussion is that regardless of the aggregation algorithm used, we observe a significant benefit by sharing multi-center clinical data for collaborative model training. We showed that training independent distributed models with information from PD patients can increase the model's generalizability on a hold-out test set without transferring patient's records to a central data store. Notwithstanding, both aggregation algorithms remain vulnerable to inference attacks, therefore, stronger privacy guarantees are needed to protect the information transferred across sites. One solution is using secure protocols [3] or differential-privacy guarantees [6, 14] to ensure that data is transferred between clients and servers safely. Alternatively, we demonstrated that distributed data not only augments the size and variety of the global training set, but also it increases clinical utility. For example, the prediction of the progression and trajectory of the disease can be achieve with less biased decisions than models trained with data in single institutes, leading to more effective treatments or preventive strategies.

An important outcome from the present study is the evaluation of two distributed aggregation algorithms for the imputation and reconstruction of missing scores in PD clinical assessments. Our study indicated that with the lowest level of corruption introduced into the training data, FedAvg can achieve better generalization on a hold-out test set (based on A1 and A2 MSE) with a weighted average, despite the subtle variations in the training data. These results can be explained by the fact that PD patients exhibit high variations in disease patterns and when the number of incomplete responses rise the heterogeneity of the data increases artificially. However, this effect is not pronounced as the corruption rate increases. With less information available to the model to perform the reconstruction, PW seems to be better suited for the reconstruction task.

In addition, it is important to highlight that a randomly initialization for FCAEs was employed in our experiments. Perhaps with a better initialization strategy we could obtain better estimations, especially in the reconstruction errors for (A2). Despite this, we observed that machine/deep learning models were able to leverage the condense information in the clinical scores and factors to provide meaningful and accurate estimations that can be used to perform imputation and reconstruction tasks with acceptable clinical outcomes. Future work may investigate methods for combining information from local datasets, such as cyclical weight transfer [4], split learning [21], or transfer learning [17] as these offer performance improvements with small local training sets.

In conclusion, we present a comparative analysis of the performance of two aggregation algorithms for distributed learning: Federated Averaging (FedAvg) and PW Federated Learning. The task explored here is data imputation and reconstruction of Parkinson's disease clinical questionnaires. We built upon the work of Peralta *et al.* and evaluate the reconstruction performance with Fully-Connected Autoencoders operating in a distributed environment. The results in this study demonstrated that FedAvg is effective in estimating the reconstruction of data with subtle differences across centers, but PW poses a better choice when data is highly heterogeneous.

Acknowledgement

The data that support the findings of the study are publicly available at www.ppmi-info.org. PPMI—a public-private partnership—is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Avid, Biogen, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Roche, Sanofi Genzyme, Servier, Teva, and UCB.

References

1. Allison, P.D.: Missing data. Sage publications (2001)
2. Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J.: Multiple imputation by chained equations: what is it and how does it work? International journal of methods in psychiatric research **20**(1), 40–49 (2011)
3. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A., Seth, K.: Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482 (2016)
4. Chang, K., Balachandar, N., Lam, C., Yi, D., Brown, J., Beers, A., Rosen, B., Rubin, D.L., Kalpathy-Cramer, J.: Distributed deep learning networks among institutions for medical imaging. Journal of the American Medical Informatics Association **25**(8), 945–954 (2018)
5. Denis, A., Mergaert, L., Fostier, C., Cleemput, I., Simoens, S.: A comparative study of european rare disease and orphan drug markets. health Policy **97**(2-3), 173–179 (2010)
6. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4), 211–407 (2014)
7. Efron, B.: Missing data, imputation, and the bootstrap. Journal of the American Statistical Association **89**(426), 463–475 (1994)
8. Ioannidis, J.P., Patsopoulos, N.A., Evangelou, E.: Heterogeneity in meta-analyses of genome-wide association investigations. PloS one **2**(9), e841 (2007)
9. Jankovic, J.: Parkinson's disease: clinical features and diagnosis. Journal of neurology, neurosurgery & psychiatry **79**(4), 368–376 (2008)
10. Lachin, J.M.: Fallacies of last observation carried forward analyses. Clinical trials **13**(2), 161–168 (2016)
11. Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S., et al.: The parkinson progression marker initiative (ppmi). Progress in neurobiology **95**(4), 629–635 (2011)
12. McGhee, D.J., Royle, P.L., Thompson, P.A., Wright, D.E., Zajicek, J.P., Counsell, C.E.: A systematic review of biomarkers for disease progression in parkinson's disease. BMC neurology **13**(1), 1–13 (2013)
13. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics. pp. 1273–1282. PMLR (2017)

14. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V.: Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 691–706. IEEE (2019)
15. Peralta, M., Jannin, P., Haegelen, C., Baxter, J.S.: Data imputation and compression for parkinson's disease clinical questionnaires. Artificial Intelligence in Medicine **114**, 102051 (2021)
16. Price, W.N., Cohen, I.G.: Privacy in the age of medical big data. Nature medicine **25**(1), 37–43 (2019)
17. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. arXiv preprint arXiv:1902.07208 (2019)
18. Reyes, J., Di Jorio, L., Low-Kam, C., Kersten-Oertel, M.: Precision-weighted federated learning. arXiv preprint arXiv:2107.09627 (2021)
19. Sterne, J.A., White, I.R., Carlin, J.B., Spratt, M., Royston, P., Kenward, M.G., Wood, A.M., Carpenter, J.R.: Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj **338** (2009)
20. Tuladhar, A., Gill, S., Ismail, Z., Forkert, N.D., Initiative, A.D.N., et al.: Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling. Journal of biomedical informatics **106**, 103424 (2020)
21. Vepakomma, P., Gupta, O., Swedish, T., Raskar, R.: Split learning for health: Distributed deep learning without sharing raw patient data. arXiv preprint arXiv:1812.00564 (2018)
22. Voon, V., Fox, S.H.: Medication-related impulse control and repetitive behaviors in parkinson disease. Archives of neurology **64**(8), 1089–1096 (2007)
23. Zhang, X., Chou, J., Liang, J., Xiao, C., Zhao, Y., Sarva, H., Henchcliffe, C., Wang, F.: Data-driven subtyping of parkinson's disease using longitudinal clinical records: a cohort study. Scientific reports **9**(1), 1–12 (2019)