TU Delft

Delft University of Technology

Document Version
Final published version

Citation (APA)
Garst, S. J. F. (2026). *Federated learning for collaborative healthcare research and aging biomarkers*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:e5619409-d337-4e07-8ca6-76ec0f44142d

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Federated learning for collaborative healthcare research and aging biomarker design



Swier Garst

Federated learning for collaborative healthcare research and aging biomarker design

Federated learning for collaborative healthcare research and aging biomarker design

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus,
Prof. dr. ir. H. Bijl,
chair of the Board for Doctorates
to be defended publicly on
Wednesday, 16 September 2026, 10:00

by

Swier Jurjen Frans GARST

This dissertation has been approved by the (co)promotors.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M.J.T. Reinders,	Delft University of Technology, *promotor*
Prof. dr. P.E. Slagboom,	Leiden University Medical Center, *promotor*
Independent members:	
Prof. dr. ir. G.J.T. Leus,	Delft University of Technology
Prof. dr. ir. G.J.P.M. Houben,	Delft University of Technology
Prof. dr. D.I. Boomsma,	VU Amsterdam
Prof. dr. M.A. Swertz,	UMCG / RUG
Dr. H. Mei,	LUMC



Keywords: Federated learning, aging biomarkers

Printed by: Ridderprint

Cover by: Swier Garst in collaboration with Gemini 3.5 / Nano Banana 2.

An electronic copy of this dissertation is available at
https://repository.tudelft.nl/.

Contents

Summary xi

Samenvatting xii

Preface xiv

I. Federated learning 1

1. Introduction 3
1.1. Distributed learning . . . 8
1.2. Federated learning platforms . . . 9
1.3. Benchmarking Federated Learning . . . 11
1.4. Federated Clustering . . . 12
1.5. Thesis Contributions on federated learning . . . 13

2. A Systematic Review on Federated Learning for Healthcare Research 17
2.1. Introduction . . . 18
2.2. Methods . . . 19
2.3. Literature Analysis . . . 20
2.3.1. Use-Cases . . . 20
2.3.2. Platforms . . . 21
2.3.3. Methodologies . . . 22
2.3.4. Privacy/Security . . . 23
2.4. Discussion and Conclusion . . . 24

3. A comprehensive experimental comparison between federated and centralized learning 39
3.1. Introduction . . . 40
3.2. Results . . . 41
3.2.1. Linear models on a binary classification problem show a relation between the learning rate and amount of clients used. . . . 41
3.2.2. Extension towards more complex models shows different relationships between federated and centralized models. . . . 42
3.2.3. Binary classification experiments show robustness to sample and class distributions. . . . 42
3.2.4. Extension to multiclass problem results in similar performance as compared to binary classification. . . . 44

3.2.5. SCAFFOLD model updating performs similar to fedAVG model updating. . . . 44
3.2.6. Increasing class distribution aggressiveness gives varying results. . . . 44
3.2.7. Experiments on high dimensional data show potential for federated data processing techniques. . . . 45
3.2.8. Usage of heterogeneous datasets highlights the need to 'think federated' . . . 46
3.2.9. Experiments on 'real life' distributed datasets show increased variance for federated models. . . . 47
3.2.10. Complex federated dataset shows capability for larger neural networks. . . . 48
3.3. Conclusion . . . 49
3.4. Methods . . . 56
3.4.1. Algorithm overview . . . 56
3.4.2. Datasets . . . 58
3.4.3. Experimental setup . . . 60
3.5. Appendices . . . 62
3.5.1. Relation between clients and learning rate (for linear models) 62
3.5.2. SCAFFOLD . . . 65
3.5.3. Federated PCA . . . 66
3.5.4. Supplementary Tables . . . 68
3.5.5. Supplementary figures . . . 71

4. Federated K-Means clustering **75**
4.1. Introduction . . . 76
4.2. Related work . . . 77
4.3. Methods . . . 78
4.3.1. Determining the amount of local clusters . . . 78
4.3.2. Cluster alignment . . . 78
4.4. Results . . . 80
4.4.1. Clients holding different parts of the data . . . 80
4.4.2. Increasing levels of noise . . . 82
4.4.3. Ablation study . . . 83
4.4.4. High variability in number of local clusters . . . 84
4.4.5. Clustering higher dimensional real data . . . 85
4.4.6. Clustering real biological data . . . 86
4.5. Discussion and conclusion . . . 88
4.6. Appendices . . . 92
4.6.1. Background on k-means and k-means++ . . . 92
4.6.2. Extra results on increasing amount of points per cluster . . . 93
4.6.3. FEMNIST distribution . . . 93

5. Discussion **95**
5.1. Challenges to federated learning deployment . . . 96
5.1.1. Privacy . . . 96

5.1.2. Lessons from distributed learning . . . 96
5.1.3. Platforms . . . 97
5.1.4. Data heterogeneity . . . 97
5.1.5. Personalized versus global federated learning . . . 98
5.2. Federated clustering . . . 99
5.2.1. Privacy in federated clustering and evaluation . . . 99
5.2.2. Performance in heterogeneous settings . . . 100
5.2.3. Beyond federated K-means . . . 100
5.3. Concluding remarks . . . 101

II. Ageing Biomarkers **105**

6. Introduction **107**
6.1. The concept of Biological ageing . . . 108
6.2. Molecular and imaging biomarkers of biological ageing . . . 109
6.3. How to train a molecular ageing biomarker . . . 110
6.4. Comparing ageing biomarkers . . . 111
6.5. Thesis contributions on ageing biomarkers . . . 112

7. ProtFI, an efficient frailty-trained proteomics-based biomarker of aging, robustly predicts age-related decline **117**
7.1. Introduction . . . 118
7.2. Results . . . 119
7.2.1. Study Outline . . . 119
7.2.2. Comparison between linear and deep learning models shows no benefit for more complex models . . . 121
7.2.3. Associations with non-training endpoints remain similar when reducing the number of proteins . . . 122
7.2.4. Associations with other health outcomes . . . 125
7.2.5. Forward Feature Selection is an effective method in selecting highly associative proteins for an aging biomarker . . . 125
7.2.6. Comparison with published aging biomarkers . . . 130
7.2.7. Validation in external cohorts shows generalizability of biomarkers . . . 132
7.3. Discussion . . . 134
7.4. Methods . . . 146
7.4.1. Study Population . . . 146
7.4.2. Data preprocessing . . . 147
7.4.3. Ascertainment of covariates and outcomes . . . 149
7.4.4. Biomarker models . . . 150
7.4.5. Forward Feature Selection . . . 153
7.4.6. Statistical model comparison . . . 153
7.4.7. Calculation of previously developed aging biomarkers . . . 153
7.4.8. Association Analysis . . . 153

7.5. Supplementary information . . . 155
7.5.1. Frailty index deficits . . . 155
7.5.2. Multi-omics . . . 156
7.5.3. Full comparison between allprot, CMB and FFS . . . 158
7.5.4. External validation of biomarkers . . . 159
7.5.5. Data splits . . . 159
7.5.6. Sex-specific analysis . . . 160
7.5.7. Neural network architectures . . . 161
7.5.8. Additional tables . . . 161

8. Multitask learning approaches for aging biomarkers 163
8.1. Introduction . . . 164
8.2. Methods . . . 165
8.2.1. Data . . . 165
8.2.2. Predicting Frailty and Mortality Simultaneously . . . 165
8.2.3. Predicting frailty groups . . . 166
8.2.4. Hyperparameter optimization . . . 167
8.2.5. Goodness-of-fit evaluation . . . 167
8.2.6. Clinical evaluations . . . 168
8.2.7. Baseline methods . . . 168
8.3. Results . . . 168
8.3.1. Predicting frailty and mortality simultaneously leads to a performance tradeoff between tasks . . . 169
8.3.2. Estimating frailty groups separately does not increase performance . . . 169
8.3.3. Multitask models can give more stable associations with health related outcomes . . . 170
8.3.4. Discussion and Conclusion . . . 171

9. Multi-cohort federated learning shows synergy in mortality prediction for MRI-based and metabo- lomics-based age scores 175
9.1. Introduction . . . 176
9.2. Methods . . . 177
9.2.1. Data Preparation . . . 177
9.2.2. Federated Learning Infrastructure . . . 180
9.2.3. Deep Learning for BrainAge prediction . . . 180
9.2.4. MetaboAge prediction . . . 182
9.2.5. MetaboHealth prediction . . . 182
9.2.6. The relation between biological age scores . . . 182
9.3. Results . . . 184
9.3.1. Population characteristics . . . 184
9.3.2. BrainAge and MetaboAge models . . . 184
9.3.3. Relation between the biological age scores . . . 188
9.4. Discussion and Conclusion . . . 192
9.5. Addendum: Practical lessons learned from NCDC useCase 1 . . . 201
9.5.1. Implementing Federated Learning . . . 201

9.5.2. Legal Aspects 202
9.5.3. Concluding remarks 203

10. Discussion **205**
10.1. Association versus causality when studying aging biomarkers 207
10.2. Combining aging biomarkers 207
10.3. Deep learning strategies for aging biomarkers 208
10.4. The promise of longitudinal analyses 208
10.5. The importance of federated learning for aging biomarker improvement 209
10.6. Clinical translation of aging biomarkers 209
10.7. Concluding Remarks 210

Epilogue **213**

Acknowledgements **215**

Curriculum Vitæ **219**

List of Publications **221**

Summary

Ageing biomarkers are machine learning models that quantify an individual's ageing process using biological data such as MRI scans or blood-based measurements. They offer the potential to detect accelerated ageing early, but their development requires large and diverse datasets—resources rarely available within a single medical institution. Federated learning (FL), which enables joint model training across sites without sharing raw data, provides a powerful solution for such collaborative biomedical research. This dissertation advances both the methodology of federated learning and the development of biological ageing biomarkers.

Part I focuses on federated learning itself. It synthesizes the landscape of FL in healthcare and highlights how federated approaches can alleviate data sharing problems. Through extensive benchmarking across different data distributions and model families, it demonstrates that FL can consistently match centralized performance even under challenging heterogeneity. This part also introduces new methodology, including a federated clustering algorithm capable of handling divergent local data structures, expanding the range of analytical tools available in privacy-preserving research.
Part II turns to biomarker development, examining how modeling choices shape molecular ageing predictors. Using large-scale proteomic and metabolomic datasets, it compares modeling strategies for predicting ageing-related traits and outcomes. A key contribution is ProteinFrailty (ProtFI), an elastic-net–based proteomic biomarker that predicts frailty while relying on a minimal and interpretable protein panel. ProtFI exhibits strong external validity across independent cohorts. This part also compares two additional ageing biomarkers trained on chronolocial age, namely BrainAge, based on brain MRI features, and MetaboAge, based on NMR-metabolomics features. It does so by integrating federated modeling into biomarker research: a federated deep learning approach for BrainAge estimation shows that pooling knowledge across restricted-access cohorts improves prediction accuracy, especially when harmonizing cohort characteristics. Joint analyses of biological ageing markers reveal that BrainAge and MetaboAge capture complementary physiological dimensions and jointly enhance prediction of all-cause mortality.

Together, these findings demonstrate that federated learning is both a technically robust alternative to centralized analysis and a practical enabler of large-scale, multi-cohort ageing research. This work contributes new FL methodology, shows how federated models overcome data-access barriers in real-world biomedical settings, and advances the development of molecular and imaging-based biomarkers that capture distinct and synergistic aspects of human ageing.

Samenvatting

Verouderingsbiomarkers zijn machine-learningmodellen die het verouderingsproces van een individu kwantificeren aan de hand van biologische gegevens, zoals MRI-scans of bloedwaardes. Ze bieden de mogelijkheid om versnelde veroudering vroegtijdig te detecteren, maar hun ontwikkeling vereist grote en diverse datasets—middelen die zelden beschikbaar zijn binnen één enkele medische instelling. Federated learning (FL), dat het gezamenlijke trainen van een model over meerdere locaties mogelijk maakt zonder het delen van ruwe data, biedt een krachtige oplossing voor dergelijk collaboratief biomedisch onderzoek. Dit proefschrift draagt bij aan zowel de methodologische ontwikkeling van federated learning als aan de ontwikkeling van biologische verouderingsbiomarkers.

Deel I richt zich op federated learning zelf. Het biedt een overzicht van het FL-landschap binnen de gezondheidszorg en laat zien hoe problemen rondom het delen van data verminderd kunnen worden door middel van federated learning. Met een uitgebreide benchmark over verschillende data distributies en modelsoorten wordt aangetoond dat FL vergelijkbare resultaten kan bereiken als gecentraliseerde aanpakken, zelfs onder uitdagende heterogeniteit. Dit deel introduceert ook nieuwe methodologie, waaronder een federatief clustering algoritme dat in staat is om met uiteenlopende lokale datastructuren om te gaan, waarmee het arsenaal aan analysetools voor privacybeschermend onderzoek wordt uitgebreid.

Deel II richt zich op de ontwikkeling van biomarkers en onderzoekt hoe modelkeuzes moleculaire verouderingsvoorspellers beïnvloeden. Met behulp van grootschalige eiwit en metaboloom datasets worden verschillende modelstrategieën vergeleken voor het voorspellen van verouderingsgerelateerde kenmerken en uitkomsten. Een hoofdbijdrage is ProteinFrailty (ProtFI), een op elastic-net gebaseerde eiwit-biomarker welke fragiliteit voorspelt met behulp van een minimaal en interpreteerbaar eiwitpanel. ProtFI vertoont een sterke externe validiteit in verschillende onafhankelijke cohorten. Dit deel vergelijkt ook twee verouderings biomarkers die op leeftijd zijn getraint, namelijk het op hersen MRI scans gebaseerde BrainAge, en MetaboAge, wat gebaseerd is op karakteristieken van NMR-metabolieten. Het doet dit door middel van federatieve modellering in biomarkeronderzoek te integreren: een federated deep-learningstrategie voor BrainAge-schatting laat zien dat het bundelen van kennis over cohorten, zonder het delen van elkaars data, voorspellingsnauwkeurigheid verbeterd, met name wanneer cohort-specifieke karakteristieken worden geharmoniseerd. Gezamenlijke analyses van biologische verouderingsmarkers tonen aan dat BrainAge en MetaboAge complementaire fysiologische dimensies vastleggen en samen de voorspelling van algehele mortaliteit verbeteren.

xii

Samen tonen deze bevindingen aan dat federated learning zowel een technisch robuust alternatief is voor gecentraliseerde analyse als een praktische katalysator voor grootschalig, multicohort verouderingsonderzoek. Dit werk levert nieuwe FL-methodologie, laat zien hoe federatieve modellen barrières rond datatoegang in realistische biomedische settings overwinnen, en bevordert de ontwikkeling van moleculaire en imaging biomarkers die onderscheiden en synergetische aspecten van menselijke veroudering vastleggen.

Preface

According to Moore's law, the number of components that can be integrated on a single chip doubles approximately every two years [1]. Named after Intel co-founder Gordon Moore, this observation implies an exponential increase in computer chip performance. Originally coined in 1965 and adjusted in 1975, Moore's law was intended as a forecast for the following decade. To the surprise of many, including Moore himself, it has remained valid to this day[1].

The persistence of Moore's law has led to a massive increase in compute power. To abide by Moore's law, transistors have decreased in size, increasing efficiency, and ultimately leading to stronger computing systems, as well as massive increases in digital storage capacity. Overall, Moore's law is one of the driving factors behind the modern day digital landscape. Moore's law's impact has also been profound in the field of machine learning, particularly in its early applications to computer vision. A subfield of machine learning, computer vision focuses on enabling computers to interpret and understand visual data, such as photos and videos. Early approaches primarily relied on rule-based algorithms and the incorporation of domain-specific knowledge into computer vision models. This changed in 2012, when Krizhevski et. al. introduced AlexNet[2]. AlexNet was a deep neural network, a large machine learning model built to detect patterns in data without any rule-based algorithms or prior induced knowledge. AlexNet was introduced in the 2012 edition of the ImageNet Large Scale Visual Recognition Challenge, a yearly competition to find the most accurate model to detect 1000 image categories in the large-scale image dataset ImageNet[3]. AlexNet won the competition by a landslide (fig. 1).

The success of AlexNet can be attributed to two primary factors. First, although the theory behind deep neural networks dated back several decades, training a network such as AlexNet had been highly impractical, if not impossible, with the compute power available prior to 2012. However, owing to the advances in computing power driven by Moore's law, Krizhevski et. al. were able to train AlexNet in six days.

The second factor contributing to AlexNet's success was the unprecedented scale of the ImageNet dataset. Fei-Fei Li, the creator of ImageNet, envisioned that a large scale dataset could increase the performance of computer vision models. Krizhevski et. al. seemed to agree. In fact, they specifically chose the ImageNet challenge in favor of the PASCAL Visual Object Classes challenge[5] due to the larger size of the ImageNet training dataset.

[1] Some, including Gordon Moore, predict that Moore's law will end by around 2025.

xiv

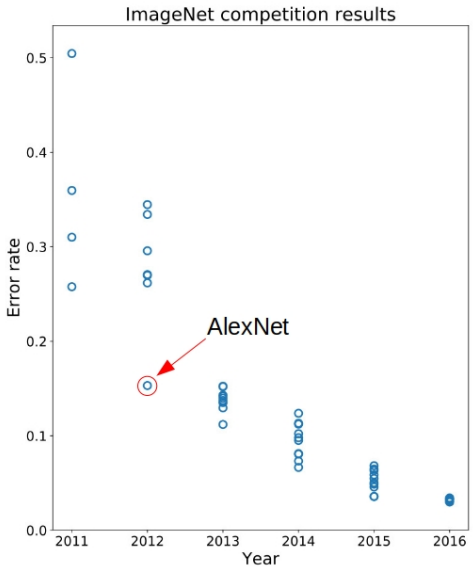


Figure 1.: A history of performance on the ImageNet contest submissions (lower is better). 2012 shows a turning point, where after the submission of AlexNet (the lowest point in the 2012 column), the adoption of deep neural networks drastically reduced the error rates in the coming years. Source: [4]

While larger machine learning models kept pushing benchmark performance, their impact on real world challenges seemed to be lagging behind somewhat. During a talk at the PyTorch Developer Conference in 2019, OpenMined founder Andrew Trask asked a room of machine learning scientists to raise their hand if they had ever trained a model on MNIST, a popular benchmark dataset containing handwritten digits. Almost the entire room raised their hand. He then asked if anyone had trained a model to predict Alzheimer's, inciting the opposite reaction: The room remained still.

Arguably, the main reason for this discrepancy is the absence of the second factor behind AlexNet's succes, larger datasets. Medical data is sensitive, and privacy regulations often discourage hospitals and medical institutes from sharing their patient data. Moreover, a single medical institute usually does not possess sufficient patient data to create a vast dataset comparable to MNIST or ImageNet. In order to create such a resource, patient data from multiple institutes have to be pooled together. Existing efforts in multi-institutional datasets exist [6], but often take years to materialize due to extensive legal and bureaucratic procedures. Even then, the resulting datasets are frequently an order of magnitude smaller than standard computer vision benchmarking datasets.

In 2017, Google was facing a similar issue for a different problem: how to pre-

dict the next word that an Android user is going to type? There are plenty of Android users, who, together, own more than enough data to train such a model. However, this data is scattered across individual phones, and downloading everyone's personal text messages into a large database would probably not be great for Google's reputation. So, in 2017, they came up with a solution, called Federated Learning (FL) [7].

Conventional (centralized) machine learning, can be described as 'bringing the data to the model'. All data is pooled together, and then the model is trained on this collection of data. In federated learning this is reversed: Instead of bringing the data to the model, the model is brought to the data. Data holders form a network over which they share machine learning model parameters, such as weights and biases of a deep neural network. Iteratively, they train the model on their local data only, and aggregate the updated model parameters into a new global model. This way, the local data never leaves the data holders' device, while the resulting model is still trained on the full ensemble of data.

This thesis explores the use of federated learning on healthcare research data. We zoom in on aging biomarker research as our application, but also explore broader aspects of federated learning to address the critical question of what is required to implement such approaches effectively within medical institutes and hospitals. This thesis is split in two parts. Part I explores various aspects of federated learning, whereas part II goes into aging biomarker design and comparison.



Federated learning

1

Introduction

In this first part of this thesis, we explore the efficacy of federated learning. From reviewing the literature (chapter 2), as well as our own work (chapter 3), we show that federated models perform comparable to centralized models under most data distributions. Furthermore, we introduce a novel K-means clustering algorithm for the federated paradigm (FKM). By iteratively improving the global cluster means, as well as automatically updating the amount of local clusters on each client, we show that FKM outperforms already existing federated K-means clustering algorithms, and attains performances similar to centralized K-means clustering.

Federated learning is the practice of iteratively training a machine learning model across multiple data sources without bringing the data from these sources together (figure 1.1). A federated learning system consists of two components: The central server and clients. The central server orchestrates the federated training, aggregating and sharing model weights each round. The clients are the data holders, who use the model weights from the central server to train on their local data, and send the updated weights back to the server. Overall, the goal is to learn a global model W^g that minimizes the local losses L^i for all N local datasets:

$$W^g = \min_{W_g} \sum_{i=1}^{N} L^i(W^g) \tag{1.1}$$

First, the central server sends out initial (random) model parameters, such as the weights and biases of a feedforward neural network, to all clients. The clients update these parameters by training the model on their local data:

$$W^i_{r+1} = \min_{W^i} L(W^i_r(X^i), Y^i) \tag{1.2}$$

where X^i and Y^i are the local data and labels of client i, respectively. W^i_r are the local model parameters at iteration r, which are initialized to the global parameters of that round (W^g_r), and $L(W, Y)$ the loss defining the fit of model W predicting labels Y . Usually, this optimization is performed using stochastic gradient descent:

$$W^i_{r+1} = W^i_r - \eta \nabla L(W^i_r(X^i), Y^i) \tag{1.3}$$

with η being the rate of adaption (i.e. the learning rate). One parameter to consider is how many rounds of stochastic gradient descent are performed, i.e. how many 'local epochs' are used. Increasing the amount of local epochs can decrease the amount of communication rounds between the clients and server, while having too many local epochs risks local models moving too far into local optima, lowering the convergence rate.

Once finished with the local update step, the clients send their updated model weights back to the server, who then aggregates the local models back into a global model (see fig. 1.1) which is one of the core components of federated

learning. The first aggregation method, introduced by McMahan et. al. in 2017, is federated averaging (*fedAvg*) [1], which takes the weighted average of all local model weights, weighted by the local dataset size:

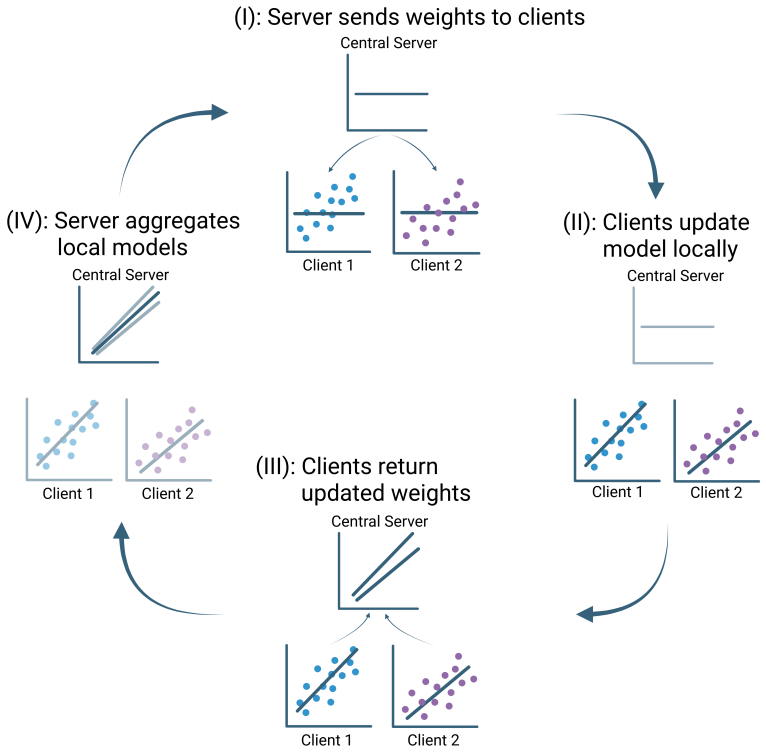


Figure 1.1.: The federated learning procedure. (I) The central server (top) sends global model parameters to both clients (bottom). During initialization, this will be randomly drawn from a distribution. (II) The clients fit the model on their local data. (III) The clients send the (partially) fitted model parameters back to the central server. (IV) the server aggregates the local models into a new global model. The cycle repeats until convergence, or a predetermined amount of iterations. (Figure made using BioRender)

$$W_r^g = \frac{1}{S}\sum_{i=1}^{N} s^i * W_r^i \tag{1.4}$$

where W_r^g is the aggregated global model in round r, s^i the dataset size of client i, N the amount of clients, and $S = \sum_{i=1}^{N} s^i$, i.e. the total data size of all clients

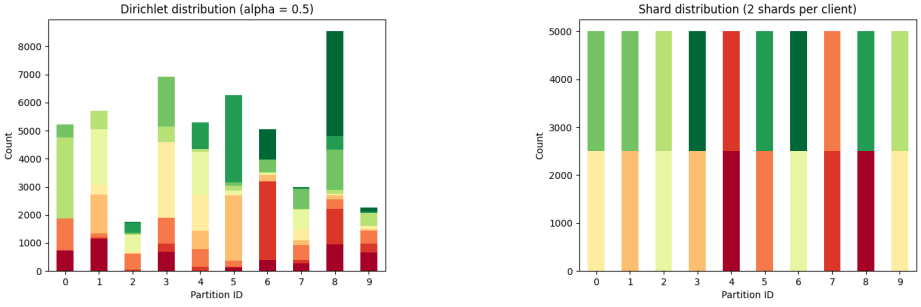
combined.

An alternative formulation shares the final gradients of the local training process instead of model weights, and then updates the global model using a weighted average of these gradients [2]:

$$W_r^g = W_{r-1}^g - \frac{\eta_g}{S} \sum_{i=1}^{N} s^i * g_r^i \tag{1.5}$$

where $g_r^i = \nabla L(W_r^g(X^i), Y^i)$, and η_g the global learning rate. If η_g is equal to η in equation 1.3, and only a single step of stochastic gradient descent is performed locally, equations 1.4 and 1.5 are equivalent. However, due to the introduction of the η_g parameter (on top of the local learning rate η, which is still used in this case), the formulation in equation 1.5 allows for more fine-grained control over the aggregation process [2].

Under most circumstances, it has been shown that fedAvg achieves performances comparable to a model trained on all data pooled together (which we also address in chapters 2 and 3 of this thesis). However, there are reasons to look beyond fedAvg for improved aggregation algorithms. Zhao et. al. introduced the term 'non-iid' (non identically and independently distributed) data, referring to label distributions that differ across clients [3]. They demonstrated that, when extreme imbalances in label distributions exists (see fig. 1.2 for two examples), models trained with fedAvg exhibit significantly degraded performance compared to a centrally trained model (table 1.1).



(a) (b)

Figure 1.2.: Two examples of distributions used by Zhao et. al. [3] in their work showing limitations to federated averaging. (a) shows a partitioning based on a Dirichlet distribution ($\alpha = 0.5$) over the clients, while (b) shows a shard partitioning. Colors represent different classes.

Table 1.1.: Adapted from [4]. Accuracy for three different datasets using various federated aggregation algorithms. Differences between an IID data distribution and two non-IID data distributions are shown, where Dir(0.5) and #C=2 correspond to the distributions shown in figure 1.2a, and 1.2b, respectively. In all cases, both non-IID data distributions show an impact on performance, with the shard distribution (#C = 2) showing a much stronger impact, across all aggregation algorithms.

Dataset	partitioning	FedAvg	FedProx	SCAFFOLD	FedNova
MNIST	IID	99.1 ± 0.1%	99.1 ± 0.1%	99.2 ± 0.0%	99.1 ± 0.1%
	Dir(0.5)	98.9 ± 0.1%	98.9 ± 0.1%	99.0 ± 0.1%	98.9 ± 0.1%
	#C=2	97.0 ± 0.4%	96.4 ± 0.3%	95.9 ± 0.3%	94.5 ± 1.5%
FMNIST	IID	89.6 ± 0.3%	89.5 ± 0.2%	89.7 ± 0.2%	89.4 ± 0.2%
	Dir(0.5)	88.1 ± 0.6%	88.1 ± 0.9%	88.4 ± 0.5%	88.5 ± 0.5%
	#C=2	77.3 ± 4.9%	74.9 ± 2.6%	42.8 ± 28.7%	70.4 ± 5.1%
CIFAR-10	IID	70.4 ± 0.2%	70.2 ± 0.1%	71.5 ± 0.3%	69.5 ± 1.0%
	Dir(0.5)	68.2 ± 0.7%	67.9 ± 0.7%	69.8 ± 0.7%	66.8 ± 1.5%
	#C = 2	49.8 ± 3.3%	50.7 ± 1.7%	49.1 ± 1.7%	46.5 ± 3.5%

This discovery led to the development of various aggregation algorithms effectively handling non-iid data distributions. Among these, some of the most widely used aggregation algorithms are SCAFFOLD[5], fedProx[6] and fedNova[7].

Stochastic Controlled Averaging for Federated Learning (SCAFFOLD) [5] introduces the notion of a global control variate c^g and a local control variate c^l. Together with the model parameters, these control variates are shared between server and clients every round. The clients use this global control variate to modify the model update step:

$$W^i_{r+1} = W^g_r - \eta \nabla g(W^g_r(X_i), y_i) + (c^g_r - c^i_r) \tag{1.6}$$

Note that both c^g and c^i are initialized at 0, so the initial step recovers gradient descent (equation 1.3).

Afterwards, the local control variate is updated based on the distance between the global model and the updated local model (i.e., based on the gradient):

$$c^i_{r+1} = c^i_r - c^g_r + \frac{1}{s_i \eta}(W^g_r - W^i_{r+1}) \tag{1.7}$$

Finally, the global control variate c^g is updated by averaging all local control variates c^i:

$$c^g_{r+1} = c^g_r + \frac{1}{N}\sum_{i=1}^{N}(c^i_{r+1} - c^i_r) \tag{1.8}$$

Notably, model parameters are still aggregated using weighted averaging (equation 1.4). By altering the gradient descent step with the control variate, SCAFFOLD aims to regularize clients based on how strongly their distribution differs from the mean distribution of all clients.

Similarly, fedProx [6] adds a proximal term to the local optimization step, in order to regularize the model based on the distance between the local and global model directly, altering equation 1.2 into:

$$W^l_{r+1} = W^g_r - \eta\nabla g(W^g_r(X_i), y_i) + \frac{\mu}{2}||W^i_r - W^g_r||_2 \tag{1.9}$$

where μ is a hyperparameter (the proximal term) which can be used to tune the regularization strength. Aside from increased performance under non-iid settings, this proximal term also aims to accommodate for FL networks where a variable amount of epochs between clients can occur.

Conversely, the Federated Normalized Averaging Algorithm (FedNova) [7] changes the weighting of the model weights when doing the global aggregation step. Instead of (only) weighing by dataset size, fedNova takes the amount of gradient descent steps taken locally (τ^i_r) into account, which can differ when e.g. a fixed batch sized is used across clients with varying dataset sizes, or when a varying amount of local epochs is permitted. FedNova's aggregation step modifies equation 1.5 into:

$$W^g_{r+1} = W^g_r - \eta_g(\sum_{i=1}^{N} p^i\tau^i_r)\sum_{i=1}^{N}\frac{p^i}{\tau^i_r}g^i_r \tag{1.10}$$

where $p^i = \frac{s^i}{S}$ is the fraction of samples on client i.

1.1. Distributed learning

Federated learning is closely related to the more established field of distributed learning [8]. Like federated learning, distributed learning aims to train a single global model based on information from multiple local clients, without sharing underlying data between those clients. The main difference between the two lies in the topology of the network connecting the clients. Federated learning typically assumes a single central server to which all clients are connected, forming a star topology. In contrast, distributed learning imposes no such restriction, allowing for any arbitrary network topology, provided all clients are connected to each other in some manner.

While such distributed networks bear advantages such as the lack of a

single point of failure, it can also complicate the learning process. Since there is not a single client that is (guaranteed to be) connected to all other clients, there is no notion of an 'orchestrator' who has an overview of the entire network. Clients can only communicate with their direct neighbors, creating a sense of locality with regards to network updates.

In order to ensure that the entire network converges to a shared global optimum, distributed learning algorithms, such as stochastic distributed gradient descent[9] (sDGD) propagate the local information throughout the network. Similar to federated learning, sDGD has both a local update and a sharing phase, in which clients exchange their parameters with their neighbours and compute a weighted average of these parameters.

Specifically, each client i updates the local model as follows:

$$W^i_{r+1} = \sum_{j=1}^{N} A_{ij} W^j_r - \eta \nabla g(W^i_r(X_i), y_i) \tag{1.11}$$

where A is the normalized adjacency matrix, meaning that it is nonzero (and equal) for each A_{ij} if clients i and j are connected (including A_{ii}), and zero otherwise. This way, the local model of client i is first averaged with the local model of all its neighbours j, and then adjusted using the gradient over the local model.

1.2. Federated learning platforms

One of the main drivers of the recent machine learning boom is the accessibility created by popular machine learning packages such as PyTorch and TensorFLow. To achieve a similar widespread adoption of federated learning, equally accessible frameworks for implementing and deploying federated learning systems are necessary. However, federated learning platforms are inherently more complex than centralized machine learning systems. Due to their distributed nature, these federated platforms require additional functionalities to ensure connectivity across the network. Furthermore, administrative tools such as access control become desirable. At the same time, these functionalities are not always required for exploratory work or methodological development. In these cases, a simulated setup on a single machine can simplify development and accelerate experimentation. For such purposes, systems that enable rapid benchmarking of existing algorithms in a simulated environment are particularly valuable.

It is therefore unsurprising that the contemporary federated learning platforms tend to specialize in either simulation or deployment, although some aim to support both. Figure 1.3 provides an overview of several popular federated learning frameworks, along this spectrum from simulation to deployment.

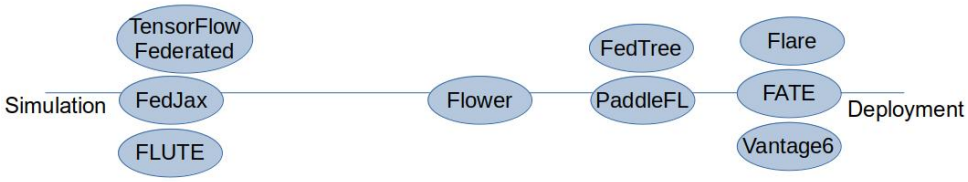


Figure 1.3.: An overview of the available FL software implementations. The X axis represents whether the platform is more focused on simulation or real world deployment.

Some of the most popular federated software platforms include:

- **FedJAX** [10]: Built on top of JAX, FedJax is a simulation platform aimed at quick simulation by providing federated learning algorithms, datasets, models and algorithms.
- **FLUTE** [11]: Introduced by Microsoft, FLUTE is a simulation platform aimed at quick prototyping of new federated learning algorithms based on pyTorch.
- **TensorFlow Federated**[1] is the federated learning API of the popular machine learning library TensorFlow. It has two levels of API (high level for quick implementation, and low level for higher customization), giving tensorFLow Federated high flexibility. However, tensorFLow Federated does not natively provide tools for distributed deployment, and remains mostly a simulation platform.
- **Flower** [12]: Providing both a simulation and a deployment environment, Flower is a platform that can be suitable for either. It provides some primitives for implementing e.g. various aggregation algorithms, but its API is quite open and flexible.
- **PaddleFL**[2] : Based on the chinese deep learning framework PaddlePaddle, PaddleFL aims to be a deployment platform that is flexible with regards to the

[1] https://www.tensorflow.org/federated
[2] https://paddlefl.readthedocs.io/en/latest/

compute it is running on. While it has specific cryptographic implementations such as secure multiparty computation, PaddleFL relies on kubernetes for deployment.

- **FedTree** [13] is a platform focused specifically on the implementation of federated tree-based algorithms. FedTree provides a configuration setup for deployment across multiple machines, as well as a simulation setup.
- **FATE**[3] prizes itself as "the world's first industrial grade federated learning open source framework". With a heavy focus on deployment, FATE provides multiple solutions for running on a network of machines, including a CLI, docker deployment and Kubernetes support.
- **Nvidia Flare**[4] is one of the most 'deployment-ready' platforms currently. With a job-submission like system, Flare uses persistent nodes on all clients and server, to which users can submit their federated computational jobs. In order to facilitate this, Flare has various tools, including a dashboard, a filter mechanism and authorization control.
- **Vantage6** [14] Is perhaps the most flexible platform built for deployment. Wrapping algorithms into docker containers, users have complete freedom in bringing their own code. Access control is included in the platform, as well as a whitelisting function that enables control over which algorithms can be run on each site.

1.3. Benchmarking Federated Learning

Although numerous barriers impede the sharing of medical data, it is not impossible. Procedures such as data sharing agreements enable, albeit through a tedious process, the exchange of medical data with collaborators. Federated learning promises to expedite this process, but it remains important to understand how the performance of federated models compare to a model trained on pooled data.

As mentioned earlier, the results of Zhao et. al. [3] highlight circumstances under which federated learning significantly underperforms compared to a central model. However, the data distributions considered by Zhao et. al. are quite extreme; in some cases, only one or two out of ten classes is present on each client. In contrast, when data is (more or less) evenly distributed among clients (e.g. when a centralized dataset is randomly split between clients), federated learning performance is generally comparable to centralized learning. This raises the question how federated learning performs under intermediate heterogeneous data distributions. While guarantees cannot be made without centralizing the data (at which point the benefits of federated learning are lost), benchmarking studies can provide valuable insights into the characteristics and limitations of federated

[3] https://fate.readthedocs.io/en/latest/
[4] https://developer.nvidia.com/flare

learning under various data distributions. Chapter 3 of this thesis is devoted to performing such a benchmark study.

1.4. Federated Clustering

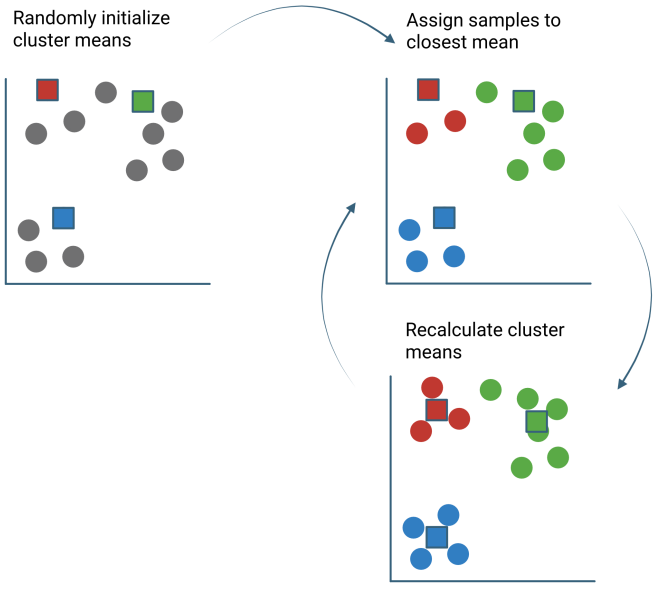


Figure 1.4.: An overview of centralized k-means clustering. (Figure made using BioRender)

Machine learning can be broadly categorized into supervised and unsupervised machine learning approaches. In supervised machine learning, the desired output for each data sample, or the label, is known. However, labeling large datasets can be labor intensive, and in many cases determining ground truth labels may not be feasible, resulting in large amounts of unlabeled data. In such scenarios, unsupervised machine learning algorithms, which identify patterns in the data without relying on labels, are more suitable.

One major class of unsupervised machine learning algorithms is clustering, which aims to group data samples into clusters such that samples within a cluster are as similar as possible. Because clustering typically does not rely on prior label information, it is particularly effective for exploring structure in unlabeled data. A popular (centralized) clustering algorithm, K-means clustering, uses a prespecified number of clusters (denoted by K) which are determined by

iteratively assigning data to the closest cluster centers and updating the cluster centers by calculating the mean of all points assigned to each cluster (see figure 1.4) [15].

While significant progress has been made in supervised federated learning, comparatively little work has focused on unsupervised federated learning [16], and even less on federated (K-means) clustering. Federated clustering presents unique challenges. Clusters may exist on only a subset of the clients, or patterns that appear minor on individual clients may emerge as significant when data from multiple clients are combined. Without data sharing, such global patterns are difficult to detect.

Instead of sharing model parameters as with supervised federated learning, we can attempt to create a federated K-means clustering method by sharing cluster means. The central server sends out global cluster means to all clients. Clients can then run one or more rounds of K-means clustering on their local data before sending their updated cluster means back to the central server. The central server then aggregates these local cluster means into new global means, before sending them back to the clients.

This approach runs into two problems. First, global cluster means can not be directly calculated. When local cluster means are shared, attempting to average these means becomes an alignment problem. If cluster means are substantially shifted due to different local distributions, it can be hard to trace back which means are supposed to be combined. Second, since clients can lack one or more clusters in their local data, it cannot be assumed that a global value for K works on every single client. Running K-means locally using the same value for K while there might be missing clusters can result in suboptimal clustering results, while finetuning local K values by hand (which is sometimes performed in centralized K-means) would scale poorly with the number of clients. In chapter 4 we present a solution to federated K-means clustering addressing these challenges.

1.5. Thesis Contributions on federated learning

The federated learning part of this thesis consists of 3 chapters. Here we summarize the contributions of each of those chapters.

Chapter 2 reviews existing literature on federated learning for healthcare research. We investigate various FL assessments, methodologies and platforms, and comment on (the lack of) privacy mechanisms.

Chapter 3 investigates the relation between centralized and federated learning. Using various machine learning algorithms and datasets, we demonstrate that federated learning is generally robust to multiple types of heterogeneous data distributions, attaining comparable performance to centralized learning even when individual clients exhibit characteristics that differ from the overall dataset. However, batch effects remain an issue and may be more difficult to detect (and therefore

correct for) in a federated setting.

Chapter 4 introduces a novel method for federated clustering, termed federated k-means clustering (FKM). Analogous to supervised federated learning, FKM identifies a global clustering over distributed datasets by sharing cluster centroids rather than the underlying data. Moreover, by comparing intermediate global cluster positions to local data, each client can automatically determine its local value of K (i.e. the number of clusters present in the local dataset). By alleviating the need for manual tuning of K locally, and leveraging the iterative nature of federated learning, FKM outperforms existing approaches for distributed K-means clustering.

Bibliography

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas. 'Communication-efficient learning of deep networks from decentralized data'. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[2] Z. Charles and J. Konečnỳ. 'On the outsized importance of learning rates in local update methods'. In: *arXiv preprint arXiv:2007.00878* (2020).

[3] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin and V. Chandra. 'Federated learning with non-iid data'. In: *arXiv preprint arXiv:1806.00582* (2018).

[4] Q. Li, Y. Diao, Q. Chen and B. He. 'Federated learning on non-iid data silos: An experimental study'. In: *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE. 2022, pp. 965–978.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich and A. T. Suresh. 'Scaffold: Stochastic controlled averaging for federated learning'. In: *International conference on machine learning*. PMLR. 2020, pp. 5132–5143.

[6] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar and V. Smith. 'Federated optimization in heterogeneous networks'. In: *Proceedings of Machine learning and systems* 2 (2020), pp. 429–450.

[7] J. Wang, Q. Liu, H. Liang, G. Joshi and H. V. Poor. 'Tackling the objective inconsistency problem in heterogeneous federated optimization'. In: *Advances in neural information processing systems* 33 (2020), pp. 7611–7623.

[8] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin and K. H. Johansson. 'A survey of distributed optimization'. In: *Annual Reviews in Control* 47 (2019), pp. 278–305. ISSN: 1367-5788. DOI: https://doi.org/10.1016/j.arcontrol.2019.05.006. URL: https://www.sciencedirect.com/science/article/pii/S1367578819300082.

[9] A. Nedic and A. Ozdaglar. 'Distributed Subgradient Methods for Multi-Agent Optimization'. In: *IEEE Transactions on Automatic Control* 54.1 (2009), pp. 48–61. DOI: 10.1109/TAC.2008.2009515.

[10] J. H. Ro, A. T. Suresh and K. Wu. 'FedJAX: Federated learning simulation with JAX'. In: *CoRR* abs/2108.02117 (2021). arXiv: 2108.02117. URL: https://arxiv.org/abs/2108.02117.

[11] M. H. Garcia, A. Manoel, D. M. Diaz, F. Mireshghallah, R. Sim and D. Dimitriadis. *FLUTE: A Scalable, Extensible Framework for High-Performance Federated Learning Simulations*. 2022. arXiv: 2203.13789 [cs.LG]. URL: https://arxiv.org/abs/2203.13789.

[12] D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. De Gusmão *et al.* 'Flower: A friendly federated learning research framework'. In: *arXiv preprint arXiv:2007.14390* (2020).

[13] Q. Li, Y. Cai, Y. Han, C. M. Yung, T. Fu and B. He. *FedTree: A Fast, Effective, and Secure Tree-based Federated Learning System*. https://github.com/Xtra-Computing/FedTree/blob/main/FedTree_draft_paper.pdf. 2022.

[14] A. Moncada-Torres, F. Martin, M. Sieswerda, J. van Soest and G. Geleijnse. 'VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange'. In: *AMIA Annual Symposium Proceedings*. 2020, pp. 870–877.

[15] S. Lloyd. 'Least squares quantization in PCM'. In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[16] A. Grammenos, R. Mendoza Smith, J. Crowcroft and C. Mascolo. 'Federated principal component analysis'. In: *Advances in neural information processing systems* 33 (2020), pp. 6453–6464.

2

A Systematic Review on Federated Learning for Healthcare Research

Federated learning (FL) has emerged as a promising approach for collaborative machine learning in healthcare, allowing institutions to jointly train models without centralizing sensitive patient data. This chapter presents a systematic review of FL applications in healthcare research, conducted following PRISMA guidelines. Of 189 records initially identified through PubMed, 87 studies satisfied the eligibility criteria and were included in the final analysis. The reviewed literature is organized into four interrelated categories: (1) use-cases, which benchmark FL performance against centralized and local training baselines and generally show that federated models approach the performance of centrally trained ones; (2) platforms, the software and infrastructure enabling FL deployment across institutions; (3) methodologies, techniques aimed at improving FL's practical utility, robustness, and performance in heterogeneous, non-IID healthcare data settings; and (4) privacy-enhancing techniques, including approaches such as differential privacy and secure aggregation that strengthen the confidentiality guarantees of federated systems. By mapping the relationships between these categories, this review offers a structured picture of the current FL-in-healthcare landscape, highlights recurring design trade-offs between privacy, performance, and utility, and identifies open challenges that warrant further methodological development.

2.1. Introduction

Advances in machine learning (ML) research over the past decade form a promising avenue to accelerate research in healthcare [1]. However, applying these methods requires large sums of training data. This forms a barrier for using ML models in healthcare, as sharing medical data comes with high administrative costs due to the privacy sensitivity of health data, often inhibiting the pooling of medical data from multiple centers. Furthermore, medical data is often scattered across various institutes, with very little medical institutes having ample data for proper ML model training, further exacerbating the need for collaborative analysis. Federated Learning (FL) [2] poses a possible solution to this problem. FL allows the training of a shared ML model on a distributed collection of data. By sharing only model updates, data can be kept local (i.e. where it was generated), vastly reducing privacy and security risks.

Applying FL in healthcare research has emerged as a popular research field. Various studies have emerged assessing FL performance, either through benchmarks or use cases, by comparing an FL model to a non-FL approach. Furthermore, training a machine learning model is often part of a larger analysis pipeline. While federated machine learning algorithms are the main focus of the FL field, other aspects of such analysis pipelines, e.g. preprocessing steps, are also being adapted to the federated paradigm, enabling broader federated data analyses. Finally, although algorithmic advances are vital in enabling FL for healthcare research, other aspects, such as data harmonization across cohorts, or addressing potential data leakage, are other challenges for enabling the deployment of federated learning. Platforms which tackle these challenges in various ways are also an emerging research area.

Finally, while it is generally accepted that FL gives a higher degree of privacy preservation compared to gathering all data in one place, the full extend of its privacy and security guarantees are still under scrutiny [3]. Earlier work has shown how to infer local data based on shared model parameter updates [4–6]. The emergence of these model inversion attacks has lead to the development of privacy preserving techniques that can be deployed on top of federated learning [7]. However, the developments of these attacks and defences seems barely reflected in real world usecases.

In this chapter, we analyze the current state of federated learning aimed at healthcare research through the means of a systematic review. We identify four categories of work, namely use-cases, methodological works, platforms and privacy preserving techniques.

2.2. Methods

This review follows the PRISMA systematic review guidelines [8]. First the research question was defined as follows:
"What are the current use-cases of federated learning within the biological and medical context?"

From this question, The following exlusion criteria were used:

- not describing a new FL experiment or work (e.g. reviews, perspective papers)
- not utilizing medical/health-related datasets, OR
- not developing methods specifically focused on the medical use-case
- not related to FL / returned by the search query by accident

The initial search query was done on the 31^{st} of May 2022 on PubMed, and consisted of the following:
("federated learning" OR "distributed learning") AND ("healthcare" OR "health" OR "medicine" OR "drug" OR "omics" OR "genomics")
Filtering the year on 2016 to current, this query returned 189 results. After screening and exclusion based on the aforementioned criteria, 87 papers were included in this review (see figure 2.1).

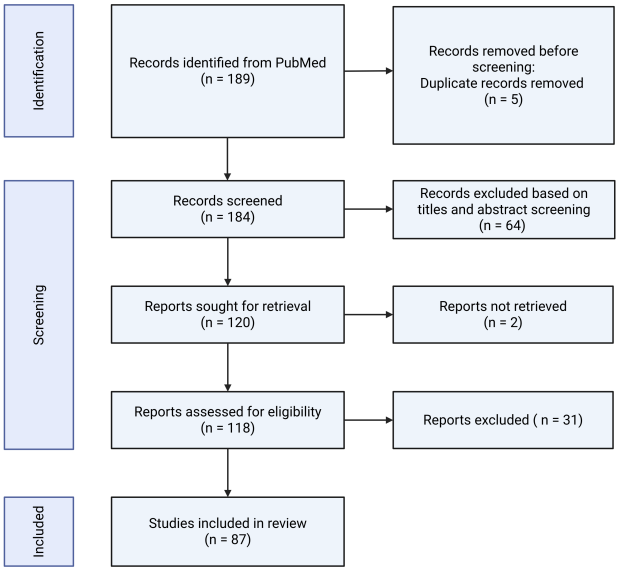


Figure 2.1.: The PRISMA guideline flowchart for this review. (Figure made using BioRender)

2.3. Literature Analysis

In this review, we distinguish four categories of works: Use-cases, platforms, methods and privacy-preserving techniques (see fig. 2.2)

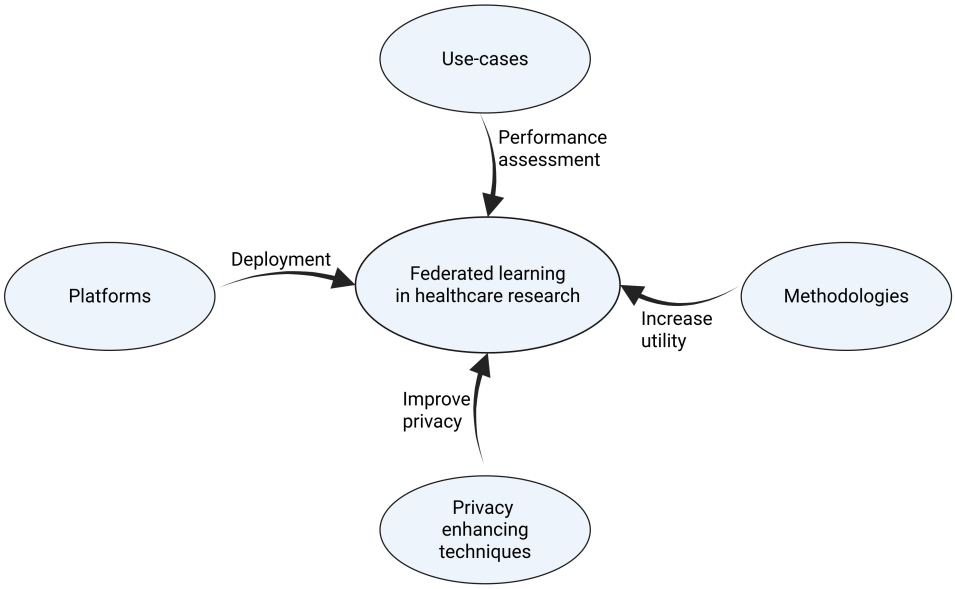


Figure 2.2.: A categorization of the works included in this review (created with bioRender).

2.3.1. Use-Cases

With the introduction of a new technology such as federated learning, use-cases and benchmarks provide valuable assessments of its performance. These use-cases are usually performed by comparing a ML model trained using federated learning to a model which was trained using the classical, centralized paradigm.

When assessing federated learning performance, there are two widely used baselines to compare to. Either, federated model performance is compared to the performance of a model trained on just one local dataset, or it is compared to a model trained on all data pooled together. These approaches will be referred to as 'local' and 'central', respectively.

Of these two options, comparing to a central model is the most popular choice. these comparisons show that, in most cases, the federated model performs on par with the central model, with only minor deviations (see table 2.1). Within the analyzed literature, only 2 out of 35 use-cases find a deviation larger than 10% between federated and centralized model performance [9, 10].

Meanwhile, when comparing to a local model, using FL seems to almost guarantee an increase in performance over single-institute models (table 2.1). From the 26 works that include a comparison to a local model, only one study shows a slight performance decrease [11], while 9 studies report their federated model outperforming local model(s) by more than ten percent [9, 10, 12–18].

Some studies compare with neither a local nor a central model. Most of these works use a single cohort as external validation set to show the generalizability of their federated learning model [19–22]. Furthermore, [23] only compares the performance of FL on different data distributions, while [24] explores how their method changes with a varying amount of clients.

Table 2.1.: overview of the use-case outcomes. deviation is calculated as federated central/local performance - federated performance, either accuracy or F1-score (between 0 and 1).

Reference	Type of comparison	average deviation (minimum - maximum)
[9–13, 15, 25–53]	federated vs. central	0.016 (-0.05 - 0.15)
[9–18, 30, 34, 35, 37, 40, 48, 50, 53–61]	federated vs. local	-0.072 (-0.23 - 0.05)
[19–24, 62]	other	NA

2.3.2. Platforms

Most of the aforementioned use-cases run multiple clients on a single machine using a simulated environment. While this is adequate for assessing federated learning performance, it removes the need for certain practical features that are required in a real world deployment. We term work which enables or describes such real world deployments 'platforms'.

The described platforms vary from theoretical designs to fully implemented international systems. Deist et. al. describe a platform (euroCAT) which was deployed among 5 institutes for detecting dyspnea [63]. Tagliaferri et. al. describe a platform currently used on six Italian centers [64]. Their paper includes a definition for the storage system as well as data analysis. Finally, Hajder et. al. describe the design of a covid detection platform using FL, with the plans of implementing said design over 2021 and 2022 [65].

Furthermore, there are various platforms with only simulated results [66–71]. Samuel et. al. describe a blockchain-based platform for the detection of COVID 19, including a mathematical model of COVID19 [66]. Chang et. al. created an Internet of Medical Things (IOMT) framework using blockchain with adaptive differential privacy [67]. Lakhan et. al. describe a blockchain-based platform for IOMT, including fraud detection and scheduling based on multiple constraints such as energy and delay [68]. Nasser et. al. propose an FL framework using Unmanned Aerial Vehicles for data collection and analysis [69]. Salim et. al. describe a platform using blockchain and InterPlanetary File System databases [70]. Mangold et. al. implemented a fully decentralized platform (i.e. without a central aggregation server) by having all clients connected to each other [71].

2.3.3. Methodologies

There are various methods, such as batch correction and clustering, that do not necessarily follow the iterative process that is associated with ML model training. Adapting such algorithms to be suitable for the federated learning paradigm is often non-trivial, yet these methods are vital parts of larger analysis pipelines, for example during preprocessing. Therefore, various works adapting existing centralized methods for distributed analysis have emerged. Ahmed et. al. propose a method for creating low-dimensional continuous embeddings in the FL setting [72]. Molina et. al. created a method that increases the explainability of an FL model by detecting association rules federatively [73]. Ma et. al. created a method to perform tensor factorization in the FL setting [74]. Li et. al. tackle domain shift for FL by using a global and local model, or by use of a Generative Adversarial Network (GAN) to align data across clients [75]. Bey et. al. created a method to get a stratification that is consistent across clients [76]. Edmonson et. al. created a FL method for modelling count outcomes, specifically for zero-inflated data, using hurdle regression [77]. Zolotareva et. al. used Secure Multiparty Computation (SMC) to create a federated version of limma voom, a tool used for analysing differential expression in genes [78]. Thakur et. al. extend the model agnostic meta-learning framework reptile [79] to the federated setting [80]. Yang et. al. utilize semi-supervised learning for datasets which are only partially labeled [81]. Huang et. al. have made a distributed calibration model by means of a global isotonic regression and SMC [7]. Nasirigerdeh et. al. also used SMC to make a federated version of PLINK [82], a tool used for Genome-Wide Association Studies [83].

Besides converting centralized tools to the federated paradigm, some works aim to alter the aggregation step in FL itself. A more in-depth review on the development of federated aggregation algorithms can be found in [84]. Nguyen et. al. utilize knowledge distillation and clustering in a ring architecture to improve performance on noisy data [85]. Finally, Remedios et. al. propose a new aggregation scheme utilizing momentum-based optimization client-side by averaging gradients instead of model parameters [86].

2.3.4. Privacy/Security

Much work has emerged on attempting to retrieve user data based on model updates which are accessible as part of the FL loop. These attacks include the inversion of gradients [3, 6], as well as adapting model parameters to recover user data [4, 5].

As a result, various defenses and mechanisms to mitigate these attacks have been emerging [87]. Perhaps the most popular privacy mechanism is differential privacy [88]. In the context of FL, differential privacy involves the clipping of gradients and adding Gaussian noise to model parameters or gradients before sending these to the central server [88]. Apart from differential privacy, other methods involve changes in the way gradients are being shared. One example of this is secure aggregation [89], which applies secure multiparty computation methods in order to mask gradients in such a way that each clients' individual gradients are not recoverable, but that the sum of gradients does not change. Shao et. al. developed an aggregation method to avoid privacy leakage from gradients by only transferring part of the model (which is stochastically chosen based on activity) [90] . Han et. al. use FL to train a generative model to synthesize image datasets based on a federation of datasets (which, due to these images being synthesized, reduces privacy risk) [91]. Han et. al. created a method for a zero watermarking scheme to preserve privacy using FL [92]. Kuo et. al. created a framework that can detect several types of model misconducts (i.e. local clients accidentaly or malicously giving poor model updates, which disturb the training process) [93]. Finally, Cui et. al. tackled the problem of doing FL in the case where the central server cannot be trusted, by having parties swap parts of their local models with each other before submitting an update to the central server [94].

Although both attacks and defensive methods regarding privacy in FL are emerging, little of these methods being used in the earlier discussed use-cases; From all the use-cases, only 11 out of 56 implemented some privacy preserving mechanism on top of their FL framework. One reason for this could be that many use cases serve as a proof of concept or (simulated) benchmark, where there is no concern of privacy leakage of real world data across institutes. However, this does not cover all of these use cases, as most utilize data from multiple centers, some of which have experiments set up across these centers.

Another explanation could be that the current defensive methodologies are insufficiently user-friendly. For example, although differential privacy is often praised for its ability to give a numeric guarantee on privacy, the interpretability of this guarantee is quite low. Furthermore, when the goal is to assess FL performance, implementing a method that inherently decreases performance can raise seem counterintuitive.

2.4. Discussion and Conclusion

This systematic review analyzed literature of federated learning in healthcare research. We distinguished four categories: use-cases, platforms, methodological works and privacy-preserving techniques. The use-cases included in this review show promising results for the use of federated learning, while advances in platforms and methodological work enable larger federated analyses beyond machine learning, Furthermore, we note that, though a vast body of research regarding privacy and security in FL exist, few use-cases decide to implement further privacy/security measures, such as differential privacy.

Use-cases mostly compare a federated model with either a centrally or a locally trained model. From a benchmarking perspective, the comparison with a central model is the most desirable. Such a model is trained with an equal amount of data as a federated model, which is often a vital aspect of machine learning model performance [95]. In a sense, a comparison to a centrally trained model serves as a 'gold standard', as it shows if there is any performance gap as a result of keeping the data distributed.

However, often the reason to use FL is exactly *because* training a central model is infeasible, especially when running a real life study instead of a simulated benchmark. In this case, comparing to models trained on single local datasets only would still give an idea of how much performance gain can be achieved by using FL, when pooling data is not possible.

From the comparisons between federated and local models analyzed in this review, it seems that FL models practically always outperform local models. This is a promising result for the adoption of federated learning, especially in cases where pooling data is infeasible. However, federated models also perform relatively similar to central models in most cases, which suggests that FL is a promising solution even when data sharing could be possible though cumbersome. However, it is critical to investigate cases in which substantial differences between federated and centralized models exist, to better characterize the limitations of current federated algorithms.

For widespread adoption of FL, larger initiatives are required, some of which are outlined in section 2.3.2 of this review. These platforms currently operate as separate initiatives. It remains to be seen how these initiatives will develop over time; whether a 'splintering' of FL platforms will occur, or whether initiatives will be fused together, creating larger systems. There are clear advantages in merging platforms, especially when similar data is used, such as increasing dataset sizes. However, larger systems come with more administrative burden, or require looser access control, which can jeopardize user privacy.

We note that most use-cases decide not to implement privacy-preserving mechanisms on top of using FL, such as differential privacy or homomorphic encryption, although a large body of work on attacking privacy guarantees in FL exists.

However, when assessing privacy guarantees in FL, many works on attacks assume in the least case a 'honest-but-curious' attack model. This means that the server (or one or more clients) complies with the FL process, but tries to learn as much about the underlying data as possible in the meantime. This is an understandable assumption when exploring the possibilities of data recovery in FL, however one can wonder whether this scenario applies in a real-world FL collaboration between medical or research institutes, where everyone is assumed to be acting in good faith.

Indeed, assuming good faith and discarding the need for any privacy preserving mechanisms (on top of FL) might not be the best approach. However, when combined with agreements/contracts between institutes, as well as digital infrastructure such as access control, this risk could partially, if not completely, be mitigated by external means. Still, the research on these attacks should warrant care when setting up these types of collaborations, especially when developing public or extensive FL platforms.

Altogether, the literature analyzed in this review shines a positive light on the adoption of FL. Most performance assessments show that using FL will increase model performance compared to local models, or attain similar performances compared to central models. Furthermore, various tools beyond ML model training are being adapted to the federated paradigm, allowing for more elaborate federated data analysis pipelines. However, due to inherent communication between parties, the privacy risks of FL should be taken seriously, and alleviated through administrative measures such as strict access control, or other privacy preserving mechanisms such as differential privacy.

Bibliography

[1] K. Shailaja, B. Seetharamulu and M. A. Jabbar. 'Machine Learning in Healthcare: A Review'. In: *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*. 2018, pp. 910–914. DOI: 10.1109/ICECA.2018.8474918.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas. 'Communication-efficient learning of deep networks from decentralized data'. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[3] J. Geiping, H. Bauermeister, H. Dröge and M. Moeller. 'Inverting gradients-how easy is it to break privacy in federated learning?' In: *Advances in neural information processing systems* 33 (2020), pp. 16937–16947.

[4] L. Fowl, J. Geiping, W. Czaja, M. Goldblum and T. Goldstein. 'Robbing the Fed: Directly Obtaining Private Data in Federated Learning with Modified Models'. In: (2021). arXiv: 2110.13057. URL: http://arxiv.org/abs/2110.13057.

[5] L. Fowl, J. Geiping, S. Reich, Y. Wen, W. Czaja, M. Goldblum and T. Goldstein. 'Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models'. In: (2022). arXiv: 2201.12675. URL: http://arxiv.org/abs/2201.12675.

[6] Y. Wen, J. Geiping, L. Fowl, M. Goldblum and T. Goldstein. 'Fishing for User Data in Large-Batch Federated Learning via Gradient Magnification'. In: (2022). arXiv: 2202.00580. URL: http://arxiv.org/abs/2202.00580.

[7] Y. Huang, S. Gupta, Z. Song, K. Li and S. Arora. 'Evaluating Gradient Inversion Attacks and Defenses in Federated Learning'. In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 7232–7241. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/3b3fff6463464959dcd1b68d0320f781-Paper.pdf.

[8] D. Moher, A. Liberati, J. Tetzlaff and D. G. Altman. 'Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement'. In: *Bmj* 339 (2009).

[9] A. Tuladhar, S. Gill, Z. Ismail and N. D. Forkert. 'Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling'. In: *Journal of Biomedical Informatics* 106 (June 2020). ISSN: 15320464. DOI: 10.1016/j.jbi.2020.103424.

[10] Z. Xiong, Z. Cheng, X. Lin, C. Xu, X. Liu, D. Wang, X. Luo, Y. Zhang, H. Jiang, N. Qiao and M. Zheng. 'Facing small and biased data dilemma in drug discovery with enhanced federated learning approaches'. In: *Science China Life Sciences* 65.3 (Mar. 2022), pp. 529–539. ISSN: 18691889. DOI: 10.1007/s11427-021-1946-0.

[11] J. C. Liu, M. ; Jack Goetz, S. Sen and A. Tewari. 'Learning From Others Without Sacrificing Privacy: Simulation Comparing Centralized and Federated Machine Learning on Mobile Health Data'. In: (). DOI: 10.2196/23728. URL: https://mhealth.jmir.org/2021/3/e23728.

[12] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen and S. Bakas. 'Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data'. In: *Scientific Reports* 10.1 (Dec. 2020). ISSN: 20452322. DOI: 10.1038/s41598-020-69250-1.

[13] O. Choudhury, Y. Park, T. Salonidis, A. Gkoulalas-Divanis, I. Sylla and A. K. Das. 'Predicting Adverse Drug Reactions on Distributed Health Data using Federated Learning'. In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2019 (2019), pp. 313–322. ISSN: 1942597X.

[14] I. Dayan, H. R. Roth, A. Zhong, A. Harouni, A. Gentili, A. Z. Abidin, A. Liu, A. B. Costa, B. J. Wood, C. S. Tsai, C. H. Wang, C. N. Hsu, C. K. Lee, P. Ruan, D. Xu, D. Wu, E. Huang, F. C. Kitamura, G. Lacey, G. C. de Antônio Corradi, G. Nino, H. H. Shin, H. Obinata, H. Ren, J. C. Crane, J. Tetreault, J. Guan, J. W. Garrett, J. D. Kaggie, J. G. Park, K. Dreyer, K. Juluru, K. Kersten, M. A. B. C. Rockenbach, M. G. Linguraru, M. A. Haider, M. AbdelMaseeh, N. Rieke, P. F. Damasceno, P. M. C. e Silva, P. Wang, S. Xu, S. Kawano, S. Sriswasdi, S. Y. Park, T. M. Grist, V. Buch, W. Jantarabenjakul, W. Wang, W. Y. Tak, X. Li, X. Lin, Y. J. Kwon, A. Quraini, A. Feng, A. N. Priest, B. Turkbey, B. Glicksberg, B. Bizzo, B. S. Kim, C. Tor-Díez, C. C. Lee, C. J. Hsu, C. Lin, C. L. Lai, C. P. Hess, C. Compas, D. Bhatia, E. K. Oermann, E. Leibovitz, H. Sasaki, H. Mori, I. Yang, J. H. Sohn, K. N. K. Murthy, L. C. Fu, M. R. F. de Mendonça, M. Fralick, M. K. Kang, M. Adil, N. Gangai, P. Vateekul, P. Elnajjar, S. Hickman, S. Majumdar, S. L. McLeod, S. Reed, S. Gräf, S. Harmon, T. Kodama, T. Puthanakit, T. Mazzulli, V. L. de Lavor, Y. Rakvongthai, Y. R. Lee, Y. Wen, F. J. Gilbert, M. G. Flores and Q. Li. 'Federated learning for predicting clinical outcomes in patients with COVID-19'. In: *Nature Medicine* 27.10 (Oct. 2021), pp. 1735–1743. ISSN: 1546170X. DOI: 10.1038/s41591-021-01506-3.

[15] M. Y. Lu, R. J. Chen, D. Kong, J. Lipkova, R. Singh, D. F. Williamson, T. Y. Chen and F. Mahmood. 'Federated learning for computational pathology on gigapixel whole slide images'. In: *Medical Image Analysis* 76 (Feb. 2022).

ISSN: 13618423. DOI: 10.1016/j.media.2021.102298. arXiv: 2009.10190.

[16] M. Flores, I. Dayan, H. Roth, A. Zhong, A. Harouni, A. Gentili, A. Abidin, A. Liu, A. Costa, B. Wood, C.-S. Tsai, C.-H. Wang, C.-N. Hsu, C. K. Lee, C. Ruan, D. Xu, D. Wu, E. Huang, F. Kitamura, G. Lacey, G. C. d. A. Corradi, H.-H. Shin, H. Obinata, H. Ren, J. Crane, J. Tetreault, J. Guan, J. Garrett, J. G. Park, K. Dreyer, K. Juluru, K. Kersten, M. A. B. C. Rockenbach, M. Linguraru, M. Haider, M. AbdelMaseeh, N. Rieke, P. Damasceno, P. M. C. E. Silva, P. Wang, S. Xu, S. Kawano, S. Sriswa, S. Y. Park, T. Grist, V. Buch, W. Jantarabenjakul, W. Wang, W. Y. Tak, X. Li, X. Lin, F. Kwon, F. Gilbert, J. Kaggie, Q. Li, A. Quraini, A. Feng, A. Priest, B. Turkbey, B. Glicksberg, B. Bizzo, B. S. Kim, C. Tor-Diez, C.-C. Lee, C.-J. Hsu, C. Lin, C.-L. Lai, C. Hess, C. Compas, D. Bhatia, E. Oermann, E. Leibovitz, H. Sasaki, H. Mori, I. Yang, J. H. Sohn, K. N. K. Murthy, L.-C. Fu, M. R. F. de Mendonça, M. Fralick, M. K. Kang, M. Adil, N. Gangai, P. Vateekul, P. Elnajjar, S. Hickman, S. Majumdar, S. McLeod, S. Reed, S. Graf, S. Harmon, T. Kodama, T. Puthanakit, T. Mazzulli, V. d. L. Lavor, Y. Rakvongthai, Y. R. Lee and Y. Wen. 'Federated Learning used for predicting outcomes in SARS-COV-2 patients.' In: *Research square* (Jan. 2021). DOI: 10.21203/rs.3.rs-126892/v1. URL: http://www.ncbi.nlm.nih.gov/pubmed/33442676%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7805458.

[17] X. Bai, H. Wang, L. Ma, Y. Xu, J. Gan, Z. Fan, F. Yang, K. Ma, J. Yang, S. Bai, C. Shu, X. Zou, R. Huang, C. Zhang, X. Liu, D. Tu, C. Xu, W. Zhang, X. Wang, A. Chen, Y. Zeng, D. Yang, M. W. Wang, N. Holalkere, N. J. Halin, I. R. Kamel, J. Wu, X. Peng, X. Wang, J. Shao, P. Mongkolwat, J. Zhang, W. Liu, M. Roberts, Z. Teng, L. Beer, L. E. Sanchez, E. Sala, D. L. Rubin, A. Weller, J. Lasenby, C. Zheng, J. Wang, Z. Li, C. Schönlieb and T. Xia. 'Advancing COVID-19 diagnosis with privacy-preserving collaboration in artificial intelligence'. In: *Nature Machine Intelligence* 3.12 (Dec. 2021), pp. 1081–1089. ISSN: 25225839. DOI: 10.1038/s42256-021-00421-z. arXiv: 2111.09461.

[18] R. W. Filice, A. Stein, I. Pan and G. Shih. 'Federated Deep Learning to More Reliably Detect Body Part for Hanging Protocols, Relevant Priors, and Workflow Optimization'. In: *Journal of Digital Imaging* 35.2 (Apr. 2022), pp. 335–339. ISSN: 1618727X. DOI: 10.1007/s10278-021-00547-x.

[19] Z. Shi, K. G. Foley, J. Pablo de Mey, E. Spezi, P. Whybra, T. Crosby, J. van Soest, A. Dekker and L. Wee. 'External Validation of Radiation-Induced Dyspnea Models on Esophageal Cancer Radiotherapy Patients.' In: *Frontiers in oncology* 9 (Dec. 2019), p. 1411. ISSN: 2234-943X. DOI: 10.3389/fonc.2019.01411. URL: http://www.ncbi.nlm.nih.gov/pubmed/31921668%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6927468.

[20] G. Defraene, F. J. Dankers, G. Price, E. Schuit, W. van Elmpt, S. Arredouani, M. Lambrecht, J. Nuyttens, C. Faivre-Finn and D. De Ruysscher. 'Multifactorial risk factors for mortality after chemotherapy and radiotherapy for non-small cell

lung cancer'. In: *Radiotherapy and Oncology* 152 (Nov. 2020), pp. 117–125. ISSN: 18790887. DOI: 10.1016/j.radonc.2019.09.005.

[21] M. Field, S. Vinod, N. Aherne, M. Carolan, A. Dekker, G. Delaney, S. Greenham, E. Hau, J. Lehmann, J. Ludbrook, A. Miller, A. Rezo, J. Selvaraj, J. Sykes, L. Holloway and D. Thwaites. 'Implementation of the Australian Computer-Assisted Theragnostics (AusCAT) network for radiation oncology data extraction, reporting and distributed learning'. In: *Journal of Medical Imaging and Radiation Oncology* 65.5 (Aug. 2021), pp. 627–636. ISSN: 17549485. DOI: 10.1111/1754-9485.13287.

[22] T. M. Deist, F. J. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin and A. Dekker. 'Distributed learning on 20 000+ lung cancer patients – The Personal Health Train'. In: *Radiotherapy and Oncology* 144 (Mar. 2020), pp. 189–200. ISSN: 18790887. DOI: 10.1016/j.radonc.2019.11.019.

[23] T.-T. Ho, K.-D. Tran and Y. Huang. 'FedSGDCOVID: Federated SGD COVID-19 Detection under Local Differential Privacy Using Chest X-ray Images and Symptom Information.' In: *Sensors (Basel, Switzerland)* 22.10 (May 2022), p. 3728. ISSN: 1424-8220. DOI: 10.3390/s22103728. URL: http://www.ncbi.nlm.nih.gov/pubmed/35632136.

[24] A. Nandi and F. Xhafa. 'A federated learning method for real-time emotion state classification from multi-modal streaming'. In: *Methods* (2022). ISSN: 10959130. DOI: 10.1016/j.ymeth.2022.03.005.

[25] G. H. Lee and S. Y. Shin. 'Federated Learning on Clinical Benchmark Data: Performance Assessment'. In: *Journal of Medical Internet Research* 22.10 (Oct. 2020). ISSN: 14388871. DOI: 10.2196/20891.

[26] E. Brophy, M. De Vos, G. Boylan and T. Ward. 'Estimation of continuous blood pressure from ppg via a federated learning approach'. In: *Sensors* 21.18 (Sept. 2021). ISSN: 14248220. DOI: 10.3390/s21186311. arXiv: 2102.12245.

[27] Y. Mou, S. Welten, M. Jaberansary, Y. U. Yediel, T. Kirsten, S. Decker and O. Beyan. 'Distributed skin lesion analysis across decentralised data sources'. In: *Public Health and Informatics: Proceedings of MIE 2021* (July 2021), pp. 352–356. DOI: 10.3233/SHTI210179.

[28] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra and K. Lekadir. 'Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease'. In: *Scientific Reports* 12.1 (Dec. 2022). ISSN: 20452322. DOI: 10.1038/s41598-022-07186-4.

[29] L. Qu, N. Balachandar, M. Zhang and D. Rubin. 'Handling data heterogeneity with generative replay in collaborative learning for medical imaging.' In: *Medical image analysis* 78 (May 2022), p. 102424. ISSN: 1361-8423. DOI: 10.1016/j.media.2022.102424. URL: http://www.ncbi.nlm.nih.gov/pubmed/35390737.

[30] Q. Dou, T. Y. So, M. Jiang, Q. Liu, V. Vardhanabhuti, G. Kaissis, Z. Li, W. Si, H. H. Lee, K. Yu, Z. Feng, L. Dong, E. Burian, F. Jungmann, R. Braren, M. Makowski, B. Kainz, D. Rueckert, B. Glocker, S. C. Yu and P. A. Heng. 'Federated deep learning for detecting COVID-19 lung abnormalities in CT: a privacy-preserving multinational validation study'. In: *npj Digital Medicine* 4.1 (Dec. 2021). ISSN: 23986352. DOI: 10.1038/s41746-021-00431-6.

[31] A. Jochems, T. M. Deist, I. El Naqa, M. Kessler, C. Mayo, J. Reeves, S. Jolly, M. Matuszak, R. Ten Haken, J. van Soest, C. Oberije, C. Faivre-Finn, G. Price, D. de Ruysscher, P. Lambin and A. Dekker. 'Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries'. In: *International Journal of Radiation Oncology Biology Physics* 99.2 (Oct. 2017), pp. 344–352. ISSN: 1879355X. DOI: 10.1016/j.ijrobp.2017.04.021.

[32] S. Baghersalimi, T. Teijeiro, D. Atienza and A. Aminifar. 'Personalized Real-Time Federated Learning for Epileptic Seizure Detection'. In: *IEEE Journal of Biomedical and Health Informatics* 26.2 (Feb. 2022), pp. 898–909. ISSN: 21682208. DOI: 10.1109/JBHI.2021.3096127.

[33] H. Islam, A. Mosa and Famia. 'A Federated Mining Approach on Predicting Diabetes-Related Complications: Demonstration Using Real-World Clinical Data'. In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2021 (2021), pp. 556–564. ISSN: 1942597X.

[34] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor and H. R. Tizhoosh. 'Federated learning and differential privacy for medical image analysis'. In: *Scientific Reports* 12.1 (Dec. 2022). ISSN: 20452322. DOI: 10.1038/s41598-022-05539-7.

[35] A. Rehman, I. Razzak and G. Xu. 'Federated Learning for Privacy Preservation of Healthcare Data from Smartphone-based Side-Channel Attacks.' In: *IEEE journal of biomedical and health informatics* PP (May 2022), pp. 1–1. ISSN: 2168-2208. DOI: 10.1109/JBHI.2022.3171852. URL: http://www.ncbi.nlm.nih.gov/pubmed/35503855.

[36] Y. Xu, L. Ma, F. Yang, Y. Chen, K. Ma, J. Yang, X. Yang, Y. Chen, C. Shu, Z. Fan, J. Gan, X. Zou, R. Huang, C. Zhang, X. Liu, D. Tu, C. Xu, W. Zhang, D. Yang, M.-W. Wang, X. Wang, X. Xie, H. Leng, N. Holalkere, N. J. Halin, I. R. Kamel, J. Wu, X. Peng, X. Wang, J. Shao, P. Mongkolwat, J. Zhang, D. L. Rubin, G. Wang, C. Zheng, Z. Li, X. Bai and T. Xia. 'A collaborative online AI engine for CT-based COVID-19 diagnosis.' In: *medRxiv : the preprint server for health sciences* (May 2020). DOI: 10.1101/2020.05.10.20096073. URL: http://

www.ncbi.nlm.nih.gov/pubmed/32511484%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7273252.

[37] S. Sav, J.-P. Bossuat, J. R. Troncoso-Pastoriza, M. Claassen and J.-P. Hubaux. 'Privacy-preserving federated neural network learning for disease-associated cell classification.' In: *Patterns (New York, N.Y.)* 3.5 (May 2022), p. 100487. ISSN: 2666-3899. DOI: 10.1016/j.patter.2022.100487. URL: http://www.ncbi.nlm.nih.gov/pubmed/35607628%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC9122966.

[38] M. N. Hossen, V. Panneerselvam, D. Koundal, K. Ahmed, F. M. Bui and S. M. Ibrahim. 'Federated Machine Learning for Detection of Skin Diseases and Enhancement of Internet of Medical Things (IoMT) Security'. In: *IEEE Journal of Biomedical and Health Informatics* (2022). ISSN: 21682208. DOI: 10.1109/JBHI.2022.3149288.

[39] A. Sadilek, L. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie, J. MacFarlane, A. Vullikanti, M. Marathe, P. Eastham, J. S. Brownstein, B. A. Y. Arcas, M. D. Howell and J. Hernandez. 'Privacy-first health research with federated learning.' In: *NPJ digital medicine* 4.1 (Sept. 2021), p. 132. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00489-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/34493770%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8423792.

[40] Ø. Sørensen, A. M. Brandmaier, D. Macià, K. Ebmeier, P. Ghisletta, R. A. Kievit, A. M. Mowinckel, K. B. Walhovd, R. Westerhausen and A. Fjell. 'Meta-analysis of generalized additive models in neuroimaging studies'. In: *NeuroImage* 224 (Jan. 2021). ISSN: 10959572. DOI: 10.1016/j.neuroimage.2020.117416. arXiv: 2002.02627.

[41] N. Balachandar, K. Chang, J. Kalpathy-Cramer and D. L. Rubin. 'Accounting for data variability in multi-institutional distributed deep learning for medical imaging'. In: *Journal of the American Medical Informatics Association* 27.5 (May 2020), pp. 700–708. ISSN: 1527974X. DOI: 10.1093/jamia/ocaa017.

[42] J. Wu, Q. Dong, J. Zhang, Y. Su, T. Wu, R. J. Caselli, E. M. Reiman, J. Ye, N. Lepore, K. Chen, P. M. Thompson and Y. Wang. 'Federated Morphometry Feature Selection for Hippocampal Morphometry Associated Beta-Amyloid and Tau Pathology.' In: *Frontiers in neuroscience* 15 (Nov. 2021), p. 762458. ISSN: 1662-4548. DOI: 10.3389/fnins.2021.762458. URL: http://www.ncbi.nlm.nih.gov/pubmed/34899166%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8655732.

[43] T. S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I. C. Paschalidis and W. Shi. 'Federated learning of predictive models from federated Electronic Health Records'. In: *International Journal of Medical Informatics* 112 (Apr. 2018), pp. 59–67. ISSN: 18728243. DOI: 10.1016/j.ijmedinf.2018.01.007.

[44] I. Feki, S. Ammar, Y. Kessentini and K. Muhammad. 'Federated learning for COVID-19 screening from Chest X-ray images.' In: *Applied soft computing* 106 (July 2021), p. 107330. ISSN: 1568-4946. DOI: 10.1016/j.asoc.2021.107330. URL: http://www.ncbi.nlm.nih.gov/pubmed/33776607%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7979273.

[45] M. Bogowicz, A. Jochems, T. M. Deist, S. Tanadini-Lang, S. H. Huang, B. Chan, J. N. Waldron, S. Bratman, B. O'Sullivan, O. Riesterer, G. Studer, J. Unkelbach, S. Barakat, R. H. Brakenhoff, I. Nauta, S. E. Gazzani, G. Calareso, K. Scheckenbach, F. Hoebers, F. W. Wesseling, S. Keek, S. Sanduleanu, R. T. Leijenaar, M. R. Vergeer, C. R. Leemans, C. H. Terhaard, M. W. van den Brekel, O. Hamming-Vrieze, M. A. van der Heijden, H. M. Elhalawani, C. D. Fuller, M. Guckenberger and P. Lambin. 'Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer'. In: *Scientific Reports* 10.1 (Dec. 2020). ISSN: 20452322. DOI: 10.1038/s41598-020-61297-4.

[46] C. Luo, M. N. Islam, N. E. Sheils, J. Buresh, M. J. Schuemie, J. A. Doshi, R. M. Werner, D. A. Asch and Y. Chen. 'dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling.' In: *Journal of the American Medical Informatics Association : JAMIA* (May 2022). ISSN: 1527-974X. DOI: 10.1093/jamia/ocac067. URL: http://www.ncbi.nlm.nih.gov/pubmed/35579348.

[47] M. A. Salam, S. Taha and M. Ramadan. 'COVID-19 detection using federated machine learning'. In: *PLoS ONE* 16.6 June (June 2021). ISSN: 19326203. DOI: 10.1371/journal.pone.0252573.

[48] A. Vaid, S. K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J. K. De Freitas, T. Wanyan, K. W. Johnson, M. Bicak, E. Klang, Y. Joon Kwon, A. Costa, S. Zhao, R. Miotto, A. W. Charney, E. Böttinger, Z. A. Fayad, G. N. Nadkarni, F. Wang and B. S. Glicksberg. 'Federated Learning of Electronic Health Records to Improve Mortality Prediction in Hospitalized Patients With COVID-19: Machine Learning Approach'. In: (). DOI: 10.2196/24207. URL: http://medinform.jmir.org/2021/1/e24207/.

[49] F. Zerka, V. Urovi, F. Bottari, R. T. Leijenaar, S. Walsh, H. Gabrani-Juma, M. Gueuning, A. Vaidyanathan, W. Vos, M. Occhipinti, H. C. Woodruff, M. Dumontier and P. Lambin. 'Privacy preserving distributed learning classifiers – Sequential learning with small sets of data'. In: *Computers in Biology and Medicine* 136 (Sept. 2021). ISSN: 18790534. DOI: 10.1016/j.compbiomed.2021.104716.

[50] H. Chen, S. Mohapatra, G. Michalopoulos, X. He and I. Mckillop. 'Federated deep learning architecture for personalized healthcare'. In: *Public Health and Informatics: Proceedings of MIE 2021* (July 2021), pp. 193–197. DOI: 10.3233/SHTI210147.

[51] H. Lee, Y. J. Chai, H. Joo, K. Lee, J. Y. Hwang, S.-M. Kim, K. Kim, I.-C. Nam, J. Y. Choi, H. W. Yu, M.-C. Lee, H. Masuoka, A. Miyauchi, K. E. Lee, S. Kim and H.-J. Kong. 'Federated Learning for Thyroid Ultrasound Image Analysis to Protect Personal Information: Validation Study in a Real Health Care Environment.' In: *JMIR medical informatics* 9.5 (May 2021), e25869. ISSN: 2291-9694. DOI: 10.2196/25869. URL: http://www.ncbi.nlm.nih.gov/pubmed/33858817%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8170555.

[52] A. Aminifar, F. Rabbi, V. K. I. Pun and Y. Lamo. 'Monitoring Motor Activity Data for Detecting Patients' Depression Using Data Augmentation and Privacy-Preserving Distributed Learning'. In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (2021), pp. 2163–2169. ISSN: 1557170X. DOI: 10.1109/EMBC46164.2021.9630592.

[53] S. Chen, D. Xue, G. Chuai, Q. Yang and Q. Liu. 'FL-QSAR: A federated learning-based QSAR prototype for collaborative drug discovery'. In: *Bioinformatics* 36.22-23 (Dec. 2020), pp. 5492–5498. DOI: 10.1093/bioinformatics/btaa1006.

[54] P. Dluhoš, D. Schwarz, W. Cahn, N. van Haren, R. Kahn, F. Španiel, J. Horáček, T. Kašpárek and H. Schnack. 'Multi-center machine learning in imaging psychiatry: A meta-model approach'. In: *NeuroImage* 155 (July 2017), pp. 10–24. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2017.03.027.

[55] A. Jochems, T. M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin and A. Dekker. 'Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept'. In: *Radiotherapy and Oncology* 121Jochems.3 (Dec. 2016), pp. 459–467. ISSN: 18790887. DOI: 10.1016/j.radonc.2016.10.002.

[56] Y. P. H. Chin, W. Song, C. E. Lien, C. H. Yoon, W.-C. Wang, J. Liu, P. A. Nguyen, Y. T. Feng, L. Zhou, Y. C. J. Li and D. W. Bates. 'Assessing the International Transferability of a Machine Learning Model for Detecting Medication Error in the General Internal Medicine Clinic: Multicenter Preliminary Validation Study.' In: *JMIR medical informatics* 9.1 (Jan. 2021), e23454. ISSN: 2291-9694. DOI: 10.2196/23454. URL: http://www.ncbi.nlm.nih.gov/pubmed/33502331%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7875695.

[57] R. R. Lopes, M. Mamprin, J. M. Zelis, P. A. L. Tonino, M. S. van Mourik, M. M. Vis, S. Zinger, B. A. J. M. de Mol, P. H. N. de With and H. A. Marquering. 'Local and Distributed Machine Learning for Inter-hospital Data Utilization: An Application for TAVI Outcome Prediction.' In: *Frontiers in cardiovascular medicine* 8 (Nov. 2021), p. 787246. ISSN: 2297-055X. DOI: 10.3389/fcvm.2021.787246. URL: http://www.ncbi.nlm.nih.gov/pubmed/34869698%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8632813.

[58] K. V. Sarma, S. Harmon, T. Sanford, H. R. Roth, Z. Xu, J. Tetreault, D. Xu, M. G. Flores, A. G. Raman, R. Kulkarni, B. J. Wood, P. L. Choyke, A. M. Priester, L. S. Marks, S. S. Raman, D. Enzmann, B. Turkbey, W. Speier and C. W. Arnold. 'Federated learning improves site performance in multicenter deep learning without data sharing'. In: *Journal of the American Medical Informatics Association* 28.6 (June 2021), pp. 1259–1264. ISSN: 1527974X. DOI: 10.1093/jamia/ocaa341.

[59] S. Warnat-Herresthal *et al.* 'Swarm Learning for decentralized and confidential clinical machine learning'. In: *Nature* 594.7862 (June 2021), pp. 265–270. ISSN: 14764687. DOI: 10.1038/s41586-021-03583-3.

[60] U. Norinder, O. Spjuth and F. Svensson. 'Synergy conformal prediction applied to large-scale bioactivity datasets and in federated learning.' In: *Journal of cheminformatics* 13.1 (Oct. 2021), p. 77. ISSN: 1758-2946. DOI: 10.1186/s13321-021-00555-7. URL: http://www.ncbi.nlm.nih.gov/pubmed/34600569%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8487527.

[61] S. Rajendran, J. S. Obeid, H. Binol, R. D'Agostino, K. Foley, W. Zhang, P. Austin, J. Brakefield, M. N. Gurcan and U. Topaloglu. 'Cloud-Based Federated Learning Implementation Across Medical Centers'. In: *JCO Clinical Cancer Informatics* 5 (Jan. 2021), pp. 1–11. ISSN: 24734276. DOI: 10.1200/cci.20.00060.

[62] A. Hanif, C. Lu, K. Chang, P. Singh, A. S. Coyner, J. M. Brown, S. Ostmo, R. V. P. Chan, D. Rubin, M. F. Chiang, J. Kalpathy-Cramer, J. P. Campbell and Imaging and Informatics in Retinopathy of Prematurity Consortium. 'Federated Learning for Multicenter Collaboration in Ophthalmology: Implications for Clinical Diagnosis and Disease Epidemiology.' In: *Ophthalmology. Retina* (Mar. 2022). ISSN: 2468-6530. DOI: 10.1016/j.oret.2022.03.005. URL: http://www.ncbi.nlm.nih.gov/pubmed/35304305.

[63] T. M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker and P. Lambin. 'Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT'. In: *Clinical and Translational Radiation Oncology* 4 (June 2017), pp. 24–31. ISSN: 24056308. DOI: 10.1016/J.CTRO.2016.12.004.

[64] L. Tagliaferri, C. Gobitti, G. F. Colloca, L. Boldrini, E. Farina, C. Furlan, F. Paiar, F. Vianello, M. Basso, L. Cerizza, F. Monari, G. Simontacchi, M. A. Gambacorta, J. Lenkowicz, N. Dinapoli, V. Lanzotti, R. Mazzarotto, E. Russi and M. Mangoni. 'A new standardized data collection system for interdisciplinary thyroid cancer management: Thyroid COBRA'. In: *European Journal of Internal Medicine* 53 (July 2018), pp. 73–78. ISSN: 18790828. DOI: 10.1016/j.ejim.2018.02.012.

[65] M. Hajder, P. Hajder, T. Gil, M. Krzywda, J. Kolbusz and M. Liput. 'Architecture and organization of a Platform for diagnostics, therapy and post-covid complications using AI and mobile monitoring.' In: *Procedia computer science* 192 (2021), pp. 3711–3721. ISSN: 1877-0509. DOI: 10.1016/j.procs.2021.09.145. URL: http://www.ncbi.nlm.nih.gov/pubmed/34630754%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8486225.

[66] O. Samuel, A. B. Omojo, A. M. Onuja, Y. Sunday, P. Tiwari, D. Gupta, G. Hafeez, A. S. Yahaya, O. J. Fatoba and S. Shamshirband. 'IoMT: A COVID-19 Healthcare System driven by Federated Learning and Blockchain.' In: *IEEE journal of biomedical and health informatics* PP (Jan. 2022). ISSN: 2168-2208. DOI: 10.1109/JBHI.2022.3143576. URL: http://www.ncbi.nlm.nih.gov/pubmed/35041615.

[67] Y. Chang, C. Fang and W. Sun. 'A blockchain-based federated learning method for smart healthcare'. In: *Computational Intelligence and Neuroscience* 2021 (2021). ISSN: 16875273. DOI: 10.1155/2021/4376418.

[68] A. Lakhan, M. A. Mohammed, J. Nedoma, R. Martinek, P. Tiwari, A. Vidyarthi, A. Alkhayyat and W. Wang. 'Federated-Learning Based Privacy Preservation and Fraud-Enabled Blockchain IoMT System for Healthcare.' In: *IEEE journal of biomedical and health informatics* PP (Apr. 2022), pp. 1–1. ISSN: 2168-2208. DOI: 10.1109/JBHI.2022.3165945. URL: http://www.ncbi.nlm.nih.gov/pubmed/35394919.

[69] N. Nasser, Z. M. Fadlullah, M. M. Fouda, A. Ali and M. Imran. 'A lightweight federated learning based privacy preserving B5G pandemic response network using unmanned aerial vehicles: A proof-of-concept.' In: *Computer networks* 205 (Mar. 2022), p. 108672. ISSN: 1872-7069. DOI: 10.1016/j.comnet.2021.108672. URL: http://www.ncbi.nlm.nih.gov/pubmed/35023995%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8702301.

[70] M. M. Salim and J. H. Park. 'Federated Learning-based secure Electronic Health Record sharing scheme in Medical Informatics.' In: *IEEE journal of biomedical and health informatics* PP (May 2022), pp. 1–1. ISSN: 2168-2208. DOI: 10.1109/JBHI.2022.3174823. URL: http://www.ncbi.nlm.nih.gov/pubmed/35560105.

[71] P. Mangold, A. Filiot, M. Moussa, V. Sobanski, G. Ficheur, P. Andrey and A. Lamer. 'A decentralized framework for biostatistics and privacy concerns'. In: *Studies in Health Technology and Informatics* 275 (Nov. 2020), pp. 137–141. ISSN: 18798365. DOI: 10.3233/SHTI200710.

[72] U. Ahmed, J. C. W. Lin and G. Srivastava. 'Hyper-Graph Attention Based Federated Learning Method For Mental Health Detection'. In: *IEEE Journal of Biomedical and Health Informatics* PP (May 2022), pp. 1–1. ISSN: 21682208. DOI: 10.1109/JBHI.2022.3172269. URL: http://www.ncbi.nlm.nih.gov/pubmed/35503851.

[73] 'Federated Mining of Interesting Association Rules over EHRs'. In: *Studies in Health Technology and Informatics*. Vol. 287. IOS Press BV, Nov. 2021, pp. 3–7. ISBN: 978-1-64368-236-5. DOI: 10.3233/SHTI210799.

[74] J. Ma, Q. Zhang, J. Lou, L. Xiong and J. C. Ho. 'Communication efficient federated generalized tensor factorization for collaborative health data analytics'. In: *The Web Conference 2021 - Proceedings of the World Wide Web Conference, WWW 2021*. 2021, pp. 171–182. ISBN: 978-1-4503-8312-7. DOI: 10.1145/3442381.3449832.

[75] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola and J. S. Duncan. 'Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results'. In: *Medical Image Analysis* 65 (Oct. 2020). ISSN: 13618423. DOI: 10.1016/j.media.2020.101765. arXiv: 2001.05647.

[76] R. Bey, R. Goussault, F. Grolleau, M. Benchoufi and R. Porcher. 'Fold-stratified cross-validation for unbiased and privacy-preserving federated learning'. In: *Journal of the American Medical Informatics Association* 27.8 (Aug. 2020), pp. 1244–1251. ISSN: 1527974X. DOI: 10.1093/jamia/ocaa096. arXiv: 2001.08090.

[77] M. J. Edmondson, C. Luo, R. Duan, M. Maltenfort, Z. Chen, K. Locke, J. Shults, J. Bian, P. B. Ryan, C. B. Forrest and Y. Chen. 'An efficient and accurate distributed learning algorithm for modeling multi-site zero-inflated count outcomes'. In: *Scientific Reports* 11.1 (Dec. 2021). ISSN: 20452322. DOI: 10.1038/s41598-021-99078-2.

[78] O. Zolotareva, R. Nasirigerdeh, J. Matschinske, R. Torkzadehmahani, M. Bakhtiari, T. Frisch, J. Späth, D. B. Blumenthal, A. Abbasinejad, P. Tieri, G. Kaissis, D. Rückert, N. K. Wenke, M. List and J. Baumbach. 'Flimma: a federated and privacy-aware tool for differential gene expression analysis'. In: *Genome Biology* 22.1 (Dec. 2021). ISSN: 1474760X. DOI: 10.1186/s13059-021-02553-2.

[79] A. Nichol and J. Schulman. 'Reptile: a scalable metalearning algorithm'. In: *arXiv preprint arXiv:1803.02999* 2.3 (2018), p. 4.

[80] A. Thakur, P. Sharma and D. A. Clifton. 'Dynamic Neural Graphs Based Federated Reptile for Semi-Supervised Multi-Tasking in Healthcare Applications'. In: *IEEE Journal of Biomedical and Health Informatics* 26.4 (Apr. 2022), pp. 1761–1772. ISSN: 21682208. DOI: 10.1109/JBHI.2021.3134835.

[81] D. Yang, Z. Xu, W. Li, A. Myronenko, H. R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang, W. Zhu, G. Carrafiello, F. Patella, M. Cariati, H. Obinata, H. Mori, K. Tamura, P. An, B. J. Wood and D. Xu. 'Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan'. In: *Medical Image Analysis* 70 (May 2021). ISSN: 13618423. DOI: 10.1016/j.media.2021.101992. arXiv: 2011.11750.

[82] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly *et al.* 'PLINK: a tool set for whole-genome association and population-based linkage analyses'. In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[83] R. Nasirigerdeh, R. Torkzadehmahani, J. Matschinske, T. Frisch, M. List, J. Späth, S. Weiss, U. Völker, E. Pitkänen, D. Heider, N. K. Wenke, G. Kaissis, D. Rueckert, T. Kacprowski and J. Baumbach. 'sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies'. In: *Genome Biology* 23.1 (Dec. 2022). ISSN: 1474760X. DOI: 10.1186/s13059-021-02562-1.

[84] H. Zhu, J. Xu, S. Liu and Y. Jin. 'Federated learning on non-IID data: A survey'. In: *Neurocomputing* 465 (2021), pp. 371–390. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.07.098. URL: https://www.sciencedirect.com/science/article/pii/S0925231221013254.

[85] T. V. Nguyen, M. A. Dakka, S. M. Diakiw, M. D. VerMilyea, M. Perugini, J. M. M. Hall and D. Perugini. 'A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical data.' In: *Scientific reports* 12.1 (May 2022), p. 8888. ISSN: 2045-2322. DOI: 10.1038/s41598-022-12833-x. URL: http://www.ncbi.nlm.nih.gov/pubmed/35614106.

[86] S. W. Remedios, J. A. Butman, B. A. Landman and D. L. Pham. 'Federated Gradient Averaging for Multi-Site Training with Momentum-Based Optimizers.' In: *Lecture notes-monograph series* 12444 (Oct. 2020), pp. 170–180. ISSN: 0749-2170. DOI: 10.1007/978-3-030-60548-3_17. URL: http://www.ncbi.nlm.nih.gov/pubmed/34531637%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8442829.

[87] X. Yin. 'A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions'. In: (2021). DOI: 10.1145/3460427. URL: https://doi.org/10.1145/3460427.

[88] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. Quek and H. V. Poor. 'Federated learning with differential privacy: Algorithms and performance analysis'. In: *IEEE transactions on information forensics and security* 15 (2020), pp. 3454–3469.

[89] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal and K. Seth. 'Practical secure aggregation for federated learning on user-held data'. In: *arXiv preprint arXiv:1611.04482* (2016).

[90] R. Shao, H. He, Z. Chen, H. Liu and D. Liu. 'Stochastic Channel-Based Federated Learning With Neural Network Pruning for Medical Data Privacy Preservation: Model Development and Experimental Validation.' In: *JMIR formative research* 4.12 (Dec. 2020), e17265. ISSN: 2561-326X. DOI: 10.2196/17265. URL: http://www.ncbi.nlm.nih.gov/pubmed/33350391%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7909896.

[91] T. Han, S. Nebelung, C. Haarburger, N. Horst, S. Reinartz, D. Merhof, F. Kiessling, V. Schulz and D. Truhn. 'Breaking medical data sharing boundaries by using synthesized radiographs'. In: *Science Advances* 6.49 (Dec. 2020). ISSN: 23752548. DOI: 10.1126/sciadv.abb7973.

[92] B. Han, R. Jhaveri, H. Wang, D. Qiao and J. Du. 'Application of Robust Zero-Watermarking Scheme Based on Federated Learning for Securing the Healthcare Data.' In: *IEEE journal of biomedical and health informatics* PP (Oct. 2021). ISSN: 2168-2208. DOI: 10.1109/JBHI.2021.3123936. URL: http://www.ncbi.nlm.nih.gov/pubmed/34714760.

[93] T.-T. Kuo and A. Pham. 'Detecting model misconducts in decentralized healthcare federated learning.' In: *International journal of medical informatics* 158 (Dec. 2021), p. 104658. ISSN: 1872-8243. DOI: 10.1016/j.ijmedinf.2021.104658. URL: http://www.ncbi.nlm.nih.gov/pubmed/34923447.

[94] J. Cui, H. Zhu, H. Deng, Z. Chen and D. Liu. 'FeARH: Federated machine learning with anonymous random hybridization on electronic medical records'. In: *Journal of Biomedical Informatics* 117 (May 2021). ISSN: 15320464. DOI: 10.1016/j.jbi.2021.103735.

[95] T. Viering and M. Loog. 'The shape of learning curves: a review'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.6 (2022), pp. 7799–7819.

3

A comprehensive experimental comparison between federated and centralized learning

Federated learning is an upcoming machine learning paradigm which allows data from multiple sources to be used for training of classifiers without the data leaving the source it originally resides. This can be highly valuable for use cases such as medical research, where gathering data at a central location can be quite complicated due to privacy and legal concerns of the data. In such cases, federated learning has the potential to vastly speed up the research cycle. Although federated and central learning have been compared from a theoretical perspective, an extensive experimental comparison of performances and learning behavior still lacks. We have performed a comprehensive experimental comparison between federated and centralized learning. We evaluated various classifiers on various datasets exploring influences of different sample distributions as well as different class distributions across the clients. The results show similar performances under a wide variety of settings between the federated and central learning strategies. Federated learning is able to deal with various imbalances in the data distributions. It is sensitive to batch effects between different datasets when they coincide with location, similar as with central learning, but this setting might go unobserved more easily. Federated learning seems robust to various challenges such as skewed data distributions, high data dimensionality, multiclass problems and complex models. Taken together, the insights from our comparison give much promise for applying federated learning as an alternative to sharing data.

S. Garst, J. Dekker and M. Reinders. 'A comprehensive experimental comparison between federated and centralized learning'. In: Database *2025 (2025), baaf016*

3.1. Introduction

Nowadays lots of data is available for the use of machine learning applications. However, in some use cases data does not naturally reside at a single location, and centralizing data might be difficult due to regulation or hardware constraints. One example is medical data [1, 2], which is collected at different hospitals or medical institutions, but cannot leave these institutions due to privacy concerns. In order to make use of this type of data, the concept of federated learning [3] was introduced. Instead of gathering the data in a centralized location before training a single model, in a federated learning environment, the model gets send to wherever the data is available: the so-called clients. At these clients, the models undergo some form of training, after which the updated model parameters are sent back to a central point, referred to as the server. The server then aggregates all the local updates in order to create a new global model, which it then sends back to all of the clients, and the cycle repeats. With this setup, only model updates are being communicated, and since the original data never has to leave its origins, the entire process becomes less privacy-sensitive.

The concept of federated learning is analogous to other fields trying to learn from distributed data, particularly distributed optimization [4]. Distributed optimization operates similarly to federated learning. The major difference is that federated learning generally uses a star network, with a centralized server connected to each client, whereas such a central point is usually not present in distributed learning, in which clients are only connected to (some) other clients. For a more in-depth analysis of distributed learning we refer the interested reader to [5].

A seminal algorithm for federated learning was Federated Averaging (fedAVG, [3]), where a weighted average of the model parameters is taken at the central server each communication round. Since the introduction of federated learning in 2017 , lots of research has been done on its efficiency, performance and privacy-preserving properties [6], [7]. Communication ends up being a bottleneck for efficiency in many federated systems [8]. As a result, techniques have been developed in order to reduce communication rounds [9], [10]. With regards to performance, many analyses have been made on the performance of federated averaging (fedAVG)[3]. These studies focus usually on performance under poorer data distributions [11], [12], [13], i.e. where data is not independent and identically distributed (IID) at the different clients, as this is the area where fedAVG performance seems to deteriorate [14]. As a result, extensions of federated averaging trying to accommodate for different (non-IID) data distributions have been developed, e.g. [13], [15]. Privacy preservation has been explored by means of constructing specific attacks on federated systems [16], [17]. As a response, extensions on the original federated algorithms that include some form of increased privacy preservation is becoming a vast area of research [18].

Although all of the aforementioned studies have developed federated learning into a vast research area, many analyses remain mostly theoretical. Besides, the aforementioned works assume that the use of a federated approach is a given. However whether or not to use federated learning is often an important design decision.

One example is the medical research field. There have been various efforts over the years to pool medical data from multiple institutes in order to create a large data resource [19] [20]. However, the efforts required to create such resources can not be understated. In these cases, one might want to know whether a federated setup will give comparable performance to central models, or whether it is beneficial to create a pooled database.

Recently, papers comparing a federated setting with centralized baselines have been emerging, e.g. [21], [22]. However, most comparisons only include one or few classifiers, and/or distributions of the data. Therefore, we set out to design a comprehensive set of experiments in which centralized and federated models are compared to one another.

We explore the use of different classifiers (six in total) on multiple datasets (five in total), distributed in several ways. Specifically, we start with a toy problem by comparing performances between federated and centralized classifiers on subsets of MNIST, as well as fashion MNIST. Next, we move to more realistic scenarios, with data collected from various locations. These datasets were selected such that they have a variety of data types: tabular (RNA abundance), sequential (kinase activation), and 3D imaging (cardiovascular disease classification). As such, we shed some light on different scenarios in which federated learning might perform similarly to a centralized model, and when to be careful in assuming such a similar performance.

3.2. Results

3.2.1. Linear models on a binary classification problem show a relation between the learning rate and amount of clients used.

The MNIST dataset (images of handwritten digits) was chosen as a first dataset, as it is known to be a relatively easy problem for modern day classifiers [23]. In order to simplify, MNIST was converted into a binary classification problem (MNIST2) by selecting only two of the classes. This dataset was distributed evenly (IID) among ten clients, meaning that each client had a similar amount of samples, with an even class distribution (see figure 3.1d, methods. A Logistic Regression (LR) and Support Vector Machine (SVM) were trained, with varying learning rates for both the central and federated model. Results for the LR model can be found in figure 3.1a. These results show that increasing the learning rate leads to faster convergence times, both in the federated and centralized case. There is a consistent factor of ten between the federated and centralized models, i.e. a federated classifier with a learning rate of 0.5 seems to give a comparable result as a central classifier with learning rate 0.05. This was observed for both the LR and SVM models (results on SVM can be found in supplementary figure 3.6). This can be explained by the amount of clients used, which is also ten.

A mathematical motivation behind this is given in supplement 3.5.1, but the intuition is as follows: a single SGD step in a central model is based on the sum of the loss function over all S available data samples. In a federated setting, these S data points are distributed over N clients, meaning that each client gives an update based on only $\frac{S}{N}$ samples (on average). As these updates are only getting averaged, this is analogous to dividing the learning rate by a factor N.

3.2.2. Extension towards more complex models shows different relationships between federated and centralized models.

We were interested whether our findings would be different when more complex classifiers were adopted. Hereto, we examined a Fully connected Neural Net (FNN), a Convolutional Neural Net (CNN), and a decision tree (GBDT) classifier (implementation details in methods). Figure 3.1b (supplementary table 3.4) shows the Area Under Curve (AUC) values for the accuracy curves (i.e. accuracy versus epochs/communication rounds). This metric is indicative of how the convergence of two classifiers compare, especially if the convergence accuracy is equal, which is shown in the top part of the plot. We ran each classifier four times, resulting in the boxplots in figure 3.1b. This figure shows that also for the neural network based models (FNN, CNN), similar performance is being reached when comparing the federated (IID) and central models of the same classifier.The GBDT shows a slight difference between the federated and central versions. This could be due to the GBDT using a different federated update scheme as all other classifiers (Algorithm 2, methods). The CNN shows some difference in AUC of the accuracy curve between the central and federated setting. This highlights a limitation of this metric, as it is caused by a slightly faster start of convergence in the federated case, as can be seen in supplementary figure 3.7. Furthermore, the convergence accuracy is similar on average, though a higher variance is observed in the central case. Curiously, the relation between learning rate and amount of clients that was observed in the first set of experiments does not seem to be present for the neural network-based classifiers, i.e. these had the same learning rate between federated and central experiments. We speculate that this difference is due to the non-convexity of the neural networks.

3.2.3. Binary classification experiments show robustness to sample and class distributions.

In the next experiment, the robustness to a change in distribution between clients was tested. Hereto, the MNIST2 dataset was split with a sample imbalance across the clients (sample imbalance, SI), i.e. the first client only holds a small fraction of the data, the second client a slightly larger fraction etc. (figure 3.1e). We also considered another imbalance, in which the distribution of classes across clients was skewed (class imbalance, CI) (figure 3.1f), i.e. some clients have more samples from one class and others have more samples from the other class, and in between.

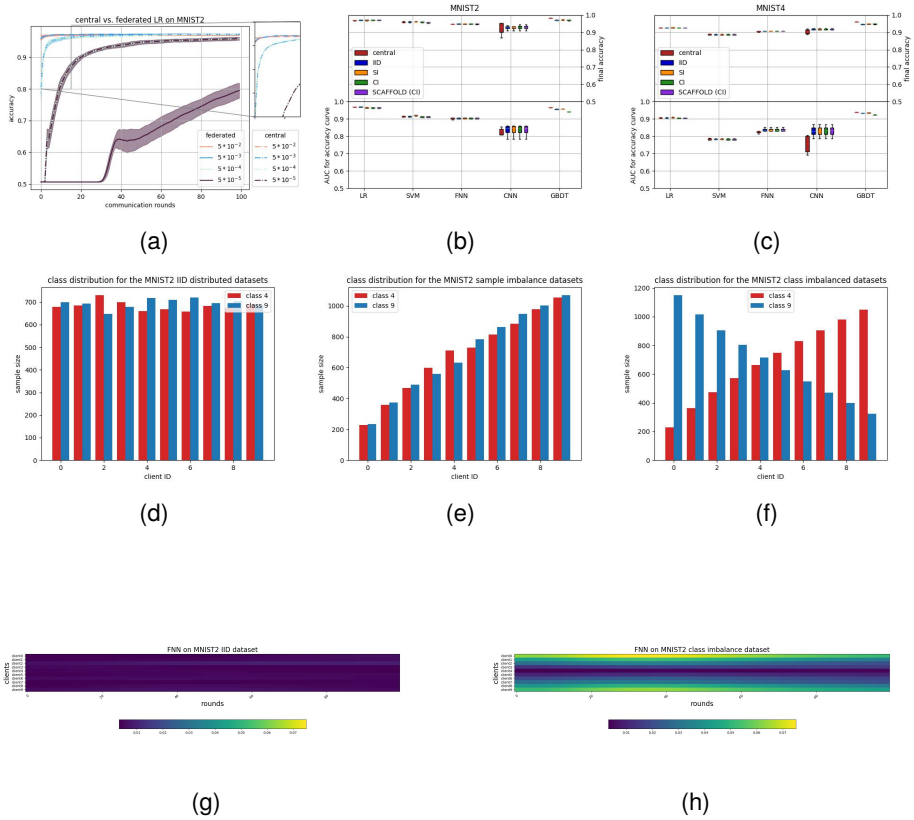


Figure 3.1.: Results on MNIST datasets. (a) shows the relation between federated and centralized learning rate for the Logistic Regressor for the MNIST2 dataset. Numbers in the legend correspond to the learning rates. (b)-(c) show a comparison between federated and centralized classifiers for MNIST2 and MNIST4, respectively. (d) - (f) show the various distributions for the MNIST2 dataset (IID, SI, CI, respectively). (g) and (h) show the heatmaps of the euclidean distance between the local and global model parameters after each communication round for the training on the IID and CI distributions, respectively.

Figure 3.1b and supplementary table 3.4 shows that, with the exception for the GBDT model, similar performances are being attained for the SI and CI settings compared to the IID setting. The SVM does seem to have a slightly higher AUC for the SI distribution, though this is not visible in final accuracy. After observing the accuracy curves (supplementary figure 3.8), this difference seems to come from a slightly faster convergence rate for the SI distribution (about 1 epoch difference).

Although no performance differences are observed in the imbalance settings, we were interested on how both imbalance settings influenced the learning process. Hereto we visualized the model parameters at the client and at the server at each communication round. In the IID setting (figure 3.1g) one can observe that the updated model parameters at the client are more or less similarly changing across the epochs. Interestingly, for the class imbalance (CI) setting this is not the case, see figure 3.1h. Here, we see that the clients with the larger class imbalance have model updates that are consistently further from the global model.

3.2.4. Extension to multiclass problem results in similar performance as compared to binary classification.

To explore the effect of having a multiclass problem, we created a four class dataset out of the original MNIST dataset; MNIST4 (see methods - datasets). When splitting the samples evenly (IID) across the clients (in number of samples as well as class distributions), the linear models achieve similarly shaped accuracy curves for the federated and central versions, as indicated by the AUC's in figure 3.1c (supplementary table 3.4). The neural network based models show differences in these AUCS. In the case of the CNN, this difference is even visible in the final accuracy. Then we distributed MNIST4 with either sample imbalance (SI) or class imbalance (CI), similar to the two class experiments (see supplementary figure 3.9). The same robustness to the difference in sample and class distributions are observed as in the two class experiments.

3.2.5. SCAFFOLD model updating performs similar to fedAVG model updating.

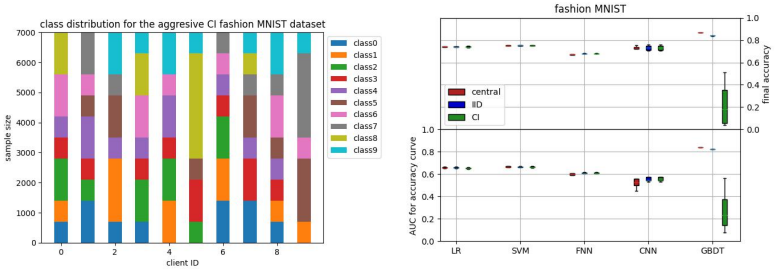
To be more robust to class imbalances across clients an alternative model updating scheme with respect to fedAVG was introduced: SCAFFOLD ([13], supplementary materials 3.5.2). We find that the performance of SCAFFOLD (3.1b and 3.1c) is comparable to fedAVG for the generated distributions of the MNIST2 and MNIST4 datasets. This might have been expected, as fedAVG shows similar performance in the CI case compared to the IID case (so no deterioration in performance that could be repaired by SCAFFOLD).

3.2.6. Increasing class distribution aggressiveness gives varying results.

After our first experiments on the MNIST datasets, we continued with the fasion MNIST dataset (images of 10 classes of different clothing) as this has been designed to be a harder classification task [24]. When distributing evenly across 10 clients (IID), all models behave similarly as for the more simple MNIST classification tasks (figure 3.2);

there is no visible difference between the central and federated versions of the linear models, the FNN shows a slight difference in AUC but not in final accuracy, and the CNN shows a larger difference in AUC, but not in final accuracy. For the GBDT, some difference can be seen in both AUC and final accuracy.

We then tested a more aggressive class imbalance by distributing the fashion MNIST across clients in such a way that the available classes for each client are not necessarily overlapping, i.e. some clients have training examples of a particular class whereas other clients do not (see figure 3.2a). Somewhat surprisingly, differences between the IID setting and the more aggressive CI setting were negligible for most classifiers, except the GBDT, which breaks down for this aggressive CI setting. This might be related to the updating scheme in which the decision tree is updated for each client sequentially (Algorithm 2, methods) instead of simultaneously for all other classifiers (Algorithm 1, methods).



(a) (b)

Figure 3.2.: Results on fashion MNIST. (a) shows the data distribution for the aggressive CI distribution for the fashion MNIST dataset. (b) shows results on both IID and CI distributions.

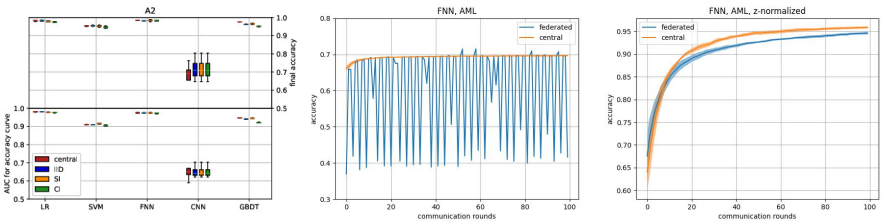
3.2.7. Experiments on high dimensional data show potential for federated data processing techniques.

Until now, all experiments have utilized a single dataset which has then been split into multiple pieces. A more realistic scenario would use multiple datasets, each gathered on a separate client. Hereto, we used datasets that were gathered to predict whether a patient has Acute Myeloid Leukemia (AML) based on their measured transcriptome ([25], methods). It consists of three different datasets (A1-A3), with in total 11.500 patients for which the expression of 12.709 genes are measured. For two of the datasets (A1, A2) the expression is measured using microarray technology, and for the third dataset (A3) this was done using current RNA sequencing technology.

In order to determine the feasibility of learning on these datasets, we first analysed A2 only. A2 was distributed in an IID fashion over ten clients. As this data has a high dimensionality (over 12.000 features), a dimensionality reduction method seemed to be essential as all classifiers could not perform above random level in the original feature domain (this was also the case for the central models). Therefore, a Principal Component Analysis (PCA) was performed beforehand, reducing the data to 100 features. For this we implemented the PCA in a federated manner (adapted from [26], see supplementary algorithm 3.5.3). With the reduced dataset, all classifiers perform similarly to what we observed in earlier experiments (figure 3.3a). Next, we also tested the SI and CI distributions (before reapplying the federated PCA). Also for this distributed A2 dataset we observe a high robustness to the different data distributions, see figure 3.3a and supplementary table 3.5.

3.2.8. Usage of heterogeneous datasets highlights the need to 'think federated'

Then we returned to the combined datasets A1-A3. In this setup, only three clients have been used, each holding one complete dataset. Also in this setting dimensionality reduction using federated PCA was done before training the classifiers. The Logistic Regressor seems to perform fine, although the relation between client amount and learning rate seems to have disappeared (supplementary figure 3.10). However, whereas the centralized neural network manages to improve slightly over time, the federated neural network now breaks down. (see figure 3.3b). We therefore dived deeper into possible causes of this breakdown.



(a) (b) (c)

Figure 3.3.: Results on the AML datasets. (a) AUC performance of the five classifiers across different splits of dataset A2 across ten clients, (b) FNN results when all three AML datasets A1-A3 are used with the original data. (c) FNN results on A1-3 using the locally scaled data.

This breakdown might have been caused by difference in measuring technology, causing the measured features to be distributed differently between dataset A3 (sequencing technology) and the datasets A1 and A2 (microarray technology).

To correct for the different feature distributions between the datasets, we normalised features for all datasets locally, similar to a batch correction scheme, by z-normalizing each gene before applying the federated PCA (supplementary algorithm 4). After this batch correction, performance of the federated and central classifiers are similar again (figure 3.3c, supplementary figure 3.11). Although technical batch effect such as these can occur in a centralized setting as well, the nature of federated data being 'unseen', means that such batch effects could go unnoticed. Moreover, it can be tempting to view a federated dataset as one dataset (which happens do be distributed). These results underline the importance of treating distributed datasets as separate entities, instead of as subsets of a larger dataset.

3.2.9. Experiments on 'real life' distributed datasets show increased variance for federated models.

Next, we explored a dataset consisting of molecules that are described by their molecular fingerprints (deMorgan fingerprints, see methods for details) which can be used to predict the activation or inhibition of certain kinases [27]. The dataset is compiled out of three different studies, which we treated as different clients. Two different kinases, KDR and ABL1, were selected as targets for classification, resulting in two datasets. Note that, due to sparsity in the data, these two datasets are not equal in size (for each client), see figures 3.4a and 3.4d for data distributions. After a brief analysis of the datasets, two interesting aspects came to mind. First, the distributions shown in figure 3.4(a) and 3.4(d) could be seen as a combination between a class -and sample imbalanced dataset, thereby increasing complexity w.r.t. earlier distributions. Second, we looked at a combined tSNE plot (figures 3.4b and 3.4e), which show a large amount of small clusters, most of which are only present at one of the datasets.

Figures 3.4c and 3.4f show the results for kinase KDR and ABL1, respectively, as well as supplementary table 3.3. For KDR, convergence behavior seems similar, albeit with a higher variance for the linear models, with the exception of CNN and GBDT. However, some difference in final accuracy can be seen for the linear models as well. CNN seems to outperform on KDR with the given metric, which can also somewhat be observed in Supplementary figure 3.12. For this experiment, minibatch learning was used for the CNN experiments (both federated and centralized), which could explain this behaviour.

For ABL1, there is a larger difference in convergence for the linear models between the federated and centralized settings. However, final accuracy is still similar, as can also be seen in supplementary figure 3.13. Curiously, both FNN and CNN seem to converge similarly this time, though there is a slightly higher final accuracy for the CNN.

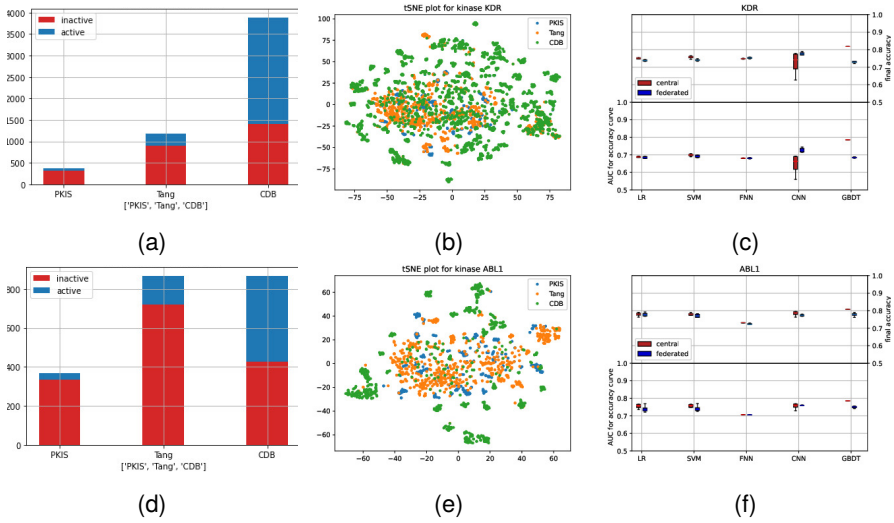


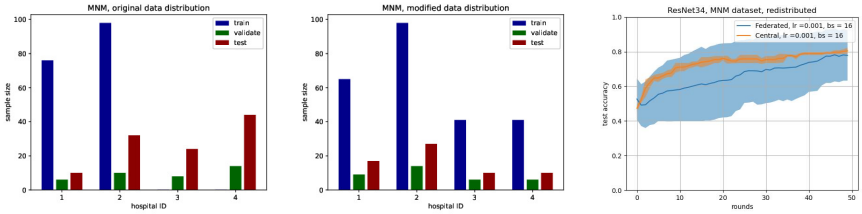
Figure 3.4.: Results on kinase dataset for two kinases, KDR and ABL1. (a) and (d) show data distributions, (b) and (e) tSNE plots (every dot is a molecule, color indicates study), and (c) and (f) show performance results for the different classifiers for the central and federated setting.

3.2.10. Complex federated dataset shows capability for larger neural networks.

To compare performances on a more complex, distributed dataset, we used the publicly available dataset for the MNM challenge from [28]. 3D MRI scans of the heart are classified as either a healthy patient, or as a patient having one of various cardiovascular diseases. We selected only MRI scans of healthy patients and patients with hypertrophic cardiomyopathy (HCM), which was the largest class of disease, simplifying the task to a binary classification.

After initial testing, it became clear that the structure of the 3D MRI scans was too complex for the simple linear and shallow neural networks to achieve reasonable performances. We therefore adapted the ResNet-34 model by adjusting the input layer to accommodate the size of the scans, as well as by modifying the output layer to account for the amount of classes. For the federated experiment, the dataset was split based on hospital of origin. This resulted in four clients (the fifth hospital did not have any samples of the classes we selected), see figure 3.5a. The train/test split chosen by the original authors proved to be exceptionally challenging without any data augmentation strategies, as performed in earlier work on the same dataset [28] [29]. We therefore changed the train/test split as shown in figure 3.5b (note that data stayed at its original 'location', only local changes to the train/test splits were made).

Figure 3.5c shows the comparison between the federated and centralized experiments. This shows, although the learning behaviour between the two settings is different, both the central and federated approach converge to a similar accuracy, indicating that federated learning also works well on larger (neural) networks, which is in line with earlier research [29].



(a) (b) (c)

Figure 3.5.: Results on the MNM dataset. (a) and (b) show distributions of the MNM dataset, originally and after redistribution, respectively. (c) shows the acccuracy curve of ResNet 34 on the distribution from (b) (bs = batch size).

3.3. Conclusion

This paper describes the results of a set of experiments exploring the differences between centralized and federated machine learning models. Multiple classifiers were used on several datasets, which have been distributed in different ways. Results show that, especially under IID circumstances, a federated classifier could reach similar performances compared to its centralized counterpart. However, a careful choice of parameters such as learning rate and batch size is vital to reach comparable performance. The exact relation between those parameters for reaching equal performance between centralized and federated models remains unclear and could be part of future work, although we have shown that the learning rate can be adapted taking into account the number of clients. Furthermore we have illustrated that when using datasets from multiple sources, one can suffer from batch effects between the datasets. Although this affects both a federated as well as a central model, we advocate to be extra careful on this aspect in the federated case as the data distributions are not trivially compared in this setting.

For all MNIST variants, fashion MNIST, A2 and both kinase datasets, the same set of classifiers was used. This allowed us to compare performance differences between central and federated classifiers across datasets for all models. It does, however, create the situation that we applied classifiers on data types for which such a specific classifier might not be a good choice. However, since our goal was not to find optimal performance on the studied datasets, but rather compare federated with central models, we deemed this situation acceptable.

Nevertheless, our results should not be seen as a recommendation on which type of classifier to use for a specific data type.

The federated GBDT-classifier breaks down on the aggressive class-imbalanced distribution of the fashion MNIST dataset, in which each client has approximately half of the total amount of classes. The inPrivate learning algorithm utilizes decision trees created by the previous clients to boost. If these trees were created using a different subset of classes, it can be understood that the resulting gradient does not serve as a helpful tool for the creation of the next decision tree.

For the AML dataset we have shown the necessity of dimension reduction to achieve reasonable performance for the classifiers studied. We should note that in the original work [30], no form of dimension reduction was used. This might have been possible because they used a deep neural network including multiple dropout layers, which are known for being helpful at dealing with highly dimensional data [31]. In contrast, the neural network used in this paper only consisted of two linear layers, combined with a relu-layer.

One of the strengths of the federated averaging algorithm is that it allows for multiple local epochs, as well as batch learning. This results in multiple learning steps per communication round, potentially reducing the total amount of communication rounds required [3]. In our work, we did not explore the amount of local epochs or batch learning in the federated setting. A further analysis on differences between federated and central models could include experiments exploring the influence of varying these parameters.

One limitation of the results on all MNIST datasets is that it is composed of one original dataset. This means that there are no concept shifts (i.e. batch effects) between the different clients, even in the class -and sample imbalanced cases, as can be observed in the AML dataset. Another limitation is the lack of calibration. Calibration is especially interesting in a federated learning case, as there can be a high variance in class distribution between clients (as has been simulated in this work). How this influences a federated classifier is an interesting direction for future work.

Another limitation to this work is the fact that we simulated a federated setting. This was done because it does not affect the performance metrics compared in this work, being the accuracy differences between federated and centralized models. However, for a comparison on runtime, including an analysis of communication overhead for a federated setup, a simulated environment falls short.

Hospitals could decide to 'personalize' their model, i.e. take a model trained using a federated approach (with or without their local dataset), to then retrain that model on their local data only. Although this might lose some generalizability, it could also result in a more accurate model for the institution itself. Exploring such strategies is an interesting venue for future work as well.

Taken together, our work has shown promising results for federated learning as an alternative to centrally organised learning. We strongly advice that, like in central learning with batch effects, it is important to align data distributions across the different clients. Nevertheless, federated learning might open new possibilities to increase data availability for data hungry machine learning methods.

Bibliography

[1] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian and F. Wang. 'Federated Learning for Healthcare Informatics'. In: *Journal of Healthcare Informatics Research* 5.1 (Mar. 2021), pp. 1–19. ISSN: 2509-498X. DOI: 10.1007/s41666-020-00082-4. URL: https://doi.org/10.1007/s41666-020-00082-4.

[2] A. Jochems, T. M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin and A. Dekker. 'Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept'. In: *Radiotherapy and Oncology* 121.3 (2016), pp. 459–467. ISSN: 0167-8140. DOI: https://doi.org/10.1016/j.radonc.2016.10.002. URL: https://www.sciencedirect.com/science/article/pii/S016781401634336

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y. Arcas. 'Communication-Efficient Learning of Deep Networks from Decentralized Data'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 20–22 Apr 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[4] A. Nedic and A. Ozdaglar. 'Distributed Subgradient Methods for Multi-Agent Optimization'. In: *IEEE Transactions on Automatic Control* 54.1 (2009), pp. 48–61. DOI: 10.1109/TAC.2008.2009515.

[5] T. Yang, X. Yi, J. Wu, Y. Yuan, D. Wu, Z. Meng, Y. Hong, H. Wang, Z. Lin and K. H. Johansson. 'A survey of distributed optimization'. In: *Annual Reviews in Control* 47 (2019), pp. 278–305. ISSN: 1367-5788. DOI: https://doi.org/10.1016/j.arcontrol.2019.05.006. URL: https://www.sciencedirect.com/science/article/pii/S1367578819300082.

[6] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, X. Liu and B. He. 'A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection'. In: *CoRR* abs/1907.09693 (2019). arXiv: 1907.09693. URL: http://arxiv.org/abs/1907.09693.

[7] T. Li, A. K. Sahu, A. Talwalkar and V. Smith. 'Federated Learning: Challenges, Methods, and Future Directions'. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60. DOI: 10.1109/MSP.2020.2975749.

[8] T. Li, A. K. Sahu, A. Talwalkar and V. Smith. 'Federated Learning: Challenges, Methods, and Future Directions'. In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60. DOI: 10.1109/MSP.2020.2975749.

[9] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh and D. Bacon. 'Federated Learning: Strategies for Improving Communication Efficiency'. In: *CoRR* abs/1610.05492 (2016). arXiv: 1610.05492. URL: http://arxiv.org/abs/1610.05492.

[10] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis and S. Kim. 'Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data'. In: *CoRR* abs/1811.11479 (2018). arXiv: 1811.11479. URL: http://arxiv.org/abs/1811.11479.

[11] A. Khaled, K. Mishchenko and P. Richtarik. 'Tighter Theory for Local SGD on Identical and Heterogeneous Data'. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. Ed. by S. Chiappa and R. Calandra. Vol. 108. Proceedings of Machine Learning Research. PMLR, 26–28 Aug 2020, pp. 4519–4529. URL: https://proceedings.mlr.press/v108/bayoumi20a.html.

[12] F. Haddadpour and M. Mahdavi. 'On the Convergence of Local Descent Methods in Federated Learning'. In: *CoRR* abs/1910.14425 (2019). arXiv: 1910.14425. URL: http://arxiv.org/abs/1910.14425.

[13] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich and A. T. Suresh. 'SCAFFOLD: Stochastic Controlled Averaging for Federated Learning'. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by H. D. III and A. Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 5132–5143. URL: https://proceedings.mlr.press/v119/karimireddy20a.html.

[14] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin and V. Chandra. 'Federated Learning with Non-IID Data'. In: *arXiv preprint arXiv:1806.00582* (2018). arXiv: 1806.00582v2.

[15] A. K. Sahu, T. Li, M. Sanjabi, M. Zaheer, A. Talwalkar and V. Smith. 'On the Convergence of Federated Optimization in Heterogeneous Networks'. In: *CoRR* abs/1812.06127 (2018). arXiv: 1812.06127. URL: http://arxiv.org/abs/1812.06127.

[16] J. Geiping, H. Bauermeister, H. Dröge and M. Moeller. 'Inverting Gradients - How easy is it to break privacy in federated learning?' In: *CoRR* abs/2003.14053 (2020). arXiv: 2003.14053. URL: https://arxiv.org/abs/2003.14053.

[17] R. Shokri, M. Stronati and V. Shmatikov. 'Membership Inference Attacks against Machine Learning Models'. In: *CoRR* abs/1610.05820 (2016). arXiv: 1610.05820. URL: http://arxiv.org/abs/1610.05820.

[18] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek and H. V. Poor. 'Federated Learning With Differential Privacy: Algorithms and Performance Analysis'. In: *IEEE Transactions on Information Forensics and Security* 15 (2020), pp. 3454–3469. DOI: 10.1109/TIFS.2020.2988575.

[19] J. Mahita, B. Ha, A. Gambiez, S. L. Schendel, H. Li, K. M. Hastie, S. M. Dennison, K. Li, N. Kuzmina, S. Periasamy, A. Bukreyev, J. E. Munt, M. Osei-Twum, C. Atyeo, J. A. Overton, R. Vita, H. Guzman-Orozco, M. Mendes, M. Kojima, P. J. Halfmann, Y. Kawaoka, G. Alter, L. Gagnon, R. S. Baric, G. D. Tomaras, T. Germann, D. Bedinger, J. A. Greenbaum, E. O. Saphire and B. Peters. 'Coronavirus Immunotherapeutic Consortium Database'. In: *Database* 2023 (Feb. 2023), baac112. ISSN: 1758-0463. DOI: 10.1093/database/baac112. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baac112/49156632/baac112.pdf. URL: https://doi.org/10.1093/database/baac112.

[20] A. Arora, D. Kaur, S. Patiyal, D. Kaur, R. Tomer and G. P. S. Raghava. 'SalivaDB—a comprehensive database for salivary biomarkers in humans'. In: *Database* 2023 (Feb. 2023), baad002. ISSN: 1758-0463. DOI: 10.1093/database/baad002. eprint: https://academic.oup.com/database/article-pdf/doi/10.1093/database/baad002/49113025/baad002.pdf. URL: https://doi.org/10.1093/database/baad002.

[21] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra and K. Lekadir. 'Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease'. In: *Scientific Reports* 12.1 (Dec. 2022). ISSN: 20452322. DOI: 10.1038/s41598-022-07186-4.

[22] M. A. Salam, S. Taha and M. Ramadan. 'COVID-19 detection using federated machine learning'. In: *PLoS ONE* 16.6 June (June 2021). ISSN: 19326203. DOI: 10.1371/journal.pone.0252573.

[23] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner. 'Gradient-based learning applied to document recognition'. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[24] H. Xiao, K. Rasul and R. Vollgraf. 'Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms'. In: *CoRR* abs/1708.07747 (2017). arXiv: 1708.07747. URL: http://arxiv.org/abs/1708.07747.

[25] S. Warnat-Herresthal, K. Perrakis, B. Taschler, M. Becker, K. Baßler, M. Beyer, P. Günther, J. Schulte-Schrepping, L. Seep, K. Klee *et al.* 'Scalable prediction of acute myeloid leukemia using high-dimensional machine learning and blood transcriptomics'. In: *Iscience* 23.1 (2020).

[26] A. Grammenos, R. Mendoza-Smith, J. Crowcroft and C. Mascolo. 'Federated Principal Component Analysis'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6453–6464.

[27] B. Merget, S. Turk, S. Eid, F. Rippmann and S. Fulle. 'Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay'. In: *Journal of Medicinal Chemistry* 60.1 (2017). PMID: 27966949, pp. 474–485. DOI: 10.1021/acs.jmedchem.6b01611. eprint: https://doi.org/10.1021/acs.jmedchem.6b01611. URL: https://doi.org/10.1021/acs.jmedchem.6b01611.

[28] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma *et al.* 'Multi-centre, multi-vendor and multi-disease cardiac segmentation: the M&Ms challenge'. In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3543–3554.

[29] A. Grammenos, R. Mendoza-Smith, J. Crowcroft and C. Mascolo. 'Federated Principal Component Analysis'. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6453–6464.

[30] S. Warnat-Herresthal, H. Schultze and K. L. e. a. Shastry. 'Swarm Learning for decentralized and confidential clinical machine learning'. In: *Nature* 594.7862 (June 2021), pp. 265–270. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03583-3. URL: https://doi.org/10.1038/s41586-021-03583-3.

[31] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever and R. R. Salakhutdinov. 'Improving neural networks by preventing co-adaptation of feature detectors'. In: *arXiv preprint arXiv:1207.0580* (2012).

[32] L. Zhao, L. Ni, S. Hu, Y. Chen, P. Zhou, F. Xiao and L. Wu. 'InPrivate Digging: Enabling Tree-based Distributed Data Mining with Differential Privacy'. In: *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 2018, pp. 2087–2095. DOI: 10.1109/INFOCOM.2018.8486352.

[33] J. Quackenbush. 'Microarray Analysis and Tumor Classification'. In: *New England Journal of Medicine* 354.23 (2006). PMID: 16760446, pp. 2463–2472. DOI: 10.1056/NEJMra042342. eprint: https://doi.org/10.1056/NEJMra042342. URL: https://doi.org/10.1056/NEJMra042342.

[34] R. J. Mackenzie. 'RNA-Seq: Basics, Applications and Protocol'. In: *Technology Networks* (2017). URL: https://www.technologynetworks.com/genomics/articles/rna-seq-basics-applications-and-protocol-299461.

[35] A. Moncada-Torres, F. Martin, M. Sieswerda, J. van Soest and G. Geleijnse. 'VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange'. In: *AMIA Annual Symposium Proceedings*. 2020, pp. 870–877.

[36] F. Martin, M. Sieswerda and e. a. Hasan Alradhi. *vantage6 repository*. Accessed: 2021-09-08. URL: https://doi.org/10.5281/zenodo.3686944.

[37] M. Galarnyk. 'PCA using Python (scikit-learn)'. In: *towards data science* (2017). URL: https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60.

3.4. Methods

3.4.1. Algorithm overview

In total, five different classifiers were implemented in a federated setting: a Logistic Regressor (LR), a Support Vector Machine (SVM), a Fully connected Neural Network (FNN), a Convolutional Neural Network (CNN) and a Gradient-Boosting Decision Tree protocol (GBDT). All classifiers except the GBDT follow the same federated learning scheme, which is given by Algorithm 1. The federated learning algorithm for the GBDT is given by Algorithm 2.

The architectures for the FNN and CNN can be found in supplementary tables 3.1 and 3.2 respectively. A performance comparison with a centralized version of the classifiers was the outcome of interest, rather than optimizing the final performance. Therefore the neural net architectures were kept simple, allowing for faster training. The linear models were implemented using sklearn's SGDClassifier, using default parameters (except for the loss function). Similarly, the GBDT protocol was implemented using sklearn's GradientBoostingClassifier, using default parameters except for *n_estimators*, which was initialized as one, and incremented each iteration to allow for the construction of subsequent trees.

Algorithm 1 works as follows: First, a model is initialized on the server with random values. This model is then sent to all clients. Upon reception, every client performs (in parallel) a local training step to update the coefficients of the local model based on Stochastic Gradient Descent (SGD). All clients then send back their updated model to the central server. At the server, these models are combined to update the global model. Then the next epoch starts, and the combined coefficients from the previous round are now used as the new values to be sent out to all clients. In Algorithm 1, the intermediate models are evaluated on the test set at the server (holding a concatenation of all test sets from all clients). Although this is not realistic, we observed no difference between calculating the performance locally at each client and then averaging the performances, or doing this centrally on the combined test sets. Therefore, for simplicity reasons, we chose to calculate the performances centrally.

Algorithm 1 The federated learning algorithm of all classifiers except GBDT

1: initialize W_0^g randomly
2: **for all** r rounds **do**
3: **On server:** Send W_r^g to all clients i
4: **On clients:** $W_r^i \leftarrow W_r^g$
5: **On clients:** $acc_r^i = acc(W_r^i, X_{test}^i, Y_{test}^i)$
6: **On clients:** $W_{r+1}^i \leftarrow localStep(W_r^i, X_{train}^i, Y_{train}^i)$
7: **On clients:** $s^i \leftarrow size(X_{train}^i)$
8: **On clients:** send W_{r+1}^i, acc_r^i, s_i back to server
9: **On server:** retrieve W_{r+1}^i, acc_r^i and s_i for all i
10: $W^l = \{W_{r+1}^1, W_{r+1}^2, .., W_{r+1}^i\}$
11: $S = \{s^1, s^2, .., s^i\}$
12: **On server:** $W_{r+1}^g = globalStep(W^l, S)$
13: **On server:** $acc_r^g = acc(W_{r+1}^g, X_{test}, Y_{test})$
14: **end for**

The functions $localStep$ and $globalStep$ in lines 6 and 12 of algorithm 1 respectively, are either an implementation of the federated averaging Algorithm ([3]), or the SCAFFOLD algorithm ([13]).

The *localStep* for federated averaging is done by means of Stochastic Gradient Descent, i.e. :

$$W_{r+1}^i = W_r^i - \eta_l * \nabla L(X_{train}^i, y_{train}^i) \tag{3.1}$$

where W_r^i is the model on client i in round r, η_l the local learning rate, $L(X_{train}^i, y_{train}^i)$ is a loss function which differs between classifiers, and X_{train}^i and Y_{train}^i are the training data available at client i. Note that the clients also have to split data into training and testing to be able to independently evaluate a learned classifier. The global step of federated averaging consists of taking the weighted average for all model parameters:

$$W_{r+1}^g = \frac{1}{S} * \sum_{i=0}^{N} s^i * W_{r+1}^i \tag{3.2}$$

where s^i is the dataset size of the i^{th} client, N the amount of clients and $S = \sum_{i=1}^{N} s^i$.

Besides federated averaging, we have also used the SCAFFOLD aggregation algorithm [13]. For both the local and global step, SCAFFOLD expands on federated averaging by means of a so-called control variate c. Intuitively, these control variates are used to compensate for model drift, where a model update does not move into

the direction of the global optimum due to local datasets being distributed in a non-IID fashion. For the local step:

$$W^i_{r+1} = W^g_r - \eta_l \nabla L(X^i_{train}, y^i_{train}) + c^g_r - c^i_r \tag{3.3}$$

in which c^g_r and c^i_r are the global and local control variates in round r which are all initialized with zero values. Next (but within the same round), c^i gets updated as:

$$c^i_{r+1} = c^i_r - c^g_r + \frac{1}{\eta_l}(W^i_r - W^i_{r+1}) \tag{3.4}$$

These updated control variates are sent back to the server together with the model parameters and accuracy for that round. Then the global update step of SCAFFOLD first updates the coefficients:

$$W^g_{r+1} = W^g_r + \frac{\eta_g}{N} * \sum_{i=0}^{N}(W^i_{r+1} - W^i_r) \tag{3.5}$$

with η_g being the global learning rate. The global control variate c^g :

$$c^g_{r+1} = c^g_r + \frac{1}{N} * \sum_{i=0}^{N}(c^i_{r+1} - c^i_r) \tag{3.6}$$

The *localStep* of both fedAVG and SCAFFOLD allows for batch learning, i.e. splitting up the local training data into multiple shards, allowing multiple consecutive gradient descent steps each local epoch. However, our main objective was comparison with the central case and not optimizing performance. Therefore, since batch learning was not required for convergence, it was not utilized. Exception to this is the CNN classifier for the kinase datasets, which needed some batch learning for convergence: For all CNN experiments, the local data was split up into 10 batches each epoch.

The federated GBDT is implemented differently from the other classifiers, as a Decision Tree does not lend itself very well for parameter averaging. Instead, for GBDT, we followed the "inPrivate Learning" algorithm from [32], as shown in Algorithm 2. In this algorithm, only one client is active per round. This client is referred to as C_a in Algorithm 2. This client uses the decision trees of its predecessors to calculate a loss on its own dataset, which it then uses to boost the decision tree it builds.

3.4.2. Datasets

All classifiers were tested on five different datasets: i) MNIST [23], images of handwritten digits, ii) fashion MNIST [24], images of clothes, iii) a dataset used for the

Algorithm 2 The inPrivate learning algorithm

init W^g randomly
for all r rounds **do**
  At server : $C_a = r \bmod N$
  At server :send W^g to client C_a
  On client C_a **:** $\tilde{y}_{pred} = W^g(X^{C_a}_{train})$
  On client C_a **:** $L(C_a) = log(1 - \tilde{y}_{pred}) + y_{train} * log(\frac{\tilde{y}_{pred}}{1-\tilde{y}_{pred}})$
  On client C_a **:**create new tree minimizing $L(C_a)$
  On client C_a **:**add tree to W^g
  On client C_a **:**$acc_r = accuracy(W^g, X^{C_a}_{test}, Y^{C_{active}}_{test})$
  On client C_a **:** send W^g, acc_r back to server
  At server : Receive W^g, acc_r from client C_a
  At server :$acc^g_i = accuracy(W^g, X_{test}Y_{test})$
end for

prediction of Acute Myeloid Leukemia (AML) based on measured gene expressions [30], iv) a dataset consisting of molecular fingerprints used to predict the activation of certain kinases, and v) a dataset of 3D MRI scans of hearts from either healthy patients or patient with a cardiovascular disease.

Two and Four Class MNIST

The first dataset is derived from MNIST [23]. In order to reduce complexity, two derivatives were made, using only a subset of the original classes: a two-class problem (MNIST2) and a four-class problem (MNIST4). The digits were chosen based on which combination is most difficult to separate according to the original work introducing MNIST. For MNIST2, we included digits 4 and 9, whereas for MNIST4, digits 2 and 8 are being added. 80% of the dataset is used for training, with the remaining 20% for testing. Three different splits were made, resulting in 6 federated datasets: an IID distribution, a distribution with imbalances in sample size (SI) and one with imbalances in classes (CI). Figure 3.1d through 3.1f and supplementary figure 3.9 show these distributions.

Fashion MNIST

The fashion MNIST dataset [24] was also split into 80% training and 20% test data. Two different distributions were made: an IID distribution and a class-imbalanced distribution. Note that this imbalanced distribution is different from the MNIST distributions as it concerns ten classes, see figure 3.2a.

AML dataset (A1-A3)

The third dataset was taken from a study done by Warnat-Herresthal et al. [30]. In their experiments, they used measured transcriptomes of gene expressions from patients

to predict the presence of acute myeloid leukaemia (AML). These transcriptomes were taken from human peripheral blood mononuclear cells (PBMC) from three different sources. In the datasets, A1 and A2 are transcriptomes that have been measured using microarrays [33], whereas dataset A3 is measured using RNA-sequencing [34]. All three datasets contain 12709 different gene expression values per sample, with labels for 25 different illnesses. Samples were labelled according to [30], i.e. all AML samples were given label (1), and all other labels were joined under one label (0) (in the original work named 'cases' and 'controls'). This approach does result in an inherent class imbalance in the combined dataset, i.e. 7500 samples have label 0 and only 4000 samples have label 1. The datasets were split into 80% training samples and 20% test samples.

Kinase dataset.

The kinase dataset was taken from [27]. It consists of two deMorgan fingerprints, connectivity and feature based, with as labels the activation values for certain kinases. We chose two of these kinases, being KDR and ABL1. Two of the three datasets were quite sparse, so there is no full overlap between molecules with data on KDR and ABL1, practically creating two datasets. The activation values were binarized between active and non-active, with a cutoff value of 6.3 (e.g. a molecule is seen as activating a kinase if it has a pIC_{50} value higher than 6.3), following the original study.

MRI dataset

The last dataset consisted of 3D MRI scans of patients' hearts, taken from [28], where it was generated as part of a challenge on using deep learning models for cardiac MRI scans. Multiple time frames of MRI scans are available for all patients'. Following instructions from [28], only the time points of the end-diastolic and the end-systolic phase were used. As various MRI scanners were used (at different hospitals), not all scans had the same resolution/dimensionality. Scans were brought to equal dimensionality by downsampling to the lowest common dimensions.

We only selected samples from healthy patients and patients with hypertrophic cardiomyopathy (HCM), making it a binary classification problem. The original dataset had a train/validate/test split set up to make the problem extra challenging. Since our main objective is not to solve this issue, but rather compare federated and central classifiers, we opted to redistribute the test/validate/train split to be more similar across all clients, see figure 3.5b. Note that this did not mix the samples across clients.

3.4.3. Experimental setup

Our goal is to compare a federated classifier performance to a central model, i.e. when all data is available at the server. To do this in a fair way, the learning rate was kept

equal between the federated and central runs. Also, the amount of communication rounds was kept equal to the amount of epochs in the central case. Only one 'local' epoch was used during all experiments, meaning all classifiers sent their updated model back to the server after only one full pass of their local data. Four runs with different model initializations were made per experiment (for both the federated and centralized case). All experiments were executed on one laptop, running all clients as well as the server. The federated learning platform vantage6 [35], [36] was used. Vantage6 uses a dockerized solution, which means that every client (and every task that every node runs) runs in its own docker, therefore being unable to alter the solution of the other clients. The experiment on the MRI dataset, however, was not performed with vantage6 due to size constraints of both data and neural network. For this case, we built a simulated federated environment to run on our high performance cluster.

3.5. Appendices

3.5.1. Relation between clients and learning rate (for linear models)

The local model update from federated averaging (as used during this paper) is given by stochastic gradient descent:

$$W_t^i = W_{t-1}^g - \eta \nabla L(W_{t-1}^g(X^i), y^i) \tag{3.7}$$

These local model updates are then combined into a new global model as follows:

$$W_t^g = \frac{1}{S}\sum_{i=1}^{N} s^i W_t^i \tag{3.8}$$

Where N is the amount of clients, s^i the dataset size of client i, and $S = \sum_{i=1}^{N} s^i$. If we combine equations 3.7 and 3.8, we get:

$$W_t^g = \frac{1}{S}\sum_{i=1}^{N} \{s^i [W_{t-1}^g - \eta \nabla L(W_{t-1}^g(X^i), y^i)]\}. \tag{3.9}$$

By doing some refactoring, equation 3.9 becomes:

$$\begin{aligned} W_t^g &= \frac{1}{S}\sum_{i=1}^{N} s^i W_{t-1}^g - \frac{1}{S}\sum_{i=0}^{n} s^i \eta \nabla L[W_{t-1}^g(X^i, y^i)] \\ &= \frac{W_{t-1}^g}{S}\sum_{i=1}^{n} s^i - \frac{1}{S}\sum_{i=1}^{n} s^i \eta \nabla L[W_{t-1}^g(X^i), y^i]. \end{aligned} \tag{3.10}$$

Since $S = \sum_{i=1}^{n} s^i$, we get:

$$W_t^g = W_{t-1}^g - \frac{1}{S}\sum_{i=0}^{N} s^i \eta \nabla L[W_{t-1}^g(X^i), y^i]. \tag{3.11}$$

The loss function can be written as a sum of the loss over all samples in the local dataset, i.e. :

$$L(W_t^g(X^i), y^i) = \sum_{j=0}^{s^i} L(W_t^g(x_j^i), y_j^i), \tag{3.12}$$

with x_j^i the j^{th} sample in dataset X^i, and y_j^i its corresponding label. Combining equation 3.11 and 3.12, we get:

$$
\begin{aligned}
W_t^g &= W_{t-1}^g - \frac{1}{S}\sum_{i=0}^{N} s^i \eta \nabla \sum_{j=0}^{s^i} L(W_t^g(x_j^i) y_j^i) \\
&= W_{t-1}^g - \frac{1}{S}\sum_{i=0}^{N} s^i \eta \nabla \sum_{j=0}^{s^i} L_j^i
\end{aligned}
\tag{3.13}
$$

where L_j^i is used as shorthand for $L(W_t^g(x_j^i), y_j^i)$. From here we can work towards showing the relation between the amount of clients and learning rate. There are two separate assumptions that can be made, which lead to similar analyses with the same result. Both analyses will be given. The options are:

1. Assume that each client holds an approximately equal amount of samples, i.e. $s^1 = s^2 = \ldots = s^i$ for each $i \in N$
2. Assume that each client gives an equivalent model update each round, i.e. $\nabla L_r^1 = \nabla L_r^2 = \ldots = \nabla L_r^i$ for each $i \in N$

Note that the second assumption is easiest to satisfy if the samples are IID distributed. However, although unlikely, a scenario is possible in which one clients holds only a few samples which result in a large loss each, whereas most clients hold many more samples with only a small loss per sample. Therefore, it is not strictly necessary to assume IID distribution with assumption 2.

Using assumption 1

With assumption 1, we can rewrite s^i as:

$$
s^1 = s^2 = \ldots = s^i = \frac{S}{N}. \tag{3.14}
$$

Now, using equation 3.14 with equation 3.13, we get:

$$
\begin{aligned}
W_t^g &= W_{t-1}^g - \frac{1}{S}\sum_{i=0}^{N} \frac{S}{N} \eta \nabla \sum_{j=0}^{s^i} L(W_t^g(x_j^i) y_j^i) \\
&= W_{t-1}^g - \frac{\eta}{N} \nabla \sum_{i=0}^{N} \sum_{j=0}^{s^i} L(W_t^g(x_j^i), y_j^i).
\end{aligned}
\tag{3.15}
$$

In order to compare with a centralized case, lets assume a setting in which one single client holds all S datapoints. A similar (no batch learning) SGD step of such a client can be described as:

$$
W_t^c = W_{t-1} - \eta \nabla L(W_{t-1}(X), y), \tag{3.16}
$$

where W_t^c is the central model at epoch t, and

$$X = \sum_{i=1}^{N} X^i, \tag{3.17}$$

is the concatenation of all local datasets X^i. Combining 3.16 and 3.17, we get:

$$W_t^c = W_{t-1}^c - \eta \nabla \sum_{i=1}^{N} L(W_{t-1}^c(X^i), y). \tag{3.18}$$

Now using equation 3.12, we can get:

$$W_t^c = W_{t-1}^c - \eta \nabla \sum_{i=1}^{N} \sum_{j=1}^{s^i} L(W_{t-1}^c(x_j^i), y). \tag{3.19}$$

Comparing equations 3.15 and 3.19, it becomes clear that they are equivalent, except for the factor $\frac{1}{N}$ in equation 3.15, which effectively lowers the learning rate of the federated case by a factor N as compared to its central counterpart.

Using assumption 2

For this analysis, it is convenient to use equation 3.11 as a starting point instead of equation 3.13:

$$W_t^g = W_{t-1}^g - \frac{1}{S} \sum_{i=0}^{N} s^i \eta \nabla L^i, \tag{3.20}$$

where we use L^i as shorthand for $L(W_{t-1}^g(X^i), y^i)$. Since we assume that $\nabla L^1 = \nabla L^2 = \nabla L^i$ for all $i \in N$, we can remove the dependency of i, i.e.

$$W_t^g = W_{t-1}^g - \frac{1}{S} \sum_{i=0}^{N} s^i \eta \nabla L^i = W_{t-1}^g - \frac{1}{S} \sum_{i=0}^{N} s^i \eta \nabla L. \tag{3.21}$$

Since $S = \sum_{i=0}^{N} s^i$, we now get:

$$W_t^g = W_{t-1}^g - \frac{1}{\sum_{i=0}^{N} s^i} \sum_{i=0}^{N} s^i \eta \nabla L = W_{t-1}^g - \eta \nabla L. \tag{3.22}$$

Comparing to the central case, we start off with equation 3.19:

$$W_t^c = W_{t-1}^c - \eta \nabla \sum_{i=1}^{N} \sum_{j=1}^{s^i} L(W_{t-1}^c(x_j^i), y) = W_{t-1}^c - \eta \nabla \sum_{i=1}^{N} L^i, \tag{3.23}$$

where we once again use $L^i = L(W^g_{t-1}(X^i), y^i)$. Once again, using $L^i = L$, we arrive at:

$$W^c_t = W^c_{t-1} - \eta\nabla\sum_{i=1}^{N} L^i = W^c_{t-1} - \eta\nabla\sum_{i=1}^{N} L = W^c_{t-1} - N\eta\nabla L. \tag{3.24}$$

3.5.2. SCAFFOLD

The SCAFFOLD algorithm was created as a means to increase the performance of the earlier discussed federated averaging, specifically with respect to a non-IID setting. A non-IID setting means that the distribution of classes between different clients is not similar, i.e. some clients only having access to a very low amount (in some cases even zero) of samples of certain classes. The problem that arises in federated averaging within the non-IID setting can be intuitively described as a drift, meaning that the global parameters do not or slowly converge to the optimum. In order to combat this drift, [13] introduce a new algorithm for Stochastic Controlled Averaging, called SCAFFOLD.

The main addition in SCAFFOLD is the introduction of a control variate for all clients, *and* for the server. This control variate consists of a set of values with the same structure as the set of parameters (i.e. it has one value per parameter). Intuitively, it denotes the direction (and magnitude) of the local update of said parameter. If this direction is different from many other clients, the control variate is used to compensate for that difference, which decreases the aforementioned drift.

The general setup of SCAFFOLD is similar to that of federated averaging. The algorithm consists of rounds, which consists of a global part at the server, and a local part which happens at all clients simultaneously. at the beginning of each round, the server sends the global model to all clients, as well as its own control variate c. Now, each client makes a pass over its local data to calculate the gradient over its loss function. This also happens in federated averaging, as part of the stochastic gradient descent step. After the gradient has been determined, the parameters get updated as follows:

$$W^i_{r+1} = W^g_r - \eta\nabla L(X^i_{train}, y^i_{train}) + c^g_r - c^i_r \tag{3.25}$$

c^g_0 and c^i_r have been initialized as all zero, as described in [13]. Note that if both c^g and c^i are all zero, equation 3.25 becomes standard SGD, and SCAFFOLD becomes federated averaging for the local step. After the model has been updated, each client also needs to update its control variate. This is done according to the following equation:

$$c^i_{r+1} = c^i_r - c^g_r + \frac{1}{\eta_l} W^i_r - W^i_{r+1} \tag{3.26}$$

Finally, W^i_{r+1} is sent back to the server. Once the server has received W^i_{r+1} for all $i \in N$, it aggregates the local updates into a new iteration of the global model. This update is executed as follows:

$$W^g_{r+1} = W^g_r + \frac{\eta_g}{N} * \sum_{i=1}^{N} W^i_{r+1} - W^i_r \tag{3.27}$$

Which is equivalent to the global update of federated averaging, if η_g is set to 1 (which is the case during all experiments described in this report). Different from federated averaging is the need to update c^g as well, which is done according to the following equation:

$$c^g_{r+1} = c^g_r + \frac{1}{N} * \sum_{i=1}^{N} (c^i_{r+1} - c^i_r) \tag{3.28}$$

Once this is done, the next round starts by the server sending out the updated model parameters and c. A big difference wich federated averaging is that SCAFFOLD is a *stateful* algorithm, with the control variates functioning as some sort of state for clients and server.

3.5.3. Federated PCA

After some initial testing, it turned out that the high dimensionality of the AML datasets was problematic for model convergence. To reduce the dimensionality, a principal component analysis (PCA) ([37]) was applied, and the first 100 principal components were chosen to represent the samples in a federated way, described in supplementary Algorithm 3. Our implementation is a specific case of the incremental federated PCA as proposed in[26] (see pseudocode 3), with the increment size being the dataset size (resulting in an approach that is not incremental), as this was fast enough for our experiments. Two versions of normalizing were explored, using a global or a local mean and variance, see pseudocode 4. The local normalization was deemed to perform a lot better, as it also serves as a z1-normalization.

Algorithm 3 Federated PCA

input: glob (bool), n_{pca} (int)
On all clients i**:** X^i_{norm} = normalize(X^i, glob) ▷ *normalize either using global or local mean, see supplementary algorithm 4*
On all clients i**:** $U^i, \sum^i, V^i = \text{SVD}(X^i_{norm})$
On all clients i**:** send U^i, $\sum^i$ to server
On server: $US^g = \text{concat}(U^1 * \sum^1, ..., U^N * \sum^N)$
On server: $U^g, \sum^g, V^g = \text{SVD}(US^g)$
On server: send $U^g_{0:n_{pca}}$ to all clients
On all clients i**:** $X^i_{pca} = X^i * U^g_{0:n_{pca}}$

Algorithm 4 Normalization options for fedPCA

input: glob(bool)
if glob == True **then**
    On Server: request metadata from all clients i
    On all Clients: $mean^i = mean(X^i)$, $s^i = length(X^i)$ ▷ *Send back local mean and dataset size*
    On server: Collect $mean^i$, std^i and s^i for all i
    On server: $S = \sum_{i=1}^{n} s^i$ ▷ *Calculate total sample size*
    On server: $mean^g = \frac{1}{S} \sum_{i=1}^{n} s^i * mean^i$ ▷ *Calculate global mean*
    On server: send $mean^g$ to all clients,
    On all Clients: $\sigma^i = \frac{1}{s^i} \sum_{j=0}^{s^i} (X^i_j - mean^g)^2$ ▷ *calculate partial variance on each client*
    On all Clients: send σ^i to server
    On server: $\sigma^g = \frac{1}{S} \sum_{i=0} Ns^i * \sigma^i$ ▷ *calculate global variance*
    On server: send σ^g to all clients
    On all Clients i**:** $X^i_{norm} = \frac{X^i - mean^g}{\sqrt{\sigma^g}}$
else
    On all clients i **do:** $mean^i = mean(X^i)$, $\sigma^i = var(X^i)$ ▷ *calculate local mean and variance*
    On all clients i **do:** $X^i_{norm} = \frac{X^i - mean^i}{\sqrt{\sigma^i}}$
end if

3.5.4. Supplementary Tables

Table 3.1.: the architectures of the FNN

	Layer 1	Layer 2	Layer 3
MNIST2	Fully Connected: 784 x 100	ReLU	Fully Connected: 100 x 2
MNIST4	Fully Connected: 784 x 100	ReLU	Fully Connected: 100 x 4
fashion MNIST	Fully Connected: 784 x 100	ReLU	Fully Connected: 100x10
AML	Fully Connected: 100 x 100	ReLU	Fully Connected: 100x2

Table 3.2.: the CNN architectures

	Layer 1	Layer 2	Layer 3
MNIST2	Convolutional: kernel size:3 stride: 1 padding: 1	Max pool: kernel = 2 stride = 2	Fully Connected: 196 x 2
MNIST4	Convolutional: kernel size:3 stride: 1 padding: 1	Max pool: kernel = 2 stride = 2	Fully Connected: 196 x 4
fashion MNIST	Convolutional: kernel size:3 stride: 1 padding: 1	Max pool: kernel = 2 stride = 2	Fully Connected: 196 x 10
AML	Convolutional: kernel size:3 stride: 1 padding: 1	Max pool: kernel = 2 stride = 2	Fully Connected: 25 x 2

Table 3.3.: Results on the two kinase datasets

		KDR		ABL1	
		Final accuracy (std)	Mean AUC (std)	Final accuracy (std)	Mean AUC (std)
LR	Central	0.75 (0.007)	0.69 (0.006)	0.78 (0.01)	0.75 (0.012)
	Federated	0.74 (0.004)	0.68 (0.005)	0.78 (0.01)	0.74 (0.018)
SVM	Central	0.76 (0.007)	0.7 (0.004)	0.78 (0.015)	0.75 (0.017)
	Federated	0.74 (0.005)	0.69 (0.007)	0.77 (0.01)	0.74 (0.017)
FNN	Central	0.75 (0.002)	0.68 (0.002)	0.73 (0.004)	0.71 (0.0)
	Federated	0.75 (0.005)	0.68 (0.004)	0.72 (0.003)	0.71 (0.0)
CNN	Central	0.72 (0.061)	0.65 (0.053	0.76 (0.014)	0.74 (0.015)
	Federated	0.78 (0.009)	0.73 (0.012)	0.77 (0.005)	0.76 (0.001)
GBDT	Central	0.82 (0.0)	0.79 (0.0)	0.81 (0.001)	0.79 (0.001)
	Federated	0.73 (0.009)	0.69 (0.003)	0.78 (0.011)	0.75 (0.005)

Table 3.4.: performances on the MNIST datasets

			MNIST2		MNIST4	
			Final accuracy (std)	Mean AUC (std)	Final accuracy (std)	Mean AUC (std)
LR	Central		0.97 (0.001)	0.97 (0.001)	0.93 (0.001)	0.91 (0.001)
	Federated	IID	0.97 (0.002)	0.97 (0.002)	0.93 (0.001)	0.91 (0.001)
		CI	0.97 (0.001)	0.96 (0.002)	0.93 (0.001)	0.91 (0.001)
		SI	0.97 (0.002)	0.97 (0.002)	0.93 (0.002)	0.91 (0.002)
SVM	Central		0.96 (0.003)	0.92 (0.005)	0.89 (0.006)	0.78 (0.008)
	Federated	IID	0.96 (0.003)	0.92 (0.005)	0.89 (0.006)	0.78 (0.008)
		CI	0.96 (0.003)	0.91 (0.005)	0.89 (0.005)	0.78 (0.008)
		SI	0.96 (0.003)	0.92 (0.005)	0.89 (0.005)	0.78 (0.008)
FNN	Central		0.95 (0.000)	0.90 (0.004)	0.90 (0.013)	0.82 (0.006)
	Federated	IID	0.95 (0.005)	0.90 (0.005)	0.91 (0.008)	0.84 (0.009)
		CI	0.95 (0.001)	0.90 (0.005)	0.91 (0.008)	0.84 (0.009)
		SI	0.93 (0.010)	0.90 (0.005)	0.91 (0.008)	0.84 (0.009)
CNN	Central		0.94 (0.027)	0.82 (0.041)	0.90 (0.013)	0.75 (0.05)
	Federated	IID	0.93 (0.011)	0.83 (0.030)	0.92 (0.008)	0.83 (0.03)
		CI	0.93 (0.013)	0.83 (0.002)	0.92 (0.008)	0.83 (0.03)
		SI	0.93 (0.010)	0.83 (0.030)	0.92 (0.008)	0.83 (0.03)
GBDT	Central		0.98 (0.000)	0.97 (0.000)	0.96 (0.0)	0.94 (0.000)
	Federated	IID	0.97 (0.002)	0.96 (0.001)	0.95 (0.002)	0.93 (0.001)
		CI	0.97 (0.002)	0.94 (0.001)	0.95 (0.001)	0.92 (0.001)
		SI	0.97 (0.002)	0.96 (0.001)	0.95 (0.001)	0.93 (0.001)

Table 3.5.: performances on the fashion MNIST and A2 of the AML dataset

			fashion MNIST		A2	
			Final accuracy (std)	Mean AUC (std)	Final accuracy (std)	Mean AUC (std)
LR	Central		0.74 (0.003)	0.66 (0.005)	0.98 (0.002)	0.98 (0.002)
	Federated	IID	0.78 (0.003)	0.66 (0.005)	0.98 (0.003)	0.98 (0.001)
		CI	0.74 (0.005)	0.65 (0.005)	0.98 (0.002)	0.98 (0.001)
		SI	-	-	0.98 (0.003)	0.98 (0.001)
SVM	Central		0.75 (0.003)	0.67 (0.005)	0.96 (0.006)	0.91 (0.001)
	Federated	IID	0.75 (0.003)	0.67 (0.005)	0.95 (0.003)	0.91 (0.001)
		CI	0.75 (0.004)	0.66 (0.005)	0.95 (0.005)	0.91 (0.003)
		SI	-	-	0.95 (0.004)	0.92 (0.002)
FNN	Central		0.67 (0.005)	0.6 (0.008)	0.98 (0.000)	0.97 (0.002)
	Federated	IID	0.68 (0.006)	0.61 (0.004)	0.98 (0.002)	0.97 (0.001
		CI	0.68 (0.006)	0.61 (0.004)	0.98 (0.004)	0.97 (0.003)
		SI	-	-	0.97 (0.002)	0.98 (0.002)
CNN	Central		0.73 (0.023)	0.54 (0.075)	0.87 (0.046)	0.65 (0.047)
	Federated	IID	0.73 (0.021)	0.56 (0.038)	0.90 (0.005)	0.65 (0.031)
		CI	0.73 (0.021)	0.56 (0.038)	0.90 (0.005)	0.65 (0.031)
		SI	-	-	0.90 (0.005)	0.65 (0.031)
GBDT	Central		0.87 (0.0)	0.84 (0.0)	0.98 (0.000)	0.95 (0.000)
	Federated	IID	0.84 (0.001)	0.83 (0.0)	0.96 (0.001)	0.94 (0.001)
		CI	0.23 (0.192)	0.28 (0.184)	0.95 (0.002)	0.92 (0.002)
		SI	-	-	0.97 (0.003)	0.94 (0.001)

3.5.5. Supplementary figures

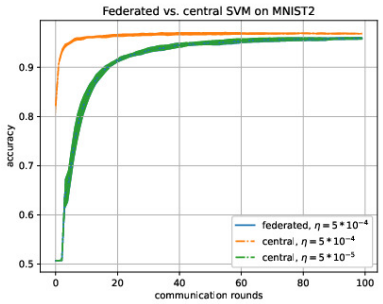


Figure 3.6.: the effect of varying learning rates for the comparison between federated and centralized SVM models.

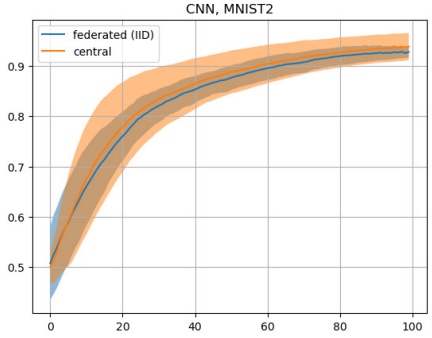


Figure 3.7.: Accuracy curves for the centralized and federated CNN on MNIST2 (IID for federated)

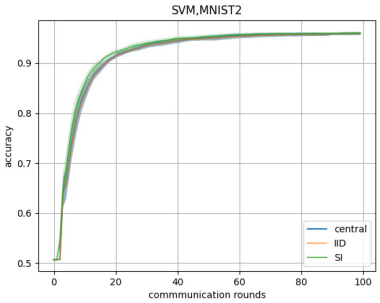
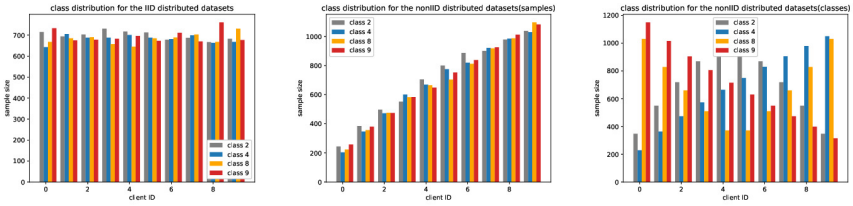


Figure 3.8.: Accuracy curves for the federated IID and SI distributions, compared with the central classifier.



(a) (b) (c)

Figure 3.9.: Data distributions on MNIST4

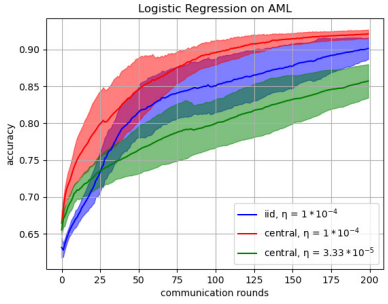


Figure 3.10.: LR on the original AML dataset.

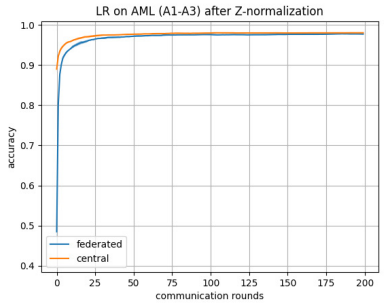


Figure 3.11.: LR on the AML dataset after Z-normalization

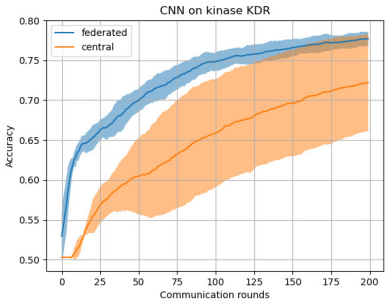


Figure 3.12.: CNN on kinase KDR

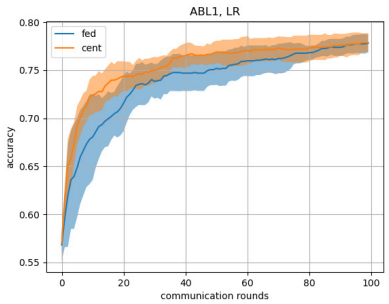


Figure 3.13.: LR on kinase ABL1

4

Federated K-Means clustering

Federated learning is a technique that enables the use of distributed datasets for machine learning purposes without requiring data to be pooled, thereby better preserving privacy and ownership of the data. While supervised FL research has grown substantially over the last years, unsupervised FL methods remain scarce. This work introduces an algorithm which implements K-means clustering in a federated manner, addressing the challenges of varying number of clusters between centers, as well as convergence on less separable datasets.

S. Garst and M. Reinders. 'Federated k-means clustering'. In: International Conference on Pattern Recognition*. Springer. 2024, pp. 107–122*

4.1. Introduction

Nowadays, lots of data is being generated in a distributed fashion. Mobile phones and other personal devices such as smart watches enable the collection of massive amounts of data. If made accessible, this data could prove useful for improving the performance of the services provided by these devices. However, due to a growing concern on data privacy, more and more users of these devices are hesitant in sharing their data. Furthermore, regulations such as the General Data Protection and Regulation (GDPR) act prevent the collection of data of this kind in bulk. Federated learning (FL) [1] was introduced as a solution to this problem. In short, instead of pooling data to train a single model, instances of a model are being shared to data owners (clients), which then train the model on their local data. Then, these trained models are sent back to the central server, which aggregates the results. Next, a new round begins with the server sending out the updated models. This cycle continues until convergence.

Over the past couple of years, research has shown FL to be a promising technique, reaching performances comparable to a central approach in which all data of the clients is pooled at a single location[2, 3]. The vast majority of the federated learning research has been focusing on the supervised learning paradigm. Little work has been done on unsupervised federated learning methods, even less so when specifically looking into clustering techniques [4, 5]. One of these clustering techniques is k-means clustering [6]. In a federated learning setting, k-means clustering can be described as trying to find overarching cluster means over data which is distributed among different datasets (clients).

Prior work has been done on creating federated k-means clustering algorithms [7–11]. We focus on solving the issue of a variable amount of local clusters. When data distributions between clients differ, it is likely that not all clients share data from all global clusters. When this is the case, the amount of clusters present per client can differ. This complicates federated clustering in two ways. First, matching cluster means between clients becomes less straightforward, as there is no one-to-one matching anymore. Second, when the local k does not correspond to the global k, the problem of finding an optimal k scales linearly with the amount of clients. Furthermore, when clients hold only part of the data, it can be hard to distinguish between outliers and samples of a different cluster, without the knowledge of the data distribution of other clients. Altogether, this makes manually determining a value for k on each client locally complicated, if not infeasible without loss of performance.

We propose an iterative federated k-means clustering algorithm (FKM) that automatically determines the local value for k. By iteratively aggregating local cluster means, and running k-means locally on all clients in parallel, we are able to create a clustering that in many cases corresponds to a central clustering, i.e. the k-means clustering that would occur if all data was pooled together. By pruning empty global clusters on local clients, we are able to deal with a variable amount of local clusters between clients, without having to set the values for k locally on each client.

4.2. Related work

Since its inception in 2017, federated learning has been applied in various scenarios. The most well known taxonomy of federated learning systems is the split between the cross-device and cross-silo settings [5]. In brief, in the cross-device setting, many , in the order of thousands or more, devices are connected. However their connection is usually unstable, and cross-device federations will have to deal with users dropping out or joining later throughout the process. This setting is mostly applicable to federations with end-user devices, such as mobile phones. On the other hand, in the cross-silo setting, there are usually only a couple to tens of clients. These clients are usually quite stable and can be assumed to be connected from the beginning of the process all the way to the end. This setting is more applicable to health care centers or companies learning a shared model without sharing their data due to privacy or competition reasons.

Different challenges emerge based on whether the federation is cross-device or cross-silo and as such different algorithms have emerged for either setting. For the cross-device setting, Liu et. al. introduced a method for federated k-means in a setting where each client is seen as a single sample [9]. More specifically, they tackle the issue of proactive caching in cellular networks, i.e. trying to predict what data to keep local based on popularity. Their method is robust against user change, and assumes a tiered architecture with base stations and sub stations, as can be found in next-generation cellular networks. Kumar et. al. also propose a federated k-means clustering algorithm that can be applied for the cross-device setting [7]. Their method assumes a central dataset available at the server, on which a k-means clustering is pre-trained. It is then distributed across all clients who update the clustering based on their local data, after which the clustering is aggregated again on the server. Finally, Hou et. al. created a k-means clustering algorithm based on homomorphic encryption and blockchain [8]. Although their method is not explicitly applied to the cross-device setting, it does share encrypted versions of the data, which is often still unacceptable for many cross-silo use-cases.

For the cross-silo setting, Servetnyk et. al. proposed a federated k-means clustering algorithm based on dual averaging and self-organising maps [10]. Although their algorithm is capable of dealing with heterogeneous data, they do not explicitly address the challenge of a variable local k, or cluster alignment. Finally, Dennis et. al. propose a one-shot federated k-means clustering algorithm, which only needs one local clustering, as well as one global aggregation step [11]. As we focus mostly on the cross-silo setting, our algorithms bears the most resemblance to the method by Dennis et. al. (kFed). The key differences are that our method does not require setting a local value for k on each client, which, as we argue in the introduction, can be difficult if not impossible without loss of performance. This also allows us to have a different k for each client, further improving performance. Finally, we show that, for less separable datasets, there is a substantial performance gain in iterating between the local data and global aggregation more than once, as is done by Dennis et. al.

4.3. Methods

Notation used throughout this section is found in table 4.1. The pseudocode for our proposed federated k-means algorithm (FKM) can be found in algorithm 5. The algorithm can be divided into two parts: an initialization step, in which we generate initial cluster means on each client using k-means++ initialization [12], and an iterative k-means step in which clients communicate their cluster means to the server, which aggregates these means into a 'global' set of means, which then gets redistributed to the clients for the next k-means iteration. See supplement 4.6.1 for background on k-means and k-means++.

Table 4.1.: notations used

symbol	description
X_i	data on client i
K_g	global number of clusters
K_i	number of clusters on client i
C_g	global cluster means
C_i	cluster means of client i
M	total amount (sum) of local clusters
S_i	amount of samples for each cluster on client i
N	total amount of clients

4.3.1. Determining the amount of local clusters

While the global amount of clusters is set (main parameter k of the k-means procedure), it is not a given that each client has data for each of these clusters. In other words, the number of clusters between clients can differ, and is not necessarily equal to the number of clusters in the pooled data. In order to solve this problem, we determine which global clusters correspond to local data in each round on each client. Before a client applies a new k-means step locally, it assigns its data to the global cluster means it has received (line 14). Next, clients check if there are empty clusters, i.e. cluster means which did not get any points assigned to them. If so, clients discard these empty clusters (line 15). The remaining (global) cluster means are then used as initialization for the next local k-means step (line 16). This way, k can locally become smaller when running k-means on the clients. Since this pruning step only happens after global aggregation, we guarantee that the discarded clusters are indeed corresponding to clusters on other clients.

4.3.2. Cluster alignment

After each client has calculated one iteration of k-means (not until convergence, to avoid local minima) on their local data (each with their own amount of local clusters), they send their cluster means as well as the amount of samples per cluster back to the server.

Algorithm 5 The federated kmeans algorithm
Input: K_g

Init:
on each client $i \in N$ **do:**
$K_i = K_g$
S_i, C_i = kmeans++_init(X_i, K_i) ▷ get cluster means using kmeans++ initialization
send S_i, C_i to server
For each round r **do:**
On server do:
$C_l = [C_1|C_2|..|C_M]$ ▷ Concatenate all local cluster means
$S_l = [S_1|S_2|..|S_M]$ ▷ Repeat for sample amounts per cluster
C_g = kmeans(C_l, K_g, weights = S_l) ▷ Obtain new global clusters using kmeans
send C_g to all clients
On each client $i \in N$ **do:**
$C_i = C_g$
S_i = kmeans_assign(X_i,C_i) ▷ Determine empty clusters
$C_i = C_i$[s != 0 for s in S_i] ▷ Drop empty global clusters
K_i = size(C_i)
S_i, C_i = kmeans(X_i, K_i, init = C_i) ▷ run kmeans from remaining global cluster means
send S_i, C_i to server

The server then concatenates all cluster means, and aggregates them. It does so by running a k-means clustering on the received local means until convergence (using the global k parameter), to align clusters from different clients to each other. This global k-means is weighted by the amount of samples per cluster found, such that a cluster with lots of samples in it will have a bigger impact on the aggregation step compared to a cluster with fewer samples. That is, we modify the k-means objective function (see supplement 4.6.1 for the original) into:

$$F_{km} = \sum_{j=0}^{M} \min_{C_i \in C_g} (S_j ||C_j - C_i||^2) \tag{4.1}$$

where S_j is the amount of samples corresponding to local cluster C_j. Note that the *cluster means* sent back by the clients are at the server used as the *samples* for clustering using k-means. Doing the aggregation with a k-means clustering, we solve the cluster alignment problem, since similar clusters will be close to each other and thus merged by the global k-means step.

Because the amount of samples have to be reported to the central server, there exists a privacy risk if a client finds a cluster with only one sample in it. To prevent this, any clusters holding less than p samples (we used $p = 2$ throughout this work) are simply omitted from the list of means sent to the server.

4.4. Results

We compared our federated k-means (FKM) with a k-means clustering that is executed on all data centrally, as well as to one-shot the method of Dennis et. al [11]. Our first set of experiments is on simulated data, such that ground truth labels of the cluster centers are known. We therefore calculate the Adjusted Rand Index (ARI) for both central and federated approaches with respect to the labelled samples data. Since there are no labels for the clustering in the FEMNIST experiment, the silhouette score was used instead. In some cases, we added an "informed" setting for Dennis et. al., in which we set K_l such as to achieve the highest ARI score by exhaustive search. In all other cases, we run their method using $K_l = K_g$, as the ARI score is only available when ground truth labels are known, which is not always the case.

4.4.1. Clients holding different parts of the data

In order to validate the FKM algorithm, a synthetic two-dimensional dataset was generated. The generation procedure is taken from Servetnyk et. al. [10]. Sixteen cluster centers were chosen with an equal distance (here 5) from one another, see fig. 4.1a. Then, 50 data points were sampled around each cluster center using a normal distribution (with variance 1). This data was then distributed among four clients in the following way: First, each client is assigned a 'location' within the field ($X_1, X_2 \in (-12.5, 12.5)$). From there, the probability P that a data point would be assigned to a certain client scales inversely with the euclidean distance d to that datapoint:

$$P = 1 - exp(-\frac{\beta}{d}) \tag{4.2}$$

where β is a parameter which can be tuned to promote more or less heterogeneity in the data separation. Differing from [10], if a data point happens to be assigned to multiple clients, it instead gets assigned at random.

We wanted to explore the influence of data heterogeneity, i.e. a varying amount of clusters per client. To do so, we generated three versions of this dataset, with $\beta = 0.1, 1, 10$. See figure 4.1 c-e for the final distributions. Note that β only changes which points get assigned to which client, meaning that it does not influence the performance for the central case. Figure 4.1b shows that our method is able to attain performance similar to a centralized k-means clustering, while outperforming Dennis et. al., regardless of tuning of the K_l parameter. Performance of our FKM approach seems to be independent of β (in contrast to the method of Dennis et. al.), meaning that our algorithm is robust to having varying cluster amounts between clients.

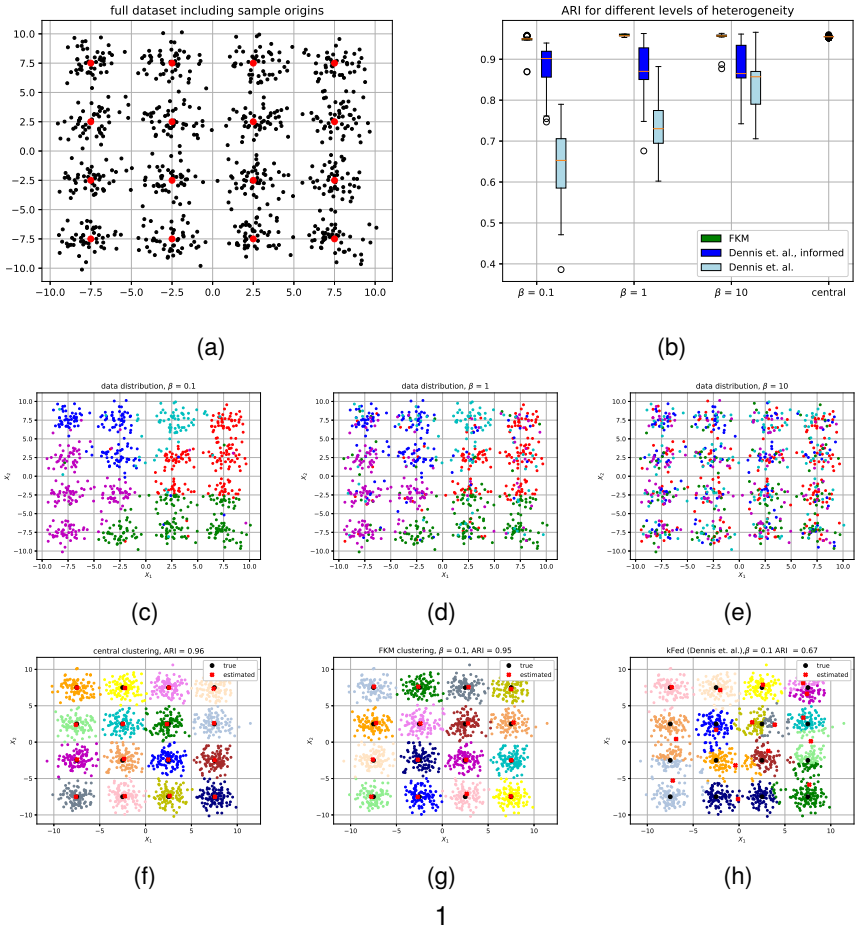


Figure 4.1.: The regular synthetic datasets. (a) shows the original sampling of the regular synthetic dataset, with the defined cluster means (from which the data are generated using a normal distribution N(0,1)) in red. (b) shows ARI results on all three datasets. (c) until (e) shows how this dataset is distributed over five different clients using different values of β. Different colors indicate the different clients. (f) to (h) show examples of a clustering on (c) as given by a centralized k-means, FKM (ours), and kFed (Dennis et. al.), respectively. Note that different colors in the last three plots denote different cluster assignments instead of different clients.

4.4.2. Increasing levels of noise

Next, we explored the effect of having noisier clusters. We recreated the regular synthetic dataset, but varied the standard deviation from which samples are being generated, from 1 to 1.5 (original used 1). Figure 4.2 shows the effect. We generated these datasets twice, once with 50 points per cluster and once with 200 points per cluster.

Results on these datasets are shown in figure 4.3. For both central and federated clustering, the ARI scores go down for higher noise levels. This is expected, as there will be more points ending up closer to the cluster they did not originally belong to, meaning that even if kmeans finds the original cluster means perfectly, the label assignment will be off. Therefore, the relative difference between federated and central clustering is more important than the absolute ARI scores. Our method attains a similar average performance; however, variance seems to increase compared to centralized clustering. Furthermore, for $\beta = 0.1$, mean ARI decreases compared to central clustering at high noise levels, meaning that a setting with high noise as well as high cluster variability is still a hard challenge for our federated k-means algorithm.

Regardless, performance does seem to increase significantly as compared to the method of Dennis et. al. [11] . This can partly be due to our ability to iterate. Figure 4.3d to 4.3f shows that, especially for noisier datasets, there is a large benefit in being able to iterate more often. The amount of points per cluster does not seem to influence ARI score significantly, see supplement 4.6.2.

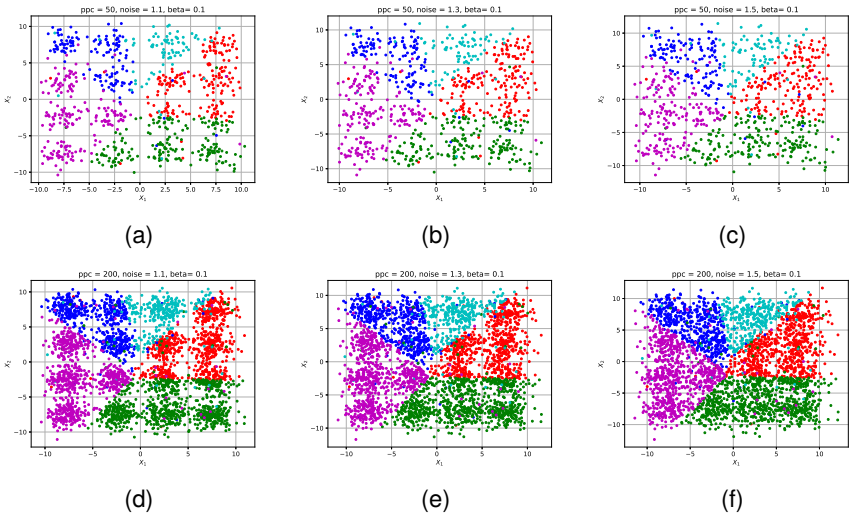


Figure 4.2.: Some of the data distributions of the simulated datasets with increasing levels of noise (columns), using 50 or 200 points per cluster (rows).

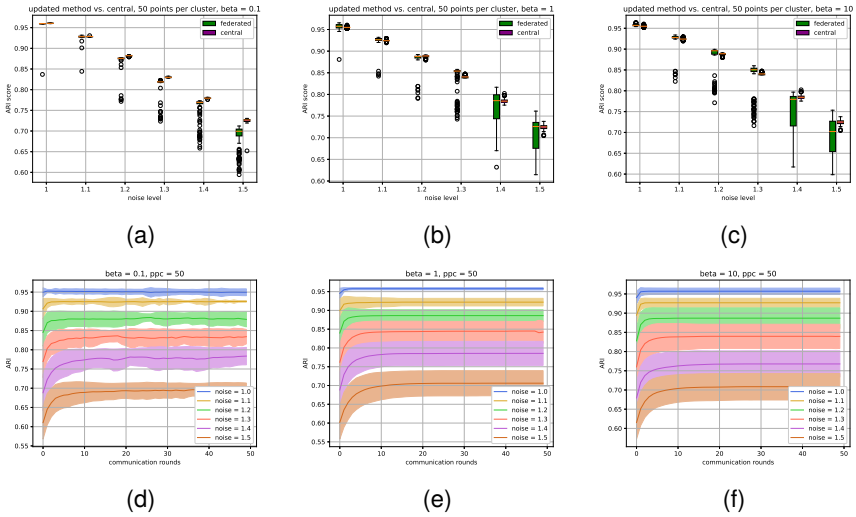


Figure 4.3.: Clustering results on the synthetic dataset when using different levels of noise for different values of β. (a) to (c) show the final ARI scores for beta = 0.1, 1 and 10, respectively. (d) to (f) show how the ARI score for FKM converges over time, each corresponding to the figure above it.

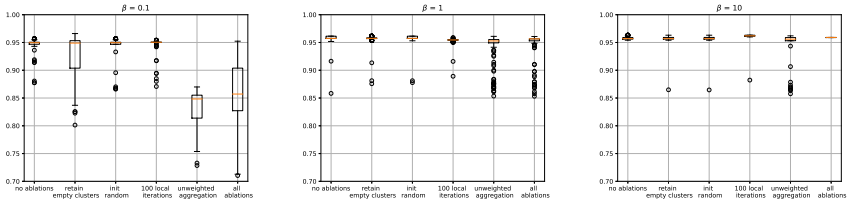
4.4.3. Ablation study

To explore the importance of several parts of our algorithm, we perform an ablation study on the two dimensional synthetic data introduced in section 4.4.1 with a noise parameter of 1. We make five separate ablations, as well as one setting in which all five modifications are included:

- **Retain empty clusters:** For this ablation, we skip the step where we prune the empty clusters, effectively fixing K_l equal to K_g in all clients.
- **Initialization:** Instead of initializing local cluster using k-means++, we initialize local clusters at random.
- **100 local iterations:** Instead of performing one iteration of k-means locally, we perform 100 iterations.
- **Unweighted aggregation:** When aggregating cluster means on the server, instead of weighting the cluster means by the amount of samples corresponding to said cluster locally, we simply give every cluster a weight of 1 (Note that we explored several values for the amount of local iterations, however we saw little difference between those values, so for sake of brevity we report only the results using the largest value that we explored (100)).

We considered the same values of β as earlier described in section 4.4.1, i.e. $\beta \in \{0.1, 1, 10\}$, as shown in figure 4.4. This figure shows that the degree to which different parameters are important depend on β, i.e. the heterogeneity of the data. For $\beta = 0.1$ the largest impact can be seen for doing weighted aggregation and local pruning.

When data is homogeneously distributed (β = 10), i.e. every client has data for each cluster, the proposed ablations seem to have little impact on the algorithm. In fact, increasing the amount of local iterations could even be beneficial in a completely heterogeneous case. This is in line with literature on supervised federated learning, where increasing the amount of local epochs can increase performance [1]. However, under heterogeneous circumstances, in supervised federated learning, clients might move too far into local optima before aggregation, decreasing performance with more local iterations [13]. We therefore hypothesize that in even more heterogeneous circumstances, a lower amount of local iterations could still be beneficial for our method as well.



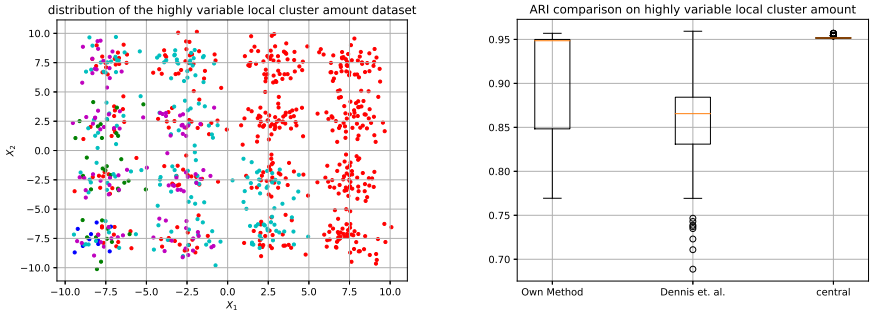
(a) (b) (c)

Figure 4.4.: Results of the ablation study.(a), (b) and (c) show results for β = 0.1, 1 and 10, respectively. (a) shows that, for a highly heterogeneous dataset, both dropping empty local clusters and especially weighted aggregation have a large impact on model performance. However, as the data becomes more homogeneously distributed, (subfigures (b)-(c)), these factors become less important.

4.4.4. High variability in number of local clusters

Next we wanted to explore the effect of having an even more variable local k. We used the same data as generated for the regular synthetic dataset, but distributed even more heterogeneously, such that each client only had data from 1, 4, 7, 10 or 16 clusters, respectively. See fig. 4.5a.

Figure 4.5b shows that our method attains a similar average performance as compared to the central case, however with a larger variation. This is probably caused by differences in initializations. If the algorithm initializes in such a way that clients assign data to more clusters than what is being present in their data, the algorithm has a hard time correcting for that. Furthermore, it does not help that one client only has ten datapoints in total, meaning it initializes ten clusters of size one, of which none are being send to the central server due to privacy issues. Regardless, our method does outperform the algorithm from Dennis et. al. [11]. This is likely due to our algorithm's ability to change the value of k for its local k-means step between clients.



(a) (b)

Figure 4.5.: Assessment of the method on data with a large variability of local clusters per client. (a) shows the distribution per client, (b) the ARI results for different methods.

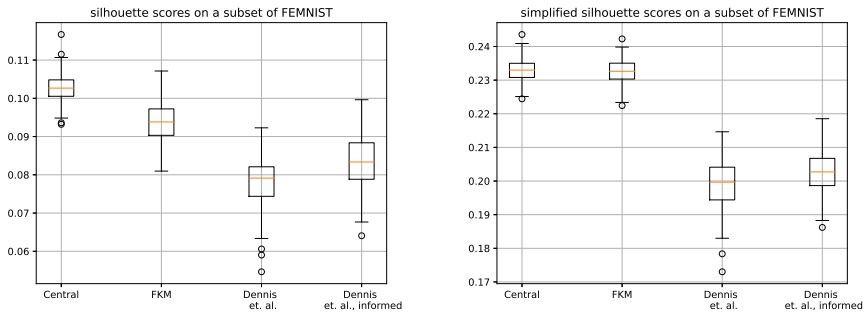
4.4.5. Clustering higher dimensional real data

So far, all our experiments have been done on two dimensional, simulated data. For many use cases, however, data has a much higher dimensionality. In order to determine performance on a higher dimensional dataset, the Federated Extended MNIST (FEMNIST) from LEAF [14] (having a dimensionality of 784) was used, which separates the original Extended MNIST [15] handwritten numbers and letters based on the person who wrote them. FEMNIST has a dimensionality of 784. This leaves approximately 110 datapoints per client; see supplement 4.6.3 for the distribution. Only 10 clients were used from the original FEMNIST, as this drastically sped up the experiments, while keeping enough data for a meaningful assessment. We set k = 60, in line with earlier experiments from Dennis et. al [11]. Figure 4.6a shows that our method outperforms both settings of the method from Dennis et. al. . There is still a difference with a central clustering, however. This could be due to the relatively small amount of samples per client compared to the amount of dimensions, decreasing the quality of the local clusters.

The FEMNIST experiments use the silhouette score [16] as their performance metric. The silhouette score involves calculating distances from each point in a dataset to each other point in a dataset. This means that, to calculate a 'global' silhouette score, distances between datapoints from different clients need to be determined, something that can not be done in a straightforward federated manner. In our case, the simulated federated environment made it possible to calculate the silhouette score for evaluation purposes.

In a real-life setting, the simplified silhouette score [17] could be a suitable alternative, as it only calculates distances between datapoints and cluster means, something which can be done on all clients separately.

We compare the simplified silhouette score with the silhouette score from the same experiments in Figure 4.6b. There seems to be a high correlation between the two scores for a given method, which is in line with previous work [18].



(a) (b)

Figure 4.6.: Results on (a subset of) FEMNIST. (a) shows the silhoutte score, (b) the simplified silhouette score.

4.4.6. Clustering real biological data

Finally, we explore a common clustering task in bioinformatics, that of cell-type identification. We use data from Bouland et. al. [19], specifically their dataset referred to as 'four cancers'. This dataset consists of single-cell RNA sequence measurements from 12 different cancer patients with one of four cancer diagnoses: Lung, endo, colon and renal cancer. In total, there are 22815 genes (features) measured in 132549 cells (samples). These samples are either from tumor tissue or from normal tissue, adjacent to tumour tissue.

Before distributing the dataset over separate clients, we run a standard single-cell pipeline protocol [1] using the Seurat R library [20]. Briefly, we first filter out genes that have less than 200 or more than 2500 feature counts. We then log-normalize the data. Afterwards, we run the Seurat function "FindVariableFeatures" to find the top 500 genes with most variance. Then, we select only the samples from tumor tissue, leaving a total of 68905 samples. Finally, we perform dimensionality reduction on these samples using the selected 500 genes, to reduce dimensionality to 5 (determined using the elbow method).

[1] https://satijalab.org/seurat/articles/pbmc3k_tutorial.html

Data is grouped by patient, meaning that all data from a single patient will end up on a single client. Data is being distributed over three clients. Two different distributions of the data are considered, denoted as IID (Identically and Independently Distributed) or non-IID. For the IID-distribution, each client gets data from every cancer type, whereas for the non-IID data, each client gets data from only one cancer type, as well as from two out of six lung patients.

We run FKM, as well as central clustering and kFed (the method by Dennis et. al. [11]), with a (global) K of 4, equivalent to the amount of cancers in the dataset. Figure 4.7 shows the silhouette and simplified silhouette scores for all algorithms. In the IID setting, our method slightly outperforms kFed, while there is still a slight gap with a centralized clustering. However, when considering the non-IID setting, we observe that the performance gap between FKM and KFed increases, whereas the difference between FKM and a centralized clustering is similar compared to the IID setting, indicating that FKM is more robust to various data distributions.

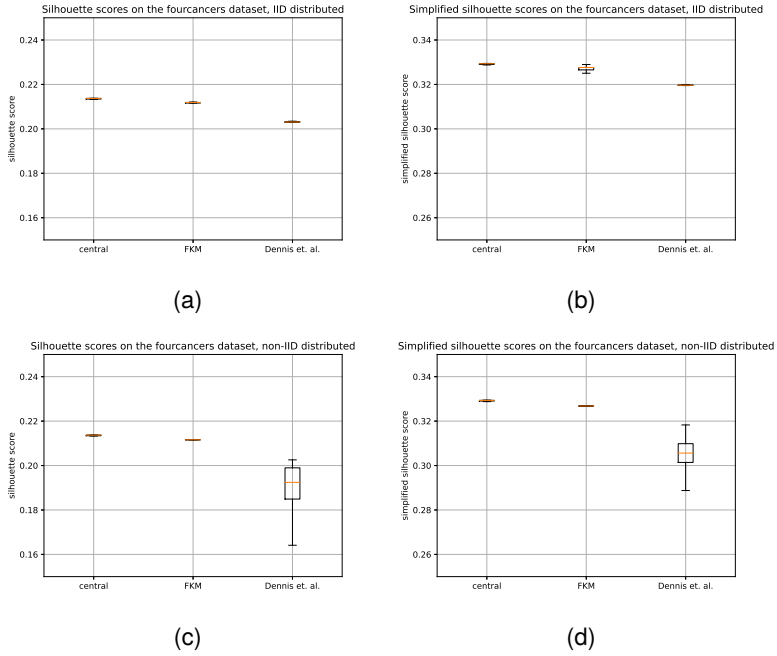


Figure 4.7.: Silhouette and simplified silhouette scores on the fourcancers dataset. Both the silhouette score in (a) as well as the simplified silhouette score in (b) show increased performance of FKM compared to kFed, though still slightly underperforming a centralized clustering. However, when the data is distributed non-IID (figure (c) and (d)), the gap between the federated methods seems to increase.

4.5. Discussion and conclusion

This work describes the implementation and validation of a federated k-means clustering algorithm (FKM), enabling clustering over multiple datasets without sharing the underlying data. Our results show performances close to a central method, in which all data is brought into a single location. There are still some scenarios in which our method shows larger variability in performance as compared to a central clustering, however. These are mostly the more difficult scenarios, such as when there is an extreme distribution in the amount of clusters present on each client, or when the data has a high dimensionality as with the FEMNIST experiment. Assessment of our method on more heterogeneous and 'real life' datasets is therefore an important direction for future work. Nevertheless, FKM has shown to be a promising method in finding similarities among distributed datasets without the need of sharing any data.

Bibliography

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y. Arcas. 'Communication-Efficient Learning of Deep Networks from Decentralized Data'. In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Ed. by A. Singh and J. Zhu. Vol. 54. Proceedings of Machine Learning Research. PMLR, 20–22 Apr 2017, pp. 1273–1282. URL: https://proceedings.mlr.press/v54/mcmahan17a.html.

[2] G. H. Lee and S. Y. Shin. 'Federated Learning on Clinical Benchmark Data: Performance Assessment'. In: *Journal of Medical Internet Research* 22.10 (Oct. 2020). ISSN: 14388871. DOI: 10.2196/20891.

[3] A. Sadilek, L. Liu, D. Nguyen, M. Kamruzzaman, S. Serghiou, B. Rader, A. Ingerman, S. Mellem, P. Kairouz, E. O. Nsoesie, J. MacFarlane, A. Vullikanti, M. Marathe, P. Eastham, J. S. Brownstein, B. A. Y. Arcas, M. D. Howell and J. Hernandez. 'Privacy-first health research with federated learning.' In: *NPJ digital medicine* 4.1 (Sept. 2021), p. 132. ISSN: 2398-6352. DOI: 10.1038/s41746-021-00489-2. URL: http://www.ncbi.nlm.nih.gov/pubmed/34493770%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC8423792.

[4] T. Li, A. K. Sahu, A. Talwalkar and V. Smith. 'Federated learning: Challenges, methods, and future directions'. In: *IEEE signal processing magazine* 37.3 (2020), pp. 50–60.

[5] Q. Li, Z. Wen, S. Hu, N. Wang, Y. Li, X. Liu and B. He. 'A Survey on Federated Learning Systems: Vision, Hype and Reality for Data Privacy and Protection'. In: *IEEE Transactions on Knowledge and Data Engineering* (2021), pp. 1–1. DOI: 10.1109/TKDE.2021.3124599.

[6] J. Hartigan and M. Wong. 'Algorithm AS 136: A K-Means Clustering Algorithm'. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28 (1979), pp. 100–108.

[7] H. H. Kumar, V. R. Karthik and M. K. Nair. 'Federated K-Means Clustering: A Novel Edge AI Based Approach for Privacy Preservation; Federated K-Means Clustering: A Novel Edge AI Based Approach for Privacy Preservation'. In: *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)* (2020). DOI: 10.1109/CCEM50674.2020.00021.

[8] R. Hou, F. Tang, S. Liang and G. Ling. 'Multi-Party Verifiable Privacy-Preserving Federated k-Means Clustering in Outsourced Environment'. In: *Security and Communication Networks* 2021.1 (2021), p. 3630312.

[9] Y. Liu, Z. Ma, Z. Yan, Z. Wang, X. Liu and J. Ma. 'Privacy-preserving federated k-means for proactive caching in next generation cellular networks'. In: *Information Sciences* 521 (June 2020), pp. 14–31. ISSN: 00200255. DOI: 10.1016/J.INS.2020.02.042.

[10] M. Servetnyk, C. C. Fung and Z. Han. 'Unsupervised Federated Learning for Unbalanced Data'. In: *GLOBECOM 2020 - 2020 IEEE Global Communications Conference* (2020). DOI: 10.1109/GLOBECOM42002.2020.9348203.

[11] D. K. Dennis, T. Li and V. Smith. 'Heterogeneity for the Win: One-Shot Federated Clustering'. In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 2611–2620. URL: https://proceedings.mlr.press/v139/dennis21a.html.

[12] D. Arthur and S. Vassilvitskii. *k-means++: The Advantages of Careful Seeding*. Technical Report 2006-13. Stanford InfoLab, June 2006. URL: http://ilpubs.stanford.edu:8090/778/.

[13] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin and V. Chandra. 'Federated Learning with Non-IID Data'. In: *CoRR* abs/1806.00582 (2018). arXiv: 1806.00582. URL: http://arxiv.org/abs/1806.00582.

[14] S. Caldas, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith and A. Talwalkar. 'LEAF: A Benchmark for Federated Settings'. In: *CoRR* abs/1812.01097 (2018). arXiv: 1812.01097. URL: http://arxiv.org/abs/1812.01097.

[15] G. Cohen, S. Afshar, J. Tapson and A. van Schaik. 'EMNIST: an extension of MNIST to handwritten letters'. In: *CoRR* abs/1702.05373 (2017). arXiv: 1702.05373. URL: http://arxiv.org/abs/1702.05373.

[16] P. J. Rousseeuw. 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis'. In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: https://doi.org/10.1016/0377-0427(87)90125-7. URL: https://www.sciencedirect.com/science/article/pii/037704278790125

[17] E. Hruschka, L. de Castro and R. Campello. 'Evolutionary algorithms for clustering gene-expression data'. In: *Fourth IEEE International Conference on Data Mining (ICDM'04)*. 2004, pp. 403–406. DOI: 10.1109/ICDM.2004.10073.

[18] F. Wang, H.-H. Franco-Penya, J. D. Kelleher, J. Pugh and R. Ross. 'An Analysis of the Application of Simplified Silhouette to the Evaluation of k-means Clustering Validity'. In: *Machine Learning and Data Mining in Pattern Recognition*. Ed. by P. Perner. Cham: Springer International Publishing, 2017, pp. 291–305. ISBN: 978-3-319-62416-7.

[19] G. A. Bouland, A. Mahfouz and M. J. T. Reinders. ‘Differential analysis of binarized single-cell RNA sequencing data captures biological variation’. In: *NAR Genomics and Bioinformatics* 3.4 (Dec. 2021), lqab118. ISSN: 2631-9268. DOI: 10.1093/nargab/lqab118. eprint: https://academic.oup.com/nargab/article-pdf/3/4/lqab118/41881586/lqab118.pdf. URL: https://doi.org/10.1093/nargab/lqab118.

[20] Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda and R. Satija. ‘Dictionary learning for integrative, multimodal and scalable single-cell analysis’. In: *Nature Biotechnology* (2023). DOI: 10.1038/s41587-023-01767-y.

4.6. Appendices

4.6.1. Background on k-means and k-means++

k-means clustering

The objective of a clustering algorithm is to partition a given dataset into several subsets with similar features. The k-means clustering algorithm does so by trying to minimize the within cluster sum-of-squares criterion:

$$F_{km} = \sum_{j=0}^{m} \min_{C_i \in C} (||X_j - C_i||^2) \tag{4.3}$$

with m the amount of samples, C_i the cluster mean of cluster i, C the set of all cluster means and $X_{j,k}$ being data point j assigned to cluster k. The procedure in which the k-means algorithm tries to minimize equation 4.3 consists of two steps. First, all data points get assigned to the cluster mean according to the lowest euclidean distance. Then, the mean center point from all points assigned to a certain cluster is calculated. This is done for every cluster, creating a new set of means to start the next round with. This process is repeated until the change within these means is smaller than a certain threshold (and the algorithm has reached convergence) [6].

k-means++

One of the drawbacks of classical k-means clustering is that its initialization is sampled uniformly from the underlying data. This means that having initial cluster means that all come from the same cluster is as probable as having initial cluster means spread across all clusters. Although the K-means algorithm itself can somewhat compensate for this, it still leads to large variability in performance. Arthur and Vassilvitskii developed an initialization method for K-Means to combat this high variability, called k-means++ [12]. Instead of sampling K cluster means from the data with uniform probability, datapoints get weighted based on their distance to the closest already mean that is chosen, with larger distances giving larger weights. This results in (on average) initializations that are more distributed over the space, and prevents (on average) initial cluster means from starting very close to each other, decreasing k-means performance.

4.6.2. Extra results on increasing amount of points per cluster

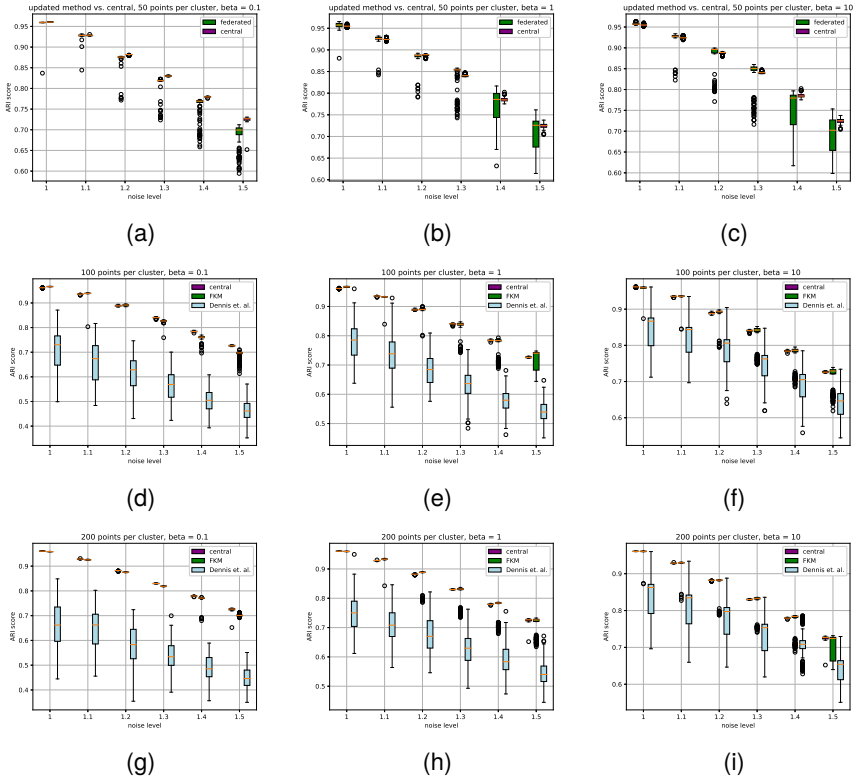


Figure 4.8.: results on using different levels of noise for different values of β , with differing amounts of points per cluster. fom left to right, the columns correspond to β = 0.1, 1 and 10 respectively. From top to bottom, the rows correspond to 50, 100, and 200 points per cluster.

4.6.3. FEMNIST distribution

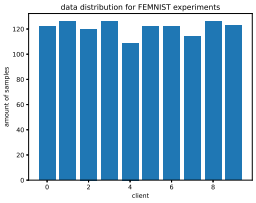


Figure 4.9.: Sample distribution for the FEMNIST dataset

5
Discussion

In this part of the thesis, we explore the use of federated learning in the medical setting. In chapter 3, we benchmark various machine learning models, under various distributions, to show that federated learning is robust to mild heterogeneous distributions. This view aligns with the reviewed literature reviewed in chapter 2, which also highlights that the field is moving beyond federated learning into a more holistic federated analysis. Finally, in line with enabling federated analysis, we introduce a novel federated K-means clustering algorithm in chapter 4, which is robust against a varying amount of local clusters and heterogeneous data splits.

5.1. Challenges to federated learning deployment

5.1.1. Privacy

As federated learning is becoming an established field of research, more data analysis methods beyond machine learning are brought into the federated paradigm (chapter 2), enabling fully distributed analysis pipelines. While these advances enable more federated learning applications, it raises the question which derivations of personal data can be shared. There is a gray area between sharing model parameters (which, in fact, also carries privacy risks [1]) or simple summary statistics such as population mean and variance on one hand, and sharing raw data on the other hand. One complicating factor is that within this gray area, whether derived data can be used to reconstruct the original data can be dependent on what other data is (publicly) available. For example, in 2006, Netflix published a large anonimised dataset for a contest they were hosting. When combined with public data from the Internet Movie DataBase (IMDB), researchers were able to recover the identities, and with it all movie ratings, of the anonimised Netflix users [2].

The combination of analysis methods into pipelines further complicates the data privacy analysis of these methods. Similar to the Netflix case, a single federated method might guarantee privacy because the data is unrecoverable based on what is shared. However, when combined with another method, which might have a similar approach of sharing permuted data, combining the two methods could allow for the reconstruction of the original data [3]. Since data analysis pipelines are modular, there is a combinatorial explosion of method combinations, making it nearly impossible to keep all these combinations in mind when constructing a new federated data analysis method. Therefore, systematic approaches to determine the degree of privacy preservation of federated analysis pipelines are required for privacy guarantees.

5.1.2. Lessons from distributed learning

While federated learning and distributed learning bear many similarities, they originate from different fields of research. Distributed learning comes from the field of distributed optimization and signal processing, whereas federated learning is more often associated with the area of machine learning.

This creates the risk that the federated learning field might redevelop algorithms similar to existing distributed learning work. Especially with the emergence of distributed federated learning [4], the risk of reinventing the distributed learning wheel becomes significant.

Some of the advantages of distributed learning over federated learning include the omission of a single point of failure, as well as better scalability to more clients [4]. While these advantages should be considered for certain use cases, their implications for the health care domain seem scarce. Even in the case of medical wearable devices, which requires high scalability and uses (unreliable) wireless communication links, using a central orchestrator is the dominant choice [5, 6]. Distributed learning requires peer-to-peer communication between devices in order to be implemented efficiently. With modern day telecommunication infrastructure, this is much harder to achieve compared to enabling internet access through WiFi or cellular networks, especially when these wearable devices are not located close to each other. Therefore, federated learning seems more advantageous than distributed learning for healthcare applications.

5.1.3. Platforms

In section 1.2, we mention various software implementations that enable federated learning. Over time, various platforms have emerged and were discontinued, significantly altering the federated learning landscape. Understanding how this landscape might change in the near future is important for researchers seeking to use these platforms for long-term use. Similar centralized deep learning, where two libraries (pyTorch and tensorFlow) enable the vast majority of applications, it seems likely that a small number of platforms will dominate and develop into the status quo. Unlike the centralized machine learning libraries, the divide between simulation and deployment of federated platforms could allow for more diversity within the federated platform software landscape. Specifically, deployment platforms should focus on the practical challenges facing federated learning (some of which we discussed in addendum 9.5), including secure remote data exploration and analysis, as well as increasing remote debugging capabilities.

5.1.4. Data heterogeneity

Under most distributions, federated models attain similar performances compared to centralized models (chapters 2 and 3). However, how to handle extreme non-iid partitionings is still an open question. Results from [7] (table 1.1) show that there is currently not a single aggregation scheme that can handle all of these partitionings, while simultaneously attaining competitive performance for IID distributions.

The current focus of federated performance research is on heterogeneity induced by label skew (such as shown in fig. 1.2). This is partially because it is relatively straightforward to induce such a distribution skew, creating benchmark settings which are used to explore novel aggregation schemes.

However, label skew as extreme as shown in fig. 1.2b rarely occur in reality. Furthermore, in a setting in which each client only has data on 20% of the labels, one can wonder whether a shared model would be beneficial to any of the collaborators.

Another type of heterogeneity is data skew, where clients have different data distributions for the same labels. Although there is relatively little literature addressing data skew, it seems much more realistic to appear in real use cases. For example, batch effects are a known phenomenon in biological data when data originates from multiple institutes. It is therefore interesting to see whether this type of data heterogeneity has a large impact on the performance of federated models or not. Li et. al. [7] attempts to induce such heterogeneity by inducing various amounts of noise, or by utilizing heterogeneous real world partitions, but their results show that these inflicted heterogeneities only have a small effect on model performance. Moreover, the small differences between federated and central models, as shown in chapter 2, suggest that this type of heterogeneity can be dealt with by federated models. This could be caused by the fact that aggressive data skew, eventually, also hampers centrally trained models. If data is skewed in such a way that it reduces performance of a model trained on pooled data, it is reasonable to expect a similar drop in performance for a federated model. In the central paradigm, data skew is often mitigated using batch effect correction methods [8]. When some of these methods have been adapted to the federated paradigm [9, 10], federated learning can handle this type of data skew similarly as the central case.

However, detecting (the severity of) batch effects in federated learning is still an open problem. In centralized learning, batch effects are often detected qualitatively, using visualization methods such as UMAP [11] or tSNE [12] on the combined data, or quantitatively with similarity metrics between and within batches. While summary statistics such as feature wise mean and standard deviation can be effective exploratory metrics, to fully characterize batch effects would require more in-depth analysis that is currently out of reach for the federated paradigm. Hence, developing metrics that can effectively characterize batch effects without sharing data is an open problem for future work.

5.1.5. Personalized versus global federated learning

Another class of solutions to the heterogeneous data problem is what is known as personalized federated learning [13]. The premise of personalized FL is to accept that there are differences between client data that are too large to create a single model that can represent the entire data distribution. Instead, we create personalized models that incorporates only relevant information from other clients, as to still benefit from the federated setup. To achieve this, it is necessary to determine to what degree other clients are beneficial to each client. Various solutions exist, for example altering the loss function [14], clustering clients based on their data distributions [15], or by interpolating between local and global models [16].

While personalized FL has shown promising results, the line between personalization and overfitting seems thin. Personalized federated learning models are by design less generalizable than a global federated model, and thus might be prone to overfitting on the local distribution, which could make them less applicable to new institutes. Furthermore, these personalized models might deteriorate on newly generated data. While client drift (i.e. changes in data distribution of clients over time) remains a challenge outside for federated learning in general [17], a personalized model might suffer more from such a drift compared to a global, more generalizable model.

These issues have lead to new methods attempting to combine the best of both worlds: maintain global models that correspond to (subsets of) the collection of clients, while also keeping local models that best fit specific clients [18]. Furthermore, if there is a collection of personalized models to choose from, new sites are able to determine which personalized model fits their data best, ultimately leading to higher performances across the entire federation. Nevertheless, in the ideal scenario, a global model accounts for the overarching data distribution, without loss of performance on parts of this distribution. While the pursuit of such a federated learning algorithm is still ongoing, personalizing a federated model seems like a stop-gap solution to the heterogeneous data distribution problem. The degree to which personalizing a federated model leans into overfitting on the local distribution, versus it simply being a finetuning step for higher utility within the context of a single client, remains a relevant discussion.

5.2. Federated clustering

5.2.1. Privacy in federated clustering and evaluation

With federated clustering algorithms still emerging, their privacy and security guarantees have not been scrutinized to the same level as for supervised federated learning. For example, in chapter 4 we create a federated version of K-means clustering, for which we share local cluster means as well as cluster sizes. We argue that these measures are similar to summary statistics, which are usually considered safe to share, because both cluster means and sizes are aggregated over samples. However, we have to note that if a cluster consists of a single (or very few) samples, the shared cluster mean might be too similar to the sample and should be filtered out before sharing. One major difference with regards to privacy between supervised and unsupervised learning is evaluation. In supervised learning, most performance metrics such as accuracy can be determined locally, shared and using a (weighted) average, be determined for the entire federation. Clustering metrics often require the use of sample level statistics, such as the distance to cluster means or the clustering assignment. Some of these metrics can be made federated in a secure way, such as the Davies-Bouldin index [19], or the simplified silhouette score (chapter 4). However, when combining the aggregate statistics required for clustering and evaluation, privacy leaks could still emerge [20]. A thorough analysis of privacy guarantees, including both clustering and evaluation stages, is vital to ensure the

privacy-preserving nature of federated clustering. Furthermore, privacy enhancing techniques, similar to differential privacy in supervised learning, should be explored.

5.2.2. Performance in heterogeneous settings

The major factor known to detrimentally influence supervised federated learning is label skew [21]. However, since federated clustering does not use label information, it is unclear whether this is also the case for federated clustering, or if a comparable setting exists. Furthermore, if label skew happens to be detrimental to federated clustering, this can be hard to detect since label information is not necessarily available. Exploring heterogeneous data settings in order to see whether federated clustering struggles with certain distributions should be explored.

In federated averaging, a key selection of hyperparameters can help alleviate issues caused by heterogeneous data. Specifically, lowering the amount of local iterations can, at the cost of a higher communication overhead, increase overall performance in the supervised federated learning setting. Although we do not experiment with the amount of local iterations of FKM in chapter 4, the parallel between local k-means iterations in FKM and the amount of local epochs in federated averaging is evident. However, we currently lack the theoretical and practical foundation to ensure whether altering the amount of local iterations in FKM induces the same tradeoff between communication efficiency and adaptability to heterogeneous data settings as is present in federated averaging.

5.2.3. Beyond federated K-means

Clustering is the process of finding groups of similar data points. In (federated) K-means clustering, these groups are characterized by the center (or mean) of this group.

This cluster representation makes K-means well suited for the federated paradigm. These cluster representations are aggregated over a region of the local data, making them less prone to privacy risks. Therefore, they can usually be shared among clients, which eases the adaptation of these algorithms from central to federated.

However, not all clustering algorithms deploy such cluster representations. Algorithms such as spectral or hierarchical clustering [22, 23] directly use the similarity between samples. In order to create such a similarity metric across all samples, data from different clients need to be compared directly, which would require raw data sharing. How to adapt these types of clustering algorithms to the federated setting remains a challenging avenue. More generally, methods which require sample similarity currently seem out of reach for the federated setting. Solving the problem of determining similarity across clients, for example by constructing a global K-nearest neighbor graph [24], could enable applications such as patient similarity networks between hospitals [25].

5.3. Concluding remarks

In the first part of this dissertation, we explore various aspects of federated learning. We assess the performance of federated learning, showing that it is robust to most mild heterogeneous distributions. Furthermore, we introduce a federated k-means clustering algorithm, paving the way towards federated learning beyond supervised learning. As the field is developing, federated learning is becoming a proven technique, and is generating more potential to speed up collaborative research on highly sensitive data. However, challenges remain in the area of federated analyses, mostly with respect to privacy and performance guarantees. Yet, facing these challenges could allow federated learning to transform into a broader distributed data analysis paradigm.

Bibliography

[1] J. Geiping, H. Bauermeister, H. Dröge and M. Moeller. 'Inverting gradients - How easy is it to break privacy in federated learning?' In: *Advances in Neural Information Processing Systems* 2020-Decem.1 (2020), pp. 1–11. ISSN: 10495258. arXiv: 2003.14053.

[2] A. Narayanan and V. Shmatikov. 'How to break anonymity of the netflix prize dataset'. In: *arXiv preprint cs/0610105* (2006).

[3] R. Nasirigerdeh, R. Torkzadehmahani, J. Baumbach and D. B. Blumenthal. 'On the privacy of federated pipelines'. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 1975–1979.

[4] L. Yuan, Z. Wang, L. Sun, P. S. Yu and C. G. Brinton. 'Decentralized Federated Learning: A Survey and Perspective'. In: *IEEE Internet of Things Journal* 11.21 (2024), pp. 34617–34638. DOI: 10.1109/JIOT.2024.3407584.

[5] K. Arikumar, S. B. Prathiba, M. Alazab, T. R. Gadekallu, S. Pandya, J. M. Khan and R. S. Moorthy. 'FL-PMI: federated learning-based person movement identification through wearable devices in smart healthcare systems'. In: *Sensors* 22.4 (2022), p. 1377.

[6] Y. Chen, X. Qin, J. Wang, C. Yu and W. Gao. 'Fedhealth: A federated transfer learning framework for wearable healthcare'. In: *IEEE Intelligent Systems* 35.4 (2020), pp. 83–93.

[7] Q. Li, Y. Diao, Q. Chen and B. He. 'Federated learning on non-iid data silos: An experimental study'. In: *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE. 2022, pp. 965–978.

[8] H. T. N. Tran, K. S. Ang, M. Chevrier, X. Zhang, N. Y. S. Lee, M. Goh and J. Chen. 'A benchmark of batch-effect correction methods for single-cell RNA sequencing data'. In: *Genome biology* 21.1 (2020), p. 12.

[9] M. Bakhtiari, S. Bonn, F. Theis, O. Zolotareva and J. Baumbach. 'FedscGen: privacy-preserving federated batch effect correction of single-cell RNA sequencing data'. In: *Genome Biology* 26.1 (2025), p. 216.

[10] S. Silva, N. P. Oxtoby, A. Altmann, M. Lorenzi and A. D. N. Initiative. 'FedComBat: A Generalized Federated Framework for Batch Effect Harmonization in Collaborative Studies'. In: *bioRxiv* (2025). DOI: 10.1101/2023.05.24.542107. eprint: https://www.biorxiv.org/content/early/2025/05/14/2023.05.24.542107.full.pdf. URL:

https://www.biorxiv.org/content/early/2025/05/14/2023.05.24.542107.

[11] L. McInnes, J. Healy and J. Melville. 'Umap: Uniform manifold approximation and projection for dimension reduction'. In: *arXiv preprint arXiv:1802.03426* (2018).

[12] L. v. d. Maaten and G. Hinton. 'Visualizing data using t-SNE'. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[13] A. Z. Tan, H. Yu, L. Cui and Q. Yang. 'Towards personalized federated learning'. In: *IEEE transactions on neural networks and learning systems* 34.12 (2022), pp. 9587–9603.

[14] A. Fallah, A. Mokhtari and A. Ozdaglar. 'Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach'. In: *Advances in neural information processing systems* 33 (2020), pp. 3557–3568.

[15] Y. Mansour, M. Mohri, J. Ro and A. T. Suresh. 'Three approaches for personalization with applications to federated learning'. In: *arXiv preprint arXiv:2002.10619* (2020).

[16] F. Hanzely and P. Richtárik. 'Federated learning of a mixture of global and local models'. In: *arXiv preprint arXiv:2002.05516* (2020).

[17] E. Jothimurugesan, K. Hsieh, J. Wang, G. Joshi and P. B. Gibbons. 'Federated learning under distributed concept drift'. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 5834–5853.

[18] M. Chen, M. Jiang, Q. Dou, Z. Wang and X. Li. 'Fedsoup: Improving generalization and personalization in federated learning via selective model interpolation'. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 318–328.

[19] M. Stallmann and A. Wilbik. 'On a framework for federated cluster analysis'. In: *Applied Sciences* 12.20 (2022), p. 10455.

[20] G. Yang, L. Luo and Q. Li. 'Breaking Privacy in Federated Clustering: Perfect Input Reconstruction via Temporal Correlations'. In: *arXiv preprint arXiv:2511.07073* (2025).

[21] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin and V. Chandra. 'Federated learning with non-iid data'. In: *arXiv preprint arXiv:1806.00582* (2018).

[22] A. Ng, M. Jordan and Y. Weiss. 'On spectral clustering: Analysis and an algorithm'. In: *Advances in neural information processing systems* 14 (2001).

[23] F. Nielsen. *Introduction to HPC with MPI for Data Science*. Springer, 2016.

[24] E. Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*. Vol. 1. USAF school of Aviation Medicine, 1985.

[25] S. Pai and G. D. Bader. 'Patient similarity networks for precision medicine'. In: *Journal of molecular biology* 430.18 (2018), pp. 2924–2938.

II

Ageing Biomarkers

6

Introduction

6.1. The concept of Biological ageing

Since 1900, the average life expectancy of a newborn has more than doubled [1]. This rise was mainly due to better nutrition and hygiene for mother, fetus and child and has generally had a positive effect on the physiology and health of individuals from birth onward, as well as on quality of life. It has, however, also lead to new challenges, such as an increase in ageing-related diseases like type 2 diabetes [2], dementia[3] and combinations of such diseases (multi-morbidity). Furthermore, a growing population of older persons increases the overall need for healthcare, leading to shortages in e.g. nursing homes and to increasing costs [4].

Ageing is a complex and heterogeneous process, meaning that, at comparable calendar age, some individuals age significantly slower, while others age at a faster pace. This was suggested to be named biological age (individuals lifetime until death) as opposed to chronological age (individuals lifetime from birth onward). Furthermore, when interventions such as lifestyle changes are applied early in the ageing process, they can have a larger compounding impact over the course of an individual's life. Therefore, being able to recognize those at risk of developing accelerated ageing, whether early in life or in middle age, may help in recommending such interventions offering motivational health checks more effectively, reducing pressure on the healthcare system.

Traditionally, biological age and the pace of ageing have been estimated using clinical and functional measures that capture an individual's physiological decline relative to chronological age [5]. One widely used concept to mark increasing overall vulnerability in older people is frailty. Different metrics of frailty exist, but originally, the scores quantify vulnerability to adverse health outcomes by integrating deficits across multiple domains, including mobility, strength, cognition, and comorbidities [6]. Other conventional measures of frailty preceding diseases include performance-based assessments such as grip strength, walking speed, and balance tests, as well as composite indices of organ system function or multi-morbidity counts [7]. These metrics provide an accessible, clinically relevant estimate of ageing and its acceleration, reflecting the cumulative impact of genetic background, lifestyle, environmental exposures and stochasticity on overall physiological resilience. However, while informative, these approaches are limited in their ability to capture molecular and cellular processes underlying ageing, motivating the development of biomolecular markers that can potentially provide deeper mechanistic insights and are generally easier to measure in epidemiological studies with blood based biobanks in case physiological test batteries are not feasible. In addition to the ageing metrics based on molecular epidemiology, imaging has also been extensively used to relate physiological and mental decline with ageing.

The second part of this thesis discusses ageing biomarkers. In chapter 7, we explore various design choices for proteomics biomarkers, resulting in our own novel ageing biomarker. Furthermore, in chapter 9 we compare metabolomics-based biomarkers with MRI based markers of biological age using a federated learning framework. In

doing so, we uncover that these markers carry non-overlapping information, in line with the different ageing aspects that these markers attempt to characterize.

6.2. Molecular and imaging biomarkers of biological ageing

Recent advancements in biological data acquisition have vastly expanded the capability to measure biological traits, especially in epidemiological studies using community based voluntary participants varying in age, health, socioeconomic status, etc. On the molecular level, measurements spanning various molecular layers, referred to as '-omics' layers, have become widely accessible. These measures span all standardized, holistic, molecular measures between the genetic code and the physiological or clinical phenotype of study, e.g. genomics, transcriptomics, proteomics, and metabolomics. Furthermore, using techniques such as Magnetic Resonance Imaging (MRI), high-fidelity three dimensional images of for example brain structure can be created. Using these widely accessible and cost-effective measuring techniques, large scale studies can be realized, measuring hundreds, if not thousands, of (molecular) features simultaneously in only mildly invasive assays.

The increase in access of -omics data has lead to the rise of molecular ageing biomarkers [8, 9]. Ageing biomarkers are machine learning models that, often using -omics measurements, predict either chronological age or a relevant ageing-related outcome (such as mortality or frailty). By comparing a biomarker score of an individual with the global population, we can assert whether an individual is ageing faster than expected, and make recommendations in order to slow down the ageing process, monitored through actionable molecular markers which are changing with for example a healthier lifestyle. The combination of actionable, motivational molecular monitoring markers with the right intervention to match the risk profile may contribute to prevention of accelerated ageing, and help increasing longevity in an ageing population.

The first generation of biomarkers were designed to predict chronological age. The difference of the chronological age prediction based on molecular data and the true age of the individual provides an indication of biological age (an older biological age refers to an individual having an older molecular profile estimate than expected for one's chronological age). These include biomarkers such as the Horvath and Hannum clocks [10, 11], which uses epigenetic data, and MetaboAge [12], which predicts chronological age based on metabolomics, or the proteomics-based ProteinAge [13]. Recently, more biomarkers predicting mortality have emerged, such as the epigenetic PhenoAge [14] and GrimAge [15], metabolomics-based MetaboHealth [16] and the proteomics-based Proteomic Aging Clock (PAC) [17]. Various studies have shown that mortality-based biomarkers generally outperform chronological age-based biomarkers when assessing their associations with ageing-related outcomes such as frailty [18, 19].

6.3. How to train a molecular ageing biomarker

Ageing biomarkers can be characterized along three dimensions: the input data, the modeling approach, and the prediction outcome. The type of input data determines the biological signal that can be exploited. When omics is measured in blood, epi -and transcriptome levels originating from blood cells are indicative of systemic physiological conditions [20, 21], while -omics layers such as proteomics and metabolomics reflect transient cellular and tissue states more directly [22–24].

Modern technologies have vastly increased the number of assessed features, such as the number of proteins or metabolites, that can be measured simultaneously. While this increase in features has the potential to uncover more intrinsic biological signals or relations between features, it also raises the risk of overfitting, where the model may rely on spurious correlations that are specific to the sampled dataset, rather than reflecting true biology. In addition, larger assays are more costly, which can limit the clinical accessibility of models trained on such comprehensive data.

While the input data refers to the type of biological signal being processed, the modeling approach determines what type of relationships can be modeled within that data. One key decision in model selection is whether to use a linear or nonlinear approach. Linear models predict the output label as a weighted linear combination of the input signals. The optimal solutions for these models can usually be calculated analytically, allowing them to be trained quickly and deterministically, but at the cost of being limited to modeling linear relationships only.

On the other hand, non-linear models, such as deep neural networks [25], characterize the relationship between input and output using non-linear functions. This allows them to capture complex patterns in the data more accurately. However, this flexibility comes at the cost of being unable to analytically determine the optimal solution, requiring optimization algorithms such as stochastic gradient descent, which are prone to getting stuck in local optima [26].

The final biomarker aspect, the target outcome, determines what the ageing biomarker is designed to predict. The ideal target outcome depends on the intended purpose of a biomarker: If its main function is to predict a specific (ageing related) disease (onset or progression), then directly predicting that disease in community or patient based data is the most straightforward approach. However, when the goal is to capture broader aspects of ageing, such as global health or combinations of vulnerabilities, a more holistic training target is preferred.

While most biomarkers are trained to predict either chronological age or mortality, other potential training targets for creating an ageing biomarker for holistic representation remain underexplored. One such candidate target is frailty [6]. Commonly used in geriatrics, frailty is used to characterize a patient's health based on diagnoses, questionnaires and/or functional tests. Frailty scores aggregate a broad spectrum of ageing-related deficits, making them a promising training target for biomarkers

aiming at providing a global indication of biological ageing.

Although a plethora of blood-based biomarkers have emerged over the last couple of years, there seems to be little consensus on best practices for the three aspects of biomarker design (data input, modeling approach and training target). There is a general trend towards using larger assays, predicting mortality instead of age, and using more complex deep learning models in favor of simpler linear machine learning models. However due to lack of systematic comparison, there is no consistent evidence that these trends result in better biomarkers. In this thesis, chapter 7 will perform such a systematic analysis.

6.4. Comparing ageing biomarkers

Since ageing is a heterogeneous process, it is no surprise that different ageing biomarkers capture different aspects of ageing. Certain blood-based biomarkers focus more on cardiovascular decline [12, 16] or immune metabolic health aspects that contribute to physiological decline of all tissues, especially when trained to predict all-cause mortality. On the other hand, MRI-based markers zoom in on cognitive decline, by using MRI scans of the brain in combination with large deep learning models to characterize brain ageing [27].

Although various biomarkers with different focuses have emerged, the relationships between them remain largely unknown. Gaining a better understanding of how these biomarkers relate to one another can guide the development of future, more holistic biomarkers, and provide insight into the underlying ageing mechanisms that different markers are characterizing.

One way to compare biomarkers is to generate their scores within a given cohort and assess their association. However, single cohorts can be prone to cohort-specific selection biases. Combining data from multiple cohorts would enable a more robust comparison. Unfortunately, since the data used to train such biomarkers is highly privacy-sensitive, sharing the raw data, or the resulting biomarker scores, is challenging.

Federated learning [28] offers a solution to this data-sharing challenge. By enabling training a single model or analysis across multiple cohorts, federated learning allows for the comparison of biomarkers across datasets without the need to share patient-level data. In chapter 9 we present such a solution in which we compare a brain MRI-based biomarker with a metabolomics-based biomarker across three different institutes using a federated approach.

6.5. Thesis contributions on ageing biomarkers

This part of the thesis focuses on two main research questions: First, what is the optimal approach for creating a proteomics-based ageing biomarker? And second, How do biomarkers which focus on different aspects of the ageing process, that are based on different data modalities, compare to each other?

To answer the first question, chapter 7 zooms in on molecular biomarker design, investigating best practices on input data, model choice and training target. In doing so, we introduce two new ageing biomarkers, protFI, and protMort, trained using a minimal, yet robust, set of proteins to predict the frailty index and mortality, respectively. In particular, ProtFI outperforms various other ageing biomarkers on a variety of ageing related endpoints, showing the efficacy of the frailty index as a training target. Chapter 8 expands on chapter 7 with additional modeling approaches using a multitask framework.

For answering the second research question, chapter 9 introduces a real-world federated learning framework to train and validate an ageing biomarker based on brain MRI scans across three medical institutes within the Netherlands. Next, we use the same framework to compare this newly trained model with molecular ageing biomarkers, based on metabolomics from blood draws. In doing so, we show that these biomarkers carry orthogonal information, highlighting the heterogeneity between the various ageing aspects that these biomarkers characterize and their potential added value when combined in studies.

Bibliography

[1] S. Dattani, L. Rodés-Guirao, H. Ritchie, E. Ortiz-Ospina and M. Roser. 'Life Expectancy'. In: *Our World in Data* (2023). https://ourworldindata.org/life-expectancy.

[2] M. Halim and A. Halim. 'The effects of inflammation, aging and oxidative stress on the pathogenesis of diabetes mellitus (type 2 diabetes)'. In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 13.2 (2019), pp. 1165–1172. ISSN: 1871-4021. DOI: https://doi.org/10.1016/j.dsx.2019.01.040. URL: https://www.sciencedirect.com/science/article/pii/S187140211930013X.

[3] M. Cortes-Canteli and C. Iadecola. 'Alzheimer's Disease and Vascular Aging: JACC Focus Seminar'. In: *Journal of the American College of Cardiology* 75.8 (2020), pp. 942–951. ISSN: 0735-1097. DOI: https://doi.org/10.1016/j.jacc.2019.10.062. URL: https://www.sciencedirect.com/science/article/pii/S0735109720300565.

[4] R. Aalbers and A.-F. Roos. *Zorguitgaven, ons een zorg?* 2022.

[5] A. Comfort. 'TEST-BATTERY TO MEASURE AGEING-RATE IN MAN'. In: *The Lancet* 294.7635 (1969). Originally published as Volume 2, Issue 7635, pp. 1411–1415. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(69)90950-7. URL: https://www.sciencedirect.com/science/article/pii/S0140673669909507.

[6] A. Clegg, J. Young, S. Iliffe, M. O. Rikkert and K. Rockwood. 'Frailty in elderly people'. In: *The Lancet* 381 (9868 Mar. 2013), pp. 752–762. ISSN: 1474547X. DOI: 10.1016/S0140-6736(12)62167-9.

[7] E. Dent, P. Kowal and E. O. Hoogendijk. 'Frailty measurement in research and clinical practice: A review'. In: *European Journal of Internal Medicine* 31 (2016), pp. 3–10. ISSN: 0953-6205. DOI: https://doi.org/10.1016/j.ejim.2016.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0953620516300279.

[8] M. Moqri, C. Herzog, J. R. Poganik, J. Justice, D. W. Belsky, A. Higgins-Chen, A. Moskalev, G. Fuellen, A. A. Cohen, I. Bautmans *et al.* 'Biomarkers of aging for the identification and evaluation of longevity interventions'. In: *Cell* 186.18 (2023), pp. 3758–3775.

[9] M. Moqri, C. Herzog, J. R. Poganik, K. Ying, J. N. Justice, D. W. Belsky, A. T. Higgins-Chen, B. H. Chen, A. A. Cohen, G. Fuellen *et al.* 'Validation of biomarkers of aging'. In: *Nature medicine* 30.2 (2024), pp. 360–372.

[10] S. Horvath. 'DNA methylation age of human tissues and cell types'. In: *Genome Biology* 14 (3156 2013). DOI: 10.1145/2820783.2820789. URL: https://doi.org/10.1186/gb-2013-14-10-r115.

[11] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao *et al.* 'Genome-wide methylation profiles reveal quantitative views of human aging rates'. In: *Molecular cell* 49.2 (2013), pp. 359–367.

[12] E. B. van den Akker, S. Trompet, J. J. B. Wolf, M. Beekman, H. E. D. Suchiman, J. Deelen, F. W. Asselbergs, E. Boersma, D. Cats, P. M. Elders, J. M. Geleijnse, M. A. Ikram, M. Kloppenburg, H. Mei, I. Meulenbelt, S. P. Mooijaart, R. G. Nelissen, M. G. Netea, B. W. Penninx, M. Slofstra, C. D. Stehouwer, M. A. Swertz, C. E. Teunissen, G. M. Terwindt, L. M. t. Hart, A. M. van den Maagdenberg, P. van der Harst, I. C. van der Horst, C. J. van der Kallen, M. M. van Greevenbroek, W. E. van Spil, C. Wijmenga, A. Zhernakova, A. H. Zwinderman, N. Sattar, J. W. Jukema, C. M. van Duijn, D. I. Boomsma, M. J. Reinders and P. E. Slagboom. 'Metabolic Age Based on the BBMRI-NL ¹H-NMR Metabolomics Repository as Biomarker of Age-related Disease'. In: *Circulation: Genomic and Precision Medicine* 13.5 (2020), pp. 541–547. DOI: 10.1161/CIRCGEN.119.002610. eprint: https://www.ahajournals.org/doi/pdf/10.1161/CIRCGEN.119.002610. URL: https://www.ahajournals.org/doi/abs/10.1161/CIRCGEN.119.002610.

[13] M. A. Argentieri, S. Xiao, D. Bennett, L. Winchester, A. J. Nevado-Holgado, U. Ghose, A. Albukhari, P. Yao, M. Mazidi, J. Lv, I. Millwood, H. Fry, R. S. Rodosthenous, J. Partanen, Z. Zheng, M. Kurki, M. J. Daly, A. Palotie, C. J. Adams, L. Li, R. Clarke, N. Amin, Z. Chen and C. M. van Duijn. 'Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations'. In: *Nature Medicine* 30 (9 2024), pp. 2450–2460. ISSN: 1546-170X. DOI: 10.1038/s41591-024-03164-7.

[14] M. E. Levine, A. T. Lu, A. Quach, B. H. Chen, T. L. Assimes, S. Bandinelli, L. Hou, A. A. Baccarelli, J. D. Stewart, Y. Li *et al.* 'An epigenetic biomarker of aging for lifespan and healthspan'. In: *Aging (albany NY)* 10.4 (2018), p. 573.

[15] A. T. Lu, A. Quach, J. G. Wilson, A. P. Reiner, A. Aviv, K. Raj, L. Hou, A. A. Baccarelli, Y. Li, J. D. Stewart *et al.* 'DNA methylation GrimAge strongly predicts lifespan and healthspan'. In: *Aging (albany NY)* 11.2 (2019), p. 303.

[16] J. Deelen, J. Kettunen, K. Fischer, A. van der Spek, S. Trompet, G. Kastenmüller, A. Boyd, J. Zierer, E. B. van den Akker, M. Ala-Korpela, N. Amin, A. Demirkan, M. Ghanbari, D. van Heemst, M. A. Ikram, J. B. van Klinken, S. P. Mooijaart, A. Peters, V. Salomaa, N. Sattar, T. D. Spector, H. Tiemeier, A. Verhoeven, M. Waldenberger, P. Würtz, G. D. Smith, A. Metspalu, M. Perola, C. Menni, J. M. Geleijnse, F. Drenos, M. Beekman, J. W. Jukema, C. M. van Duijn and P. E. Slagboom. 'A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals'. In: *Nature Communications*

10 (1 2019), pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-019-11311-9.

[17] C.-L. Kuo, Z. Chen, P. Liu, L. C. Pilling, J. L. Atkins, R. H. Fortinsky, G. A. Kuchel and B. S. Diniz. 'Proteomic aging clock (PAC) predicts age-related outcomes in middle-aged and older adults'. In: *Aging Cell* 23.8 (2024), e14195. DOI: https://doi.org/10.1111/acel.14195.

[18] L. M. Kuiper, H. A. Polinder-Bos, D. Bizzarri, D. Vojinovic, C. L. Vallerga, M. Beekman, M. E. T. Dollé, M. Ghanbari, T. Voortman, M. J. T. Reinders, W. M. M. Verschuren, P. E. Slagboom, E. B. van den Akker and J. B. J. van Meurs. 'Epigenetic and Metabolomic Biomarkers for Biological Age: A Comparative Analysis of Mortality and Frailty Risk'. In: *The Journals of Gerontology: Series A* 78 (10 June 2023), pp. 1753–1762. ISSN: 1758-535X. DOI: 10.1093/gerona/glad137.

[19] C. McCrory, G. Fiorito, B. Hernandez, S. Polidoro, A. M. O'Halloran, A. Hever, C. N. Cheallaigh, A. T. Lu, S. Horvath, P. Vineis and R. A. Kenny. 'GrimAge Outperforms Other Epigenetic Clocks in the Prediction of Age-Related Clinical Phenotypes and All-Cause Mortality'. In: *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 76 (5 May 2021), pp. 741–749. ISSN: 1758535X. DOI: 10.1093/gerona/glaa286.

[20] S. Horvath and K. Raj. 'DNA methylation-based biomarkers and the epigenetic clock theory of ageing'. In: *Nature Reviews Genetics* 19 (6 June 2018), pp. 371–384. ISSN: 14710064. DOI: 10.1038/s41576-018-0004-3.

[21] M. J. Peters *et al.* 'The transcriptional landscape of age in human peripheral blood'. In: *Nature Communications* 6 (Oct. 2015). ISSN: 20411723. DOI: 10.1038/ncomms9570.

[22] P. Würtz, A. J. Kangas, P. Soininen, D. A. Lawlor, G. D. Smith and M. Ala-Korpela. 'Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies'. In: *American Journal of Epidemiology* 186 (9 2017), pp. 1084–1096. ISSN: 14766256. DOI: 10.1093/aje/kwx016.

[23] B. Lehallier, D. Gate, N. Schaum, T. Nanasi, S. E. Lee, H. Yousef, P. M. Losada, D. Berdnik, A. Keller, J. Verghese, S. Sathyan, C. Franceschi, S. Milman, N. Barzilai and T. Wyss-Coray. 'Undulating changes in human plasma proteome profiles across the lifespan'. In: *Nature Medicine* 25 (12 Dec. 2019), pp. 1843–1850. ISSN: 1546170X. DOI: 10.1038/s41591-019-0673-2.

[24] F. Bogaards, T. Gehrmann, M. Beekmam, N. Lakenberg, E. Suchiman, R.-U. Mueller, J. Deelen, J.-W. Lackmann, L. de Groot, M. Reinders *et al.* 'Blood Proteomic Biomarkers indicate reduced system and tissue level inflammation responses to The GOTO Lifestyle Intervention in Older Adults'. In: *medRxiv* (2025), pp. 2025–10.

[25] D. Ciregan, U. Meier and J. Schmidhuber. 'Multi-column deep neural networks for image classification'. In: *2012 IEEE conference on computer vision and pattern recognition*. IEEE. 2012, pp. 3642–3649.

[26] K. C. Kiwiel. 'Convergence and efficiency of subgradient methods for quasiconvex minimization'. In: *Mathematical Programming* 90.1 (Mar. 2001), pp. 1–25. ISSN: 1436-4646. DOI: 10.1007/PL00011414. URL: https://doi.org/10.1007/PL00011414.

[27] J. Wang, M. J. Knol, A. Tiulpin, F. Dubost, M. De Bruijne, M. W. Vernooij, H. H. Adams, M. A. Ikram, W. J. Niessen and G. V. Roshchupkin. 'Gray matter age prediction as a biomarker for risk of dementia'. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.42 (Oct. 2019), pp. 21213–21218. ISSN: 10916490. DOI: 10.1073/PNAS.1902376116/-/DCSUPPLEMENTAL.

[28] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas. 'Communication-efficient learning of deep networks from decentralized data'. In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

7

ProtFI, an efficient frailty-trained proteomics-based biomarker of aging, robustly predicts age-related decline

Chronological age overlooks the heterogeneity in aging. In response, a wide range of molecular aging biomarkers has been developed to better capture an individual's aging rate. Yet, a comprehensive comparison of modeling choices in the development of these biomarkers is lacking. In this study, we trained aging biomarkers on the Rockwood frailty index (FI) and all-cause mortality using UK Biobank Olink proteomics and metabolomics (^{1}H-NMR) data (n=40,696). We systematically established the impact of model choice, target outcome, and molecular data source on several age-related outcomes. From this, we developed two novel aging biomarkers, ProteinFrailty (ProtFI), and ProteinMortality (ProtMort), both elastic net models that use a minimal set of proteins to predict FI and mortality, respectively. In particular, ProtFI outperformed established aging biomarkers in relation to diverse outcomes, including incident cardiovascular disease, handgrip strength, and self-rated health, both in internal validation and two Dutch external cohorts (n=995, n=500). Our findings show that an efficient frailty-trained proteomic biomarker robustly predicts age-related decline.

S. Garst, L. Kuiper, E. van den Akker et al. *'ProtFI, an efficient frailty-trained proteomics-based biomarker of aging, robustly predicts age-related decline'. In:* medRxiv *(2025), pp. 2025–09*

7.1. Introduction

As individuals age, their risk of functional decline and disease increases [1, 2]. However, aging is a heterogeneous process: individuals of the same chronological age may age at different rates and in different ways, resulting in varying risks for physical, cognitive, and functional outcomes [3]. Chronological age is commonly used as a proxy for aging but cannot, by definition, account for individual variation in age-related risk. As a result, the concept of biological age was introduced, aiming to reflect the pace and trajectory of aging on an individual level [4, 5].

Alternatively, the concept of frailty describes a state of increased vulnerability to stressors, resulting from decline across multiple physiological systems [6]. Among these, the Rockwood frailty index, further referred to as the frailty index, is one of the most widely used frailty measures [7]. Although distinct in concept, quantified frailty scores are frequently used in research and clinical practice as practical proxies for biological age [5].

In addition to clinical assessments such as frailty, molecular or *omics* biomarkers have more recently been developed to reflect different aspects of biological aging [4, 8]. Aging biomarkers are often derived from blood measurements and can span various molecular layers, each reflecting different aspects of the aging process [4]. Compared to the frailty index, which can include more than 40 age-related deficits [9], blood-based biomarkers are less time-consuming to obtain. Since the conception of the Horvath Clock in 2013 [10], numerous epigenetic clocks have been introduced, some of which incorporate clinical markers to enhance predictive accuracy [11]. In parallel, aging biomarkers derived from other omics data —including metabolomics [12], transcriptomics [13], proteomics [14], and microbiome profiles [15]— have emerged as alternative approaches to assess biological aging. Among these, transcriptome and epigenome signatures derived from blood cells are particularly indicative of cellular states that reflect systemic physiological conditions [11, 13]. Proteins and metabolite signatures, on the other hand, more directly reflect the functional state of peripheral organs [16–18].

First-generation *omics*-based aging biomarkers, known as “aging clocks”, were designed to estimate chronological age. In contrast, second-generation aging biomarkers are trained on health-related metrics, such as composite scores of disease burden or mortality risk[12, 19]. Studies have demonstrated that biomarkers trained on mortality and age-related health data outperform those solely trained on chronological age in predicting health outcomes [20, 21]. Given the broad spectrum of physiological impairments captured by the frailty index, we hypothesize that a proteomics-based aging biomarker trained to predict frailty could provide a robust and mechanistically relevant tool for assessing biological aging. Recently, using frailty as a training outcome has received considerable attention. Some works focus on the use of alternative measures to estimate frailty itself, but do not evaluate frailty as a training target for aging biomarkers [22–24]. Other studies have attempted developing biomarkers using frailty, but either lack analysis regarding associations

beyond frailty itself[25], or are missing external validation [25, 26].

More recently, advances in machine learning have transformed aging biomarker development by enabling more complex, multi-variate approaches. Despite methodological advancements in aging biomarker development, practical implementations of aging biomarkers vary widely across the field. More critically, the choice for using more complex models is often undermotivated, lacking comparison to simpler model baselines, while challenges using these more complex models persist. Systematic work exploring key steps in biomarker development, such as model choice and training target selection, is lacking. Therefore, such choices in existing biomarker work remain mostly underscrutinized.

Furthermore, challenges remain in validation and standardization. Many new biomarkers lack rigorous external replication, and are rarely benchmarked against established biomarkers, making it difficult to assess their relative effectiveness. Consequently, a growing array of biomarkers based on single or multi *omic*-data has emerged [4, 8], but without systematic comparisons, their clinical relevance and reliability remain uncertain.

In this study, we developed aging biomarkers based on using Olink proteomics and Nightingale Health (^{1}H-NMR) metabolomics data. We assessed the impact of model choice, target outcome, and molecular data source, identifying key factors that influence biomarker performance. Based on this analysis, we introduce two novel aging biomarkers: ProteinFrailty (ProtFI) and ProteinMortality (ProtMort), which leverage a minimal subset of proteins to predict frailty and mortality, respectively. By comparing ProtFI and ProtMort to previously established aging biomarkers, we demonstrate that an efficient protein-based approach, trained to predict frailty, provides robust and generalizable insights into aging phenotypes. By validating both aging biomarkers in two independent cohorts, we underscore the need to establish biomarker robustness across independent cohorts and populations

7.2. Results

7.2.1. Study Outline

We systematically investigated the use of proteomics for developing aging biomarkers. We evaluated key design choices in this process, structured into three main components: 1) model selection, 2) protein selection, and 3) prediction target. To assess model selection, we constructed aging biomarkers using ElasticNet (EN) models and deep Feedforward Neural Networks (FNN). These models were selected because they represent popular choices within two categories of models, being simple but linear (EN), or complex and non-linear (FNN). These models were trained on various subsets of plasma proteins from the UK Biobank (UKB) [27, 28], ranging from four Olink panels (1428 proteins), to the cardiometabolic panel only (344 proteins), to a subset of the cardiometabolic panel (20/27 proteins). Each biomarker was designed to predict either the frailty index (determined at the same time point as the protein

Table 7.1.: Population characteristics. Values represent mean and standard deviation between parentheses for continuous variables (chronological age, frailty index, BMI, maximum handgrip strength, percentage of lymphocytes, and percentage of monocytes), median and the interquartile range between brackets for both the mortality as well as the incident cardiovascular disease follow-up time, and the number of participants and the percentage between brackets for the categorical variables (number of women, died during follow-up, highest attained eduction, current smoker, prevalent cardiovascular disease, incident cardiovascular disease, being part of the multi-omics set, and being non-randomly selected by the UK Biobank Pharma Proteomics Project (PPP) as participant).

	UK Biobank	Rotterdam Study	Leiden Longevity Study
Number of participants	40,696	995	500
Age in years	57.5 (8.2)	62.7 (5.8)	66.0 (6.7)
Number of women	21,941 (53.9)	568 (57.1)	252 (50.4)
Frailty index (0-1)	0.13 (0.07)	0.13 (0.07)	Not available
Died during follow-up	4,497 (11.1)	67 (6.7)	95 (19.0)
Years of follow-up for all-cause mortality	13.7 [13.0, 13.5]	10.3 [9.6, 10.7]	13.8 [13.4, 14.1]
Highest attained education			
Primary education	7,293 (17.9)	68 (6.8)	17 (3.4)
Lower vocational or intermediate general education	11,079 (27.2)	323 (32.5)	119 (23.8)
Intermediate vocational or secondary education	8,844 (21.7)	280 (28.1)	190 (38.0)
Higher vocational education or university	12,764 (31.4)	321 (32.3)	168 (33.6)
BMI in kg/m2	27.5 (4.8)	27.3 (4.4)	25.5 (3.7)
Current smoker	4,357 (10.7)	133 (13.4)	31 (6.2)
Maximum handgrip strength in kilogram	32.5 (11.3)	31.2 (10.3)	40.3 (11.7)
Self-rated health poor	11,058 (27.2)	85 (8.5)	49 (9.8)
High blood pressure	11,563 (28.4)	570 (57.3)	119 (23.8)
Prevalent cardiovascular disease	2,040 (5.0)	83 (8.3)	31 (6.2)
Incident cardiovascular disease	3,587 (8.8)	60 (6.6)	28 (7.3)
Years of follow-up for incident cardiovascular disease	11.7 [11.0, 12.5]	6.5 [6.0, 6.9]	5.0 [4.4, 5.9]
Percentage of lymphocytes	28.8 (7.7)	35.0 (7.8)	Not available
Percentage of monocytes	7.1 (2.6)	6.4 (2.0)	Not available
Part of multi-omics set	21,434 (52.7)	583 (58.6)	500 (100)
Selected as PPP participant	4,898 (12.0)	Not applicable	Not applicable

measurement) or all-cause mortality. In the UK Biobank, the frailty index consisted of 49 categorical deficits, which are categorized into sensory, cranial, mental well-being, infirmity, cardiometabolic, respiratory, musculoskeletal, immunological, cancer, pain and gastrointestinal related deficits [9, 29] (specific frailty index deficits used in this study are detailed in Methods and Supplementary Tables 7.1 and 7.6).

All aging biomarkers were trained, tested, and internally and independently validated on the UKB data, and externally validated in the Rotterdam Study (RS) [30] and the Leiden Longevity Study (LLS) [31]. Table 7.1 summarizes the population characteristics. The UKB actively recruited National Health Service (NHS) patients [27], and exhibits a strong healthy volunteer bias [32]. In contrast, the RS is a population-based study [30], while the LLS comprises offspring of long-lived individuals and their partners [31]. For further details, see the study population section in the Methods. Fig. 7.1 shows a graphical representation of the study outline. For a selected set of aging biomarkers, we conducted association analysis with a range of age-related health outcomes, chosen based on their availability with statistical power across the three cohorts. Initial analyses on sex-specific modeling showed sizeable overlap in protein selection for the ElasticNet model on the full 1428 protein set, with the exception for the model trained on only female participants (Supplementary Fig. 7.4). Furthermore, stratification showed no improvement in model performance (Supplementary Table 7.3). We therefore decided not to stratify by sex. All reported effect estimates in the UKB refer to those obtained from the internal validation set.

7.2.2. Comparison between linear and deep learning models shows no benefit for more complex models

We first evaluated the impact of model complexity on aging biomarker performance. We started by using the set of four Olink panels, together making up 1428 proteins (allprot), and trained aging biomarkers based on outcomes and modelling approach. Markers were trained to either predict all-cause mortality or the frailty index, using two approaches: a linear ElasticNet (EN) model and a Feedforward Neural Network (FNN). This resulted in four biomarkers: EN trained on the frailty index and mortality (EN-allprot-FI and EN-allprot-mort), and FNN trained on the frailty index and mortality (FNN-allprot-FI and FNN-allprot-mort).

To assess model performance, we employed bootstrapping with replacement, generating 100 resampled training sets. For each resampled set, we applied the predetermined hyperparameters and evaluated performance metrics on the independent validation set, including R^2 score for the frailty index as outcome and the Concordance-index (C-index) for mortality as outcome. We then compared the distribution of these performance metrics between the resulting EN and FNN aging biomarkers. Bootstrapping revealed a significant difference in performance (p-value after false discovery rate adjustment (pFDR) < 0.001), favoring the EN model over the FNN model for both the frailty index and mortality (Fig. 7.2). On average, the EN biomarkers exhibited an R^2 that was 0.034 (standard deviation (SD) 0.008) higher and

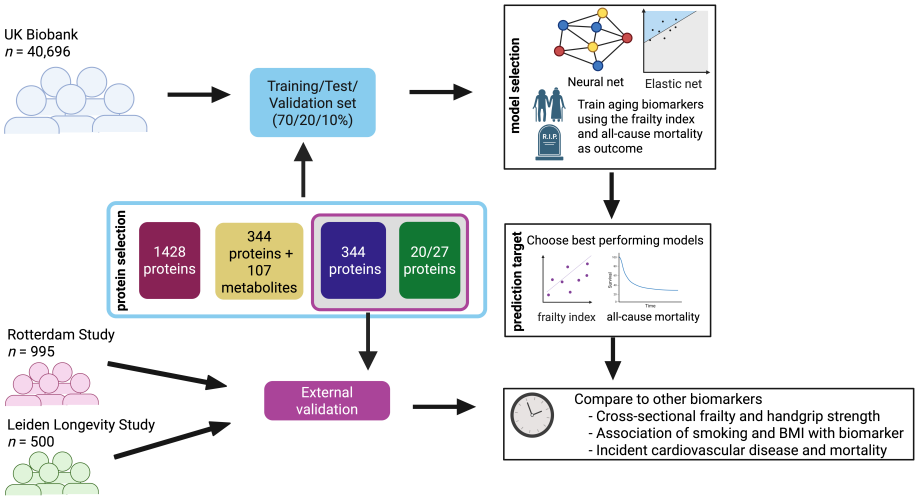


Figure 7.1.: Overview of the study design. We used proteomics data from the UK Biobank as training data for all biomarkers. By creating a 70/20/10 training/test/validation split, we investigated the effect of the number of proteins included in a biomarker, as well as the effect of including metabolomics data. Furthermore, we compared linear ElasticNet and non-linear Feedforward Neural Network models to investigate the effect of using more complex models for biomarker training. We trained all our models to either predict the frailty index or mortality, in order to investigate the effect of training outcome. Finally, biomarkers based on the two smallest protein subsets were externally validated in two cohorts, the Rotterdam Study and the Leiden Longevity Study (as well as the validation set in the UK Biobank), by means of association analyses on several aging-related phenotypes.

a C-index that was 0.015 (SD 0.010) higher than their FNN counterparts. These findings suggest that increased model complexity, as in FNNs, does not enhance performance.

7.2.3. Associations with non-training endpoints remain similar when reducing the number of proteins

For theoretical and future practical reasons, we asked the question how much redundancy exists in the information of the full set of proteins used so far. Therefore, we evaluated the impact of reducing the number of proteins used for aging biomarker training, decreasing the input from 1428 to the cardiometabolic (CMB) panel of 344 proteins (based on Olink panel distribution, see Methods for details). This selection was guided by the availability of this panel in the RS and LLS cohorts, enabling external validation. As in the previous analysis, we trained aging biomarkers using

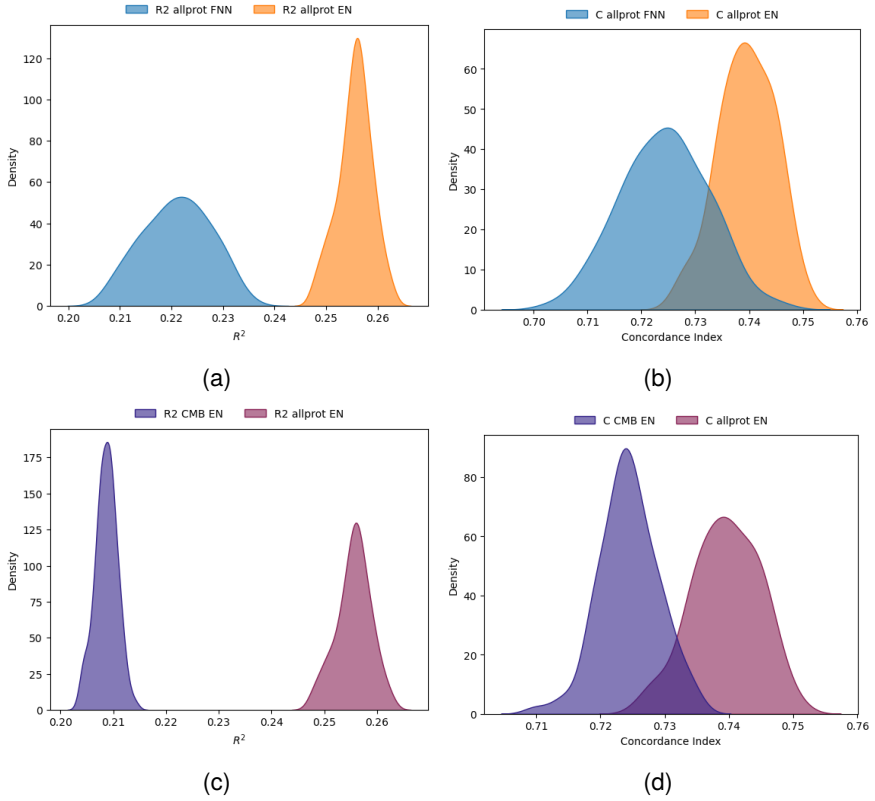


Figure 7.2.: Density plot showing (a) & (c) R^2 score (when trained on frailty) and (b) & (d) C-index (when trained on mortality) for aging biomarkers, comparing ElasticNet (EN, orange) and Feedforward Neural Network (FNN, blue) in (a) & (b), and the allprot (purple) and cardiometabolic (CMB, red) protein sets when training EN in (c) & (d). Distributions reflect performance on the fixed 10% independent validation set across models trained with the pre-selected hyperparameters on 100 bootstrap resamples of the combined training and test set (90% of the original UKB data on which the models were optimized).

EN and FNN models on the UKB training set to predict either the frailty index or all-cause mortality, resulting in four CMB-based biomarkers: EN-CMB-FI, EN-CMB-mort, FNN-CMB-FI and FNN-CMB-mort. The performance metrics (R^2 score or C-index) of the resulting CMB-based aging biomarkers on the independent UKB validation set were slightly, yet significantly, lower compared to the aging biomarkers trained on the full set of proteins (Fig. 7.2c and 7.2d, pFDR < 0.001). Specifically, the EN aging biomarker trained on all proteins (EN allprot) demonstrated, on average, a 0.047 (SD 0.003) higher R^2 for the frailty index prediction and a 0.015 (SD 0.005)

higher C-index for all-cause mortality prediction compared to the EN aging biomarker trained on the CMB panel (EN CMB).

In the UKB validation set, we then examined the associations per standard deviation (SD) increase in the age-accelerated biomarkers, which reflects the component of the biomarker independent of chronological age and indicates whether individuals are biologically older or younger than the population average, trained on either the allprot or CMB protein set with various age-related health indicators (see Methods, section 7.4.8 for details). These analyses were adjusted for sex, socio-economic status, smoking status, BMI, and cell counts. Figure 7.3 illustrates the associations between the aging biomarkers and their respective training targets: the frailty index and all-cause mortality. As expected, aging biomarkers generally exhibited stronger associations with the specific health indicator they were trained to predict—the frailty index for frailty index-trained biomarkers (top left, Fig. 7.3) and mortality for mortality-trained biomarkers (bottom right, Fig. 7.3). However, none of these differences were statistically significant after FDR correction (pFDR $>$ 0.10).

For frailty index-trained biomarkers, the adjusted percentage increase in the frailty index per SD in the aging biomarker was 3.94 (95% confidence interval (CI): (3.60, 4.28)) for EN-allprot-FI and 3.38 (95% CI: (3.02, 3.74)) for EN-CMB-FI; the FNN-based aging biomarkers exhibited similar associations.

Similarly, mortality-trained biomarkers demonstrated slightly lower but comparable associations with mortality. The EN-allprot-mort biomarker was associated with a hazard ratio (HR) of 1.83 (95% CI: (1.64, 2.05)), while EN-CMB-mort had an HR of 1.67 (95% CI: (1.51, 1.85)). The FNN-based biomarkers followed a similar trend.

Notably, this pattern was not observed for the alternative training outcome (i.e., frailty when predicting mortality and vice versa; top right and bottom left, Fig. 7.3a, left panel), where differences between aging biomarkers trained on allprot or CMB were even smaller (pFDR $>$ 0.99). When assessing associations with all-cause mortality, frailty index-trained biomarkers demonstrated equivalent performance when comparing the allprot and CMB protein sets. The EN-allprot-FI aging biomarker yielded a HR of 1.68 (95% CI: (1.50, 1.89)), while the EN-CMB-FI biomarker had an HR of 1.76 (95% CI: (1.57, 1.98)). The FNN-based aging biomarkers exhibited slightly lower HRs, though still comparable between the two protein sets.

The same held true in the reverse case (Fig. 7.3a, right panel), where mortality-trained biomarkers were evaluated against the frailty index. Both EN- and FNN-based biomarkers showed similar performances across protein sets; the EN-allprot-mort biomarker showed an adjusted percentage increase of 1.99 (95% CI: (1.62, 2.35)), while the EN-CMB-mort biomarker had an increase of 2.03 (95% CI: (1.66, 2.39)). Altogether, this suggests that, although using more proteins increases performance on the training target, it does not benefit associations with other age-related health indicators.

Subsequently, we evaluated the impact of data source on biomarker performance. We investigated the integration of Olink proteomics data (CMB panel only) with 107 non-derived Nightingale Health (^{1}H-NMR) metabolite markers using three approaches: (i) direct concatenation of proteins and metabolites, (ii) applying principal component analysis (PCA) separately to each data type followed by concatenation of the first ten or twenty principal components of each modality, and (iii) AJIVE, which finds a common subspace between the two data modalities (methods detailed in Supplement 7.5.2). However, none of the integration methods, whether using FNN or EN, resulted in higher aging biomarkers' performance metrics (R^2 or C-index); see Supplement 7.5.2.

7.2.4. Associations with other health outcomes

Next, we investigated to what extend the frailty index or mortality trained markers predicted relevant age-related health indicators. Fig. 7.3b demonstrates that a model trained on the frailty index predicts these health outcomes better than a model trained on mortality, with the exception of smoking. Secondly, regardless of the training method, biomarkers derived from the reduced set of proteins yield comparable associations (pFDR > 0.14).

Furthermore, for some health indicators, such as the association between frailty index-trained FNN biomarkers and self-rated poor health, the FNN-CMB biomarker showed slightly higher point estimates (odds ratio (*(OR)*) = 2.13, 95% CI: (1.85, 2.45) for FNN-allprot-FI vs. *OR* = 2.18, 95% CI: (1.90, 2.50) for FNN-CMB-FI). Similarly, for incident cardiovascular disease, the FNN-CMB-FI biomarker yielded higher point estimates. This indicates that this subset of proteins is able to retain important information with respect to broader aging factors.

7.2.5. Forward Feature Selection is an effective method in selecting highly associative proteins for an aging biomarker

Biological age markers ideally capture a range of aging-related health outcomes to indicate overall health of individuals with rather a minimum than a maximum of molecular features. We observed that the subset of CMB proteins as compared to allprot can retain equivalent associations with a range of health outcomes beyond those they were trained on. Therefore, we aimed to identify a 'minimal selection' of proteins that maintain robust associations with various age-related health indicators tested earlier. To achieve this, we applied a Forward Feature Selection (FFS) approach to identify proteins from the CMB protein set that incrementally increase goodness-of-fit metrics by more than 0.001, i.e. R^2 in the case of the frailty index (27 proteins plus chronological age) or the C-index in the case of all-cause mortality (20 proteins) in the UKB training set (see Methods for details). The final protein lists for ProtFI and ProtMort are shown in Table 7.2. Of these proteins, only the MET protein has been discussed in earlier GWAS studies [[33–36]. In total, only 6 proteins from the full CMB panel were overlapping with loci found in these GWAS studies.

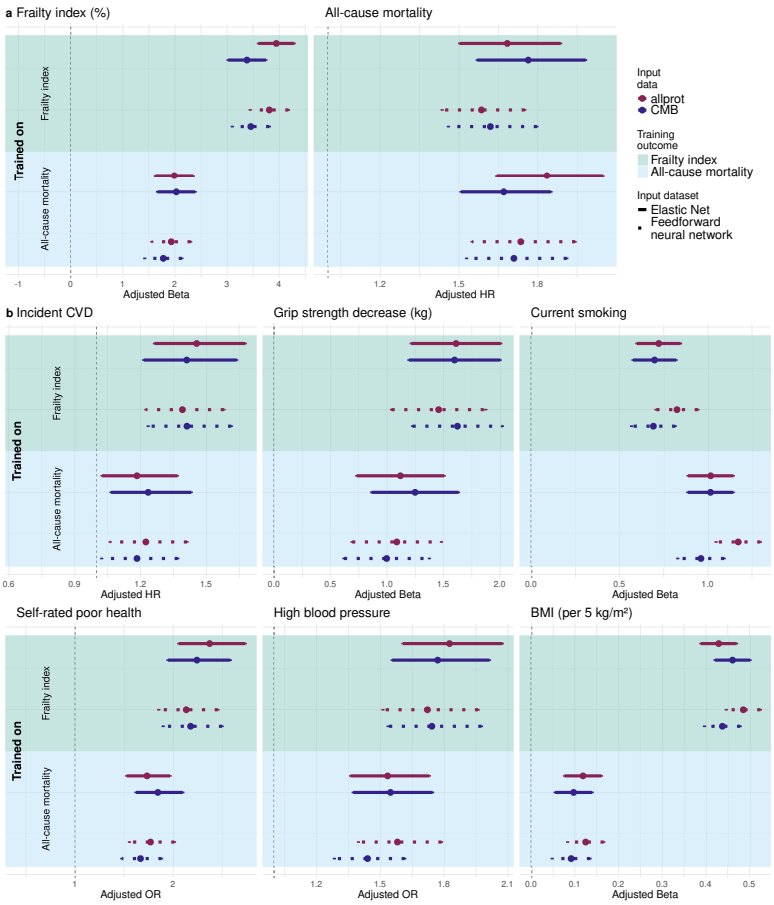


Figure 7.3.: Health indicator associations of aging biomarkers trained on allprot and CMB proteins on the UK Biobank (UKB). The figure shows associations of protein-based aging biomarkers trained using four Olink panels (allprot, red) or only the cardiometabolic panel (CMB, blue) with: (a) the training endpoints, namely the frailty index and all-cause mortality, whereas (b) other age-related health indicators. Dots represent the point estimates. Solid and dotted error bars indicate 95% confidence intervals for EN and FNN biomarkers, respectively. The background color indicates the training outcome, with green (upper) for the frailty index and blue (lower) for all-cause mortality. All performance metrics were measured in the independent UKB validation set

Using these minimal protein sets, we created aging biomarkers using ElasticNet, named ProtFI and ProtMort, predicting the frailty index and all-cause mortality, respectively. For comparative reasons, we also trained Feedforward Neural Network

Table 7.2.: The proteins selected in the Forward Feature Selection procedure, together with the accompanying beta values of the linear regressor used for selection.

Frailty index		All-cause Mortality	
Protein name	Beta value	Protein name	Beta value
ACE2	0.00566	BOC	0.147
ADGRE5	-0.00634	BPIFB1	0.142
ALCAM	0.00706	CCL27	-0.093
ART3	-0.00689	CHRDL2	0.120
CHRDL2	-0.00575	CNTN1	-0.077
COL6A3	0.00412	COL1A1	-0.111
CTSB	-0.00354	COL6A3	0.284
CTSL	0.00321	CRTAC1	-0.127
EGFR	-0.00501	DCN	-0.101
ENPP2	-0.00411	DLK1	-0.151
FAP	-0.00531	DPT	-0.095
FETUB	0.00366	GDF15	0.256
GDF15	0.00876	ICAM1	0.115
IGFBP3	-0.00430	ITGB2	-0.156
LEP	0.00927	KITLG	-0.096
LGALS3	0.00528	PCSK9	0.097
LRP11	0.00378	REN	0.090
LTBP2	0.00506	SPARCL1	-0.084
MET	-0.00345	SPP1	0.211
NRP1	0.00366	XG	-0.091
PCSK9	0.00410		
PLA2G1B	-0.00297		
PLTP	-0.00354		
REN	0.00611		
SOST	-0.00348		
TNFSF13B	0.00477		
TSHB	-0.00409		
age	-0.00486		

models on the same reduced protein sets to predict frailty (FNN-FFS-FI) and mortality (FNN-FFS-mort). We assessed the associations of the resulting FFS aging biomarkers with the frailty index, all-cause mortality, and various health indicators in the independent UKB validation set.

Reducing the number of proteins from 344 to 27 for frailty index-based models, and to 20 for all-cause mortality models, resulted in some general attenuation of associations with the respective training outcomes (Fig. 7.4a). However, these differences were not statistically significant for any aging biomarker (all pFDR > 0.86). When

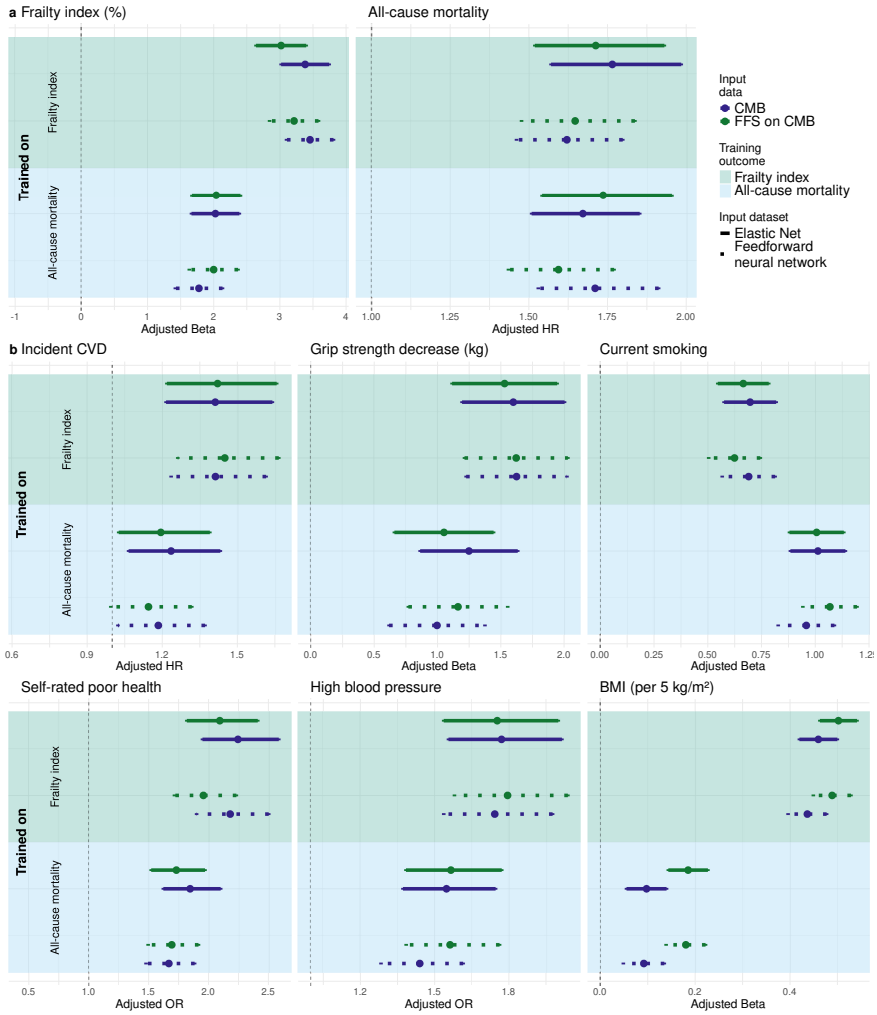


Figure 7.4.: Health indicator associations of aging biomarkers trained on CMB and FFS-selected proteins. The figure shows associations of protein-based aging biomarkers trained using the cardiometabolic (CMB) panel or the forward feature selected CMB proteins (FFS) with: (a) the training endpoints, namely the frailty index and all-cause mortality, and (b) other age-related health indicators. See Fig. 7.3 for figure elaboration. All performance metrics were measured in the independent UKB validation set

restricting the analysis to 27 proteins and age, the adjusted effect size (β) on the frailty index itself for the frailty index-trained EN biomarker decreased from 3.38 (95% CI: (3.02, 3.74)) for EN-CMB-FI to 3.02 (95% CI: (2.64, 3.39)) for ProtFI. A

similar attenuation was observed for the FNN-based aging biomarkers.

For all-cause mortality trained biomarkers, the HR for mortality prediction for ProtMort was slightly higher than for the EN-CMB-mort estimate (ProtMort: *HR* = 1.73, 95% CI: (1.54, 1.95) vs. EN-CMB-mort: *HR* = 1.67, 95% CI: (1.51, 1.85)), though this difference was not statistically significant. In contrast, for mortality-trained FNN biomarkers, a greater number of proteins resulted in a stronger association with the training outcome. This suggests that, overall, using a larger protein set enhances predictive performance for the training outcome, consistent with findings presented in Fig. 7.3.

Fig. 7.4b further compares the associations of FFS and CMB aging biomarkers with various health outcomes. The full comparison between allprot, CMB and FFS based biomarkers can be found in Supplementary Fig. 7.1. Despite being trained on fewer than one tenth of the proteins used in the CMB aging biomarkers, both EN- and FNN-trained FFS aging biomarkers showed no significant attenuation in their associations with health outcomes compared to CMB-trained biomarkers within the same modeling framework and training outcome (all pFDR $>$ 0.71). In some cases, the point estimates for the FFS aging biomarkers were even higher than those for the CMB aging biomarkers, such as for the association with incident cardiovascular disease in frailty index-trained aging biomarkers (*HR* = 1.45, 95% CI: (1.26, 1.66) for FNN-FFS-FI, vs. *HR* = 1.41, 95% CI: (1.23, 1.61) for FNN-CMB-FI) and prevalent high blood pressure in mortality-trained aging biomarkers (*OR* = 1.56, 95% CI: (1.39, 1.76) for FNN-FFS-mort vs. *OR* = 1.44, 95% CI: (1.28, 1.62) for FNN-CMB-mort). These findings further suggest that protein-based aging biomarkers trained on a reduced protein set maintain their ability to capture key aspects of the aging process, despite the substantial reduction in the number of proteins. Furthermore, in a post-hoc sensitivity analysis adjusting ProtFI for cross-sectional frailty index, the association with all-cause mortality remained stable (*HR* 1.47, 95% CI: 1.28, 1.69).

7.2.6. Comparison with published aging biomarkers

Given that biomarkers developed using less complex models with a reduced number of proteins demonstrated robust associations with aging phenotypes, we proceeded with the EN-trained FFS aging biomarkers, ProtFI (frailty index-trained) and ProtMort (mortality-trained). We then compared ProtFI and ProtMort with previously published aging biomarkers (see Methods for details). Specifically, we benchmarked ProtFI and ProtMort against the proteomics-based aging biomarker ProteinAge and ProteinAge20, trained on chronological age [37], and two mortality-trained biomarkers, the ProteinMortality Score [38], and the Proteomic Aging Clock (PAC) [39]. Additionally, we included two metabolomics-based aging biomarkers, both of which predict mortality: The Metabo Mortality Score [38] and MetaboHealth [12], which —unlike the other biomarkers— was not trained in the UK Biobank. Since previous studies identified GDF15 as a key protein associated with mortality [18], and given that it was the most important protein selected by our FFS procedures for both the frailty index and all-cause mortality, we also evaluated GDF15 alone.

Fig. 7.5 demonstrates that, despite using fewer proteins than PAC (128 proteins, derived from 2923 proteins), the ProteinMortality Score (201 proteins) or ProteinAge (204 proteins), ProtFI exhibited the strongest associations with the frailty index, incident cardiovascular disease, self-rated poor health, prevalent high blood pressure, and was most prominently associated with BMI. The differences in association between ProtFI and the next most associative model (per endpoint) remain significant for the frailty index and the association with BMI. While stronger associations for ProtFI might be expected for indicators included in the frailty index (e.g., high blood pressure and self-rated poor health), similar trends were observed for independent measures like incident cardiovascular disease and handgrip strength.

In contrast, for all-cause mortality, PAC showed the strongest association (ProtFI: HR = 1.71, 95% CI: (1.52, 1.93) vs. PAC: HR = 2.05, 95% CI: (1.85, 2.28)). Furthermore, smoking was strongest associated with mortality-trained proteomics-based biomarkers, with ProtMort showing the highest effect size (ProtMort: β = 1.00, 95% CI: (0.88, 1.13) vs. Protein Mortality Score β = 0.92, 95% CI: (0.79, 1.05); PAC: β = 0.98, 95% CI: (0.86, 1.05)).

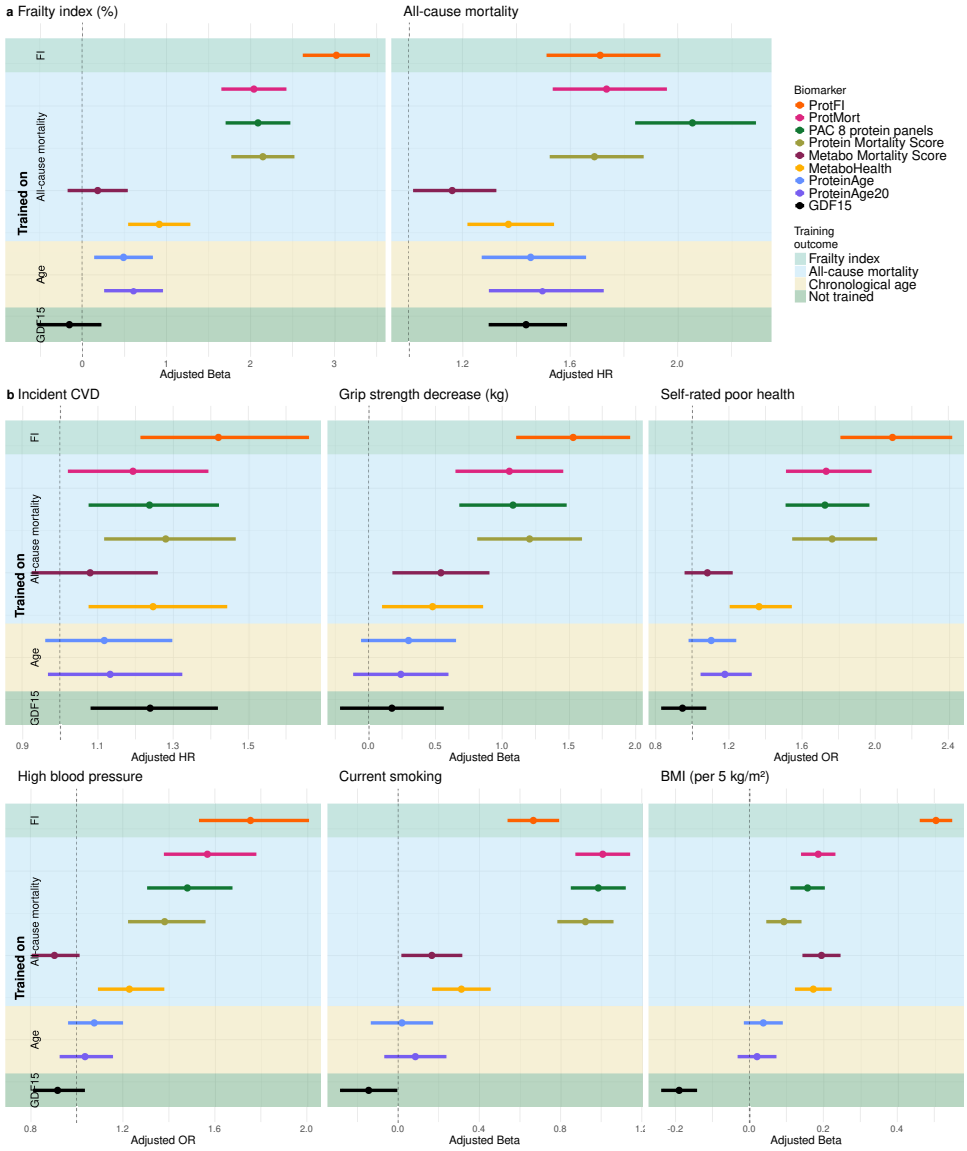


Figure 7.5.: Comparison of associations of ProtFI and ProtMort with previously published aging biomarkers. Dots represent the point estimates. Error bars are displayed to show the 95%-confidence interval. The background color indicates the training outcome, with green (upper) for the frailty index and blue (second) for all-cause mortality, yellow (third) chronological age, and dark green (bottom) GDF15, a non-trained aging biomarker. Training outcome is also displayed on the y-axis. Adjusted beta and OR are estimated in the independent UKB validation set.

7.2.7. Validation in external cohorts shows generalizability of biomarkers

We validated our aging biomarkers in two independent external cohorts, the Rotterdam Study (RS) and the Leiden Longevity Study (LLS). Given that the CMB panel was the only proteomic dataset available in these cohorts, our validation efforts focus on aging biomarkers trained on this subset of proteins, as well as ProtFI and ProtMort. External validation in these cohorts demonstrated that biomarkers trained using FFS-selected proteins exhibited effect sizes comparable to those obtained using the full CMB panel (all pFDR > 0.60), as shown in Supplementary Fig. 7.2 and Supplementary Table 7.7.

Consistent with our approach in the UKB, we also compared ProtFI and ProtMort with the metabolomics-based aging biomarker MetaboHealth. Additionally, in a subset of RS, we extended our comparisons to epigenetic biomarkers trained on chronological age [10, 40], or mortality-associated phenotypes [19], providing a broader context for evaluating their performance relative to well-established aging biomarkers.

Fig. 7.6 (Supplementary Table 7.7) presents the associations of the aging biomarkers in both the multi-omics subset of RS and LLS, while results from the full RS cohort can be found in Supplementary Table 7.8. Findings in these external validation cohorts were consistent with those in the UKB, with ProtFI demonstrating stronger associations with aging-related phenotypes compared to age-trained epigenetic biomarkers, including DNAm Horvath and DNAm Hannum. In particular, ProtFI exhibited the strongest associations with aging phenotypes regardless of whether they were self-reported or objectively measured. The only exception was the association with current smoking, which was consistent with our findings in the UKB. These results highlight the generalizability of ProtFI and ProtMort to an independent cohort and across diverse populations, reinforcing their ability to robustly reflect and predict the aging process.

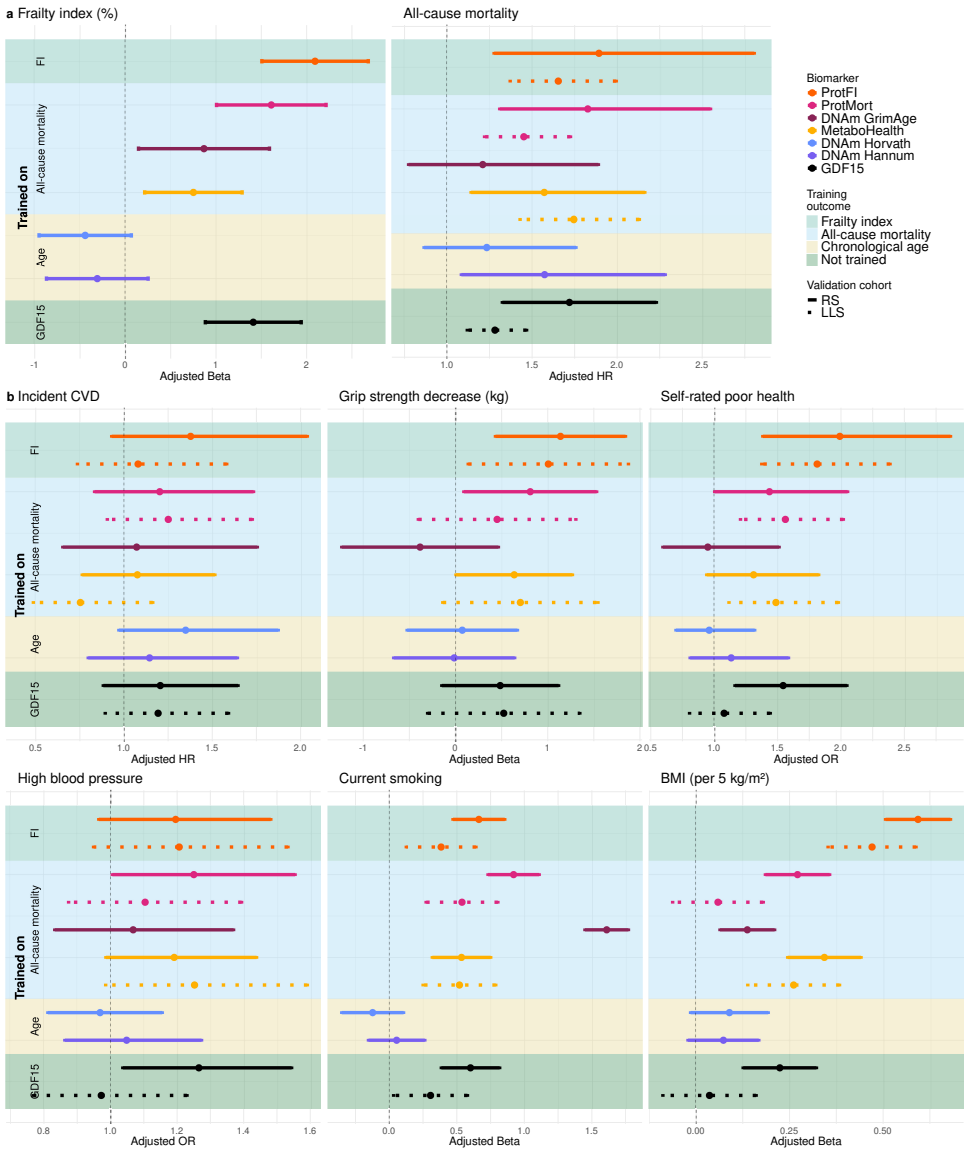


Figure 7.6.: External validation of aging biomarkers in LLS and RS with different health outcomes. The figure shows associations of different aging biomarkers. Dots represent the point estimates. Solid and dotted error bars indicate 95% confidence intervals for the Rotterdam Study (RS) and Leiden Longevity Study (LLS), respectively. The background color indicates the training outcome, with green (upper) for the frailty index and blue (second) for all-cause mortality, only in RS, yellow (third for RS) chronological age, and dark green (bottom) GDF15, a non-trained aging biomarker. Training outcome is also displayed on the y-axis.

7.3. Discussion

This study systematically compared aging biomarker development with respect to model selection, training outcome, and input data. As a result, we present ProtFI, a proteomics-based aging biomarker trained to predict the frailty index, showing promising associations with various health indicators.

We did not observe clear advantages of neural networks over simpler linear models. Furthermore, when training on either the frailty index or all-cause mortality and evaluating associations with the same outcome, biomarkers performed best when including as many proteins as possible. When assessing associations with outcomes other than the training outcome (e.g., handgrip strength or self-rated poor health), however, biomarkers trained on the frailty index outperformed those trained on mortality, suggesting that the frailty index may provide a more comprehensive representation of biological aging. Moreover, simpler biomarkers based on a subset of cardiometabolic proteins or a small number of proteins selected through forward feature selection showed broadly similar associations with health outcomes beyond the training outcomes compared to aging biomarkers trained on the full proteomics set. In fact, when predicting outcomes different from the one the biomarkers were trained on, such as using frailty-trained biomarkers to predict mortality, those based on the full set of proteins performed slightly worse than the cardiometabolic ones, suggesting that the full proteomics set biomarkers may be somewhat overtrained or capture less relevant features. We introduced two biomarkers, ProtFI and ProtMort, which are based on a minimal set of proteins —27 for ProtFI and 20 for ProtMort— from the forward feature selected protein set to predict the frailty index and mortality, respectively. ProtFI outperformed established molecular aging biomarkers including methylation, metabolomic and previously published proteomic biomarkers in association with aging phenotypes. Furthermore, the association of ProtFI with all-cause mortality outperformed those of a previously developed frailty polygenetic risk score where a standard deviation increase resulted in 8 percent additional all-cause mortality risk, compared to 71, 89, and 65 percent for ProtFI in the UK Biobank, Rotterdam Study, and Leiden Longevity Study, respectively. Importantly, after adjusting ProtFI for the cross-sectional frailty index, the association with all-cause mortality remained strong. This indicates that ProtFI captures prognostic information on the aging process beyond the conventional frailty index. Finally, both CMB- and FFS-based biomarkers were validated in independent cohorts, supporting the robustness and generalizability of their associations across populations.

Our findings reveal that within this dataset, the use of more complex models, such as neural networks, did not improve predictive performance for age-related clinical phenotypes in broad. This result is consistent with existing literature suggesting that neural networks struggle with tabular data, where traditional statistical models or simpler machine learning approaches tend to perform more effective [41]. While deep learning for tabular data remains an active area of research, with recent advances exploring the use of foundation models for such data [42], further developments are needed to assess its broader applicability in this context. Furthermore, our frailty

models were trained using a mean squared error loss function. While this is the standard choice for regression models, it does assume a gaussian distribution of the residual, which in case of frailty, is not the case, since frailty is constrained between 0 and 1. Exploring adaptations of the loss function to take this constraint into account could be an interesting avenue to augment our frailty-based models.

Training biomarkers on the frailty index resulted in superior predictive performance for most endpoints, even for those not directly related to the frailty index, such as handgrip strength and incident cardiovascular disease. This is most likely due to the broader and more nuanced nature of frailty as a training target compared to mortality. We specifically chose the frailty index as a training outcome because it can detect differences even among relatively healthy individuals, in contrast to more physical frailty measures such as the frailty phenotype, which categorizes individuals as frail or not frail per component [43–46]. However, we did not compare predictive performance on various health outcomes among biomarkers trained on different frailty measures, which would be an important direction for future research. For outcomes with a strong association to mortality, such as smoking [47, 48], biomarkers trained directly on mortality outperformed those trained on frailty. This finding underscores the importance of determining the objective of developing an aging biomarker in advance, as the choice of training target should align with the biomarker's intended function. It's important to note that differences in statistical power between mortality and frailty index-based models may also influence these results, as the number of mortality cases impacts the training power of mortality, and the high degree of censoring in the training data could affect model training and evaluation.

A similar influence of power is likely reflected in our downstream analyses: BMI was consistently associated with higher ProtFI scores across all cohorts, whereas associations with incident CVD reached significance only in the UK Biobank, and has lower effect size in LLS. This likely reflects the longer follow-up and greater number of CVD cases in UK Biobank, rather than a lack of relevance of ProtFI to cardiovascular health. Supporting this, ProtFI was associated with prevalent CVD in all cohorts, with the strongest effects in UK Biobank, partly due to the inclusion of cardiometabolic conditions in the frailty index. In contrast, BMI is consistently measured, which may explain its robust association with ProtFI.

Besides the selection of training target, our results emphasize the significance of considering the intended use of a biomarker when choosing the number of proteins to include. In both our internal validation set as well as two independent validation cohorts, increasing the number of proteins generally improved performance for the specific endpoint the biomarker was trained on, such as mortality or frailty. However, this improvement did not always lead to better associations with other health outcomes. In some cases, the point estimates of the FFS biomarkers were even higher than those of CMB. This may be explained by the targeted selection of proteins in FFS that are more consistently associated with age-related decline, which may have improved the signal-to-noise ratio by excluding proteins with weaker or more diffuse

associations.

While associations with aging related health outcomes remained or even increased when narrowing down the protein selection, the relevant goodness-of-fit metric (e.g., the concordance index or R^2 value) were lower. Moreover, the scoring of these metrics was rather low in general. However, when performing association analysis, our biomarkers compared competitively to existing markers. This highlights the point that a lower goodness-of-fit score (e.g., the concordance index or R^2 value) does not necessarily indicate that the biomarker performs poorly for other health outcomes. While these metrics are useful for model comparison, they should not be the sole criterion for evaluating a biomarker's relevance in risk assessment, particularly in population health management or clinical settings. Reducing the number of proteins, for example, may lower costs and facilitate broader implementation in the population, even if it slightly diminishes predictive accuracy in a population-based context. Future research should carefully balance the trade-off between accuracy and practicality, considering the biomarker's translational potential in real-world applications.

In our study, only one of the 43 proteins included in ProtFI and/or ProtMort showed a possible correspondence with previously reported frailty-related genetic loci (MET, Mak et al., [34]). This limited overlap is expected, as genetic associations reflect inherited predisposition, whereas circulating proteins capture downstream biological processes shaped by environmental exposures, comorbidities, and aging. Given the modest heritability of frailty and the restricted proteomic coverage in our replication panel (344 proteins), substantial overlap with GWAS findings is unlikely. Moreover, our feature-selection approach optimizes predictive performance rather than biological completeness, leading to the exclusion of highly correlated or redundant proteins. Thus, the low overlap with genetic loci does not diminish the relevance of the identified proteins; instead, it suggests that proteomic signatures provide complementary information for prediction of frailty and related mortality.

Although previous studies have shown differences between men and women when selecting proteins associated with a disease (e.g. in the integration of a biomarker PRS on the standard PRS for prediction of coronary heart disease [49]), we decided not to stratify our biomarkers by sex. This decision was based on our analyses that showed no differences in model performance when stratifying by sex. The proteins that were selected by sex-specific models were, however, largely non-overlapping. While this might suggest underlying differences in biological mechanisms, the causes for this difference are equally likely due to modeling limitations (that filters correlated proteins). We therefore decided to abstain from biological interpretation of the differences in protein selection between men and women, and focus on optimizing for predictive performance when deciding whether to stratify or not.

In our study, multi-omics biomarkers did not outperform single-omics biomarkers in terms of predictive accuracy. However, both MetaboHealth and protein-based biomarkers demonstrated independent associations with outcomes, suggesting that

integrating multiple omics layers may still offer valuable complementary insights. This inability to create multi-omics aging biomarkers that consistently outperform single-omic aging biomarkers developed on one molecular layer is consistent with previous efforts to create multi-omics aging biomarkers [38]. The added noise from integrating different molecular layers may dilute the additional information provided by these layers. Furthermore, our results align with previous work emphasizing that multi-omics approaches may enhance biological understanding rather than directly improving predictive performance [50, 51]. As the field of multi-omics integration methods is rapidly evolving, its potential to improve aging biomarker performance may change in the future.

Several limitations of the current study warrant consideration. The UKB is relatively young and healthy, resulting in lower statistical power for mortality-based biomarkers and a less frail population for frailty index prediction. This raises questions about the generalizability of these results to diseased or older populations. However, the external validation in older cohorts (RS and LLS) demonstrates the robustness of the aging biomarkers developed in this study. With the exception of MetaboHealth, all aging biomarkers we compared were developed using data from the UKB. On the other hand, RS and LLS were included in the training of MetaboHealth, accounting for 13.7 % of its training set [12]. The stronger associations of MetaboHealth in LLS and RS compared to UKB, and the reverse trend for the other biomarkers we compared, suggest that the effect sizes observed for the other biomarkers may be inflated due to overlap with their training data. Additionally, while we incorporate Olink proteomic and ^{1}H-NMR metabolomic information, we did not include other molecular layers to fully assess the robustness of our methodological findings across different molecular platforms.

Besides the cost reductions associated with using fewer proteins, selecting a minimal set of proteins may also enhance robustness. By focusing only on proteins strongly associated with the training target, we reduce the risk of overfitting to noisy or less relevant proteins, thereby improving the generalizability of the biomarker to external cohorts. Our external validation results in the RS and LLS cohorts support this hypothesis, further highlighting the benefits of using a smaller, carefully selected protein set.

In conclusion, this study presents a framework to systematically develop and compare aging biomarkers. Our approach underscores the utility of biomarkers that are designed to predict outcomes beyond those used in their training. We introduce robust aging biomarkers, ProtFI and ProtMort, which trained on only 27 and 20 proteins, demonstrate that using more proteins or more complex models does not necessarily lead to a better biomarker performance. Furthermore, our results highlight that in a rapidly expanding field, generalizability and external validation must take priority in aging biomarker development. This study provides a foundation for future research on aging biomarkers and potential applications in both population health and clinical settings.

Bibliography

[1] L. Partridge, J. Deelen and P. E. Slagboom. *Facing up to the global challenges of ageing*. 2018. DOI: 10.1038/s41586-018-0457-8.

[2] T. Niccoli and L. Partridge. *Ageing as a risk factor for disease*. Sept. 2012. DOI: 10.1016/j.cub.2012.07.024.

[3] D. J. Lowsky, S. J. Olshansky, J. Bhattacharya and D. P. Goldman. 'Heterogeneity in Healthy Aging'. In: *The Journals of Gerontology: Series A* 69 (6 June 2014), pp. 640–649. ISSN: 1079-5006. DOI: 10.1093/gerona/glt162. URL: https://doi.org/10.1093/gerona/glt162.

[4] J. Rutledge, H. Oh and T. Wyss-Coray. 'Measuring biological age using omics data'. In: *Nat Rev Genet* 23 (2022), pp. 715–727. DOI: 10.1038/s41576-022-00511-7. URL: www.nature.com/nrg.

[5] J. Félix, I. Martínez de Toda, E. Díaz-Del Cerro, M. González-Sánchez and M. De la Fuente. 'Frailty and biological age. Which best describes our aging and longevity?' In: *Molecular Aspects of Medicine* 98 (2024), p. 101291. ISSN: 0098-2997. DOI: https://doi.org/10.1016/j.mam.2024.101291. URL: https://www.sciencedirect.com/science/article/pii/S0098299724000505.

[6] A. Clegg, J. Young, S. Iliffe, M. O. Rikkert and K. Rockwood. 'Frailty in elderly people'. In: *The Lancet* 381 (9868 Mar. 2013), pp. 752–762. ISSN: 1474547X. DOI: 10.1016/S0140-6736(12)62167-9.

[7] E. Dent, P. Kowal and E. O. Hoogendijk. 'Frailty measurement in research and clinical practice: A review'. In: *European Journal of Internal Medicine* 31 (2016), pp. 3–10. ISSN: 0953-6205. DOI: https://doi.org/10.1016/j.ejim.2016.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0953620516300279.

[8] M. Moqri, C. Herzog, J. R. Poganik, J. Justice, D. W. Belsky, A. Higgins-Chen, A. Moskalev, G. Fuellen, A. A. Cohen, I. Bautmans, M. Widschwendter, J. Ding, A. Fleming, J. Mannick, J. D. J. Han, A. Zhavoronkov, N. Barzilai, M. Kaeberlein, S. Cummings, B. K. Kennedy, L. Ferrucci, S. Horvath, E. Verdin, A. B. Maier, M. P. Snyder, V. Sebastiano and V. N. Gladyshev. *Biomarkers of aging for the identification and evaluation of longevity interventions*. 2023. DOI: 10.1016/j.cell.2023.08.003.

[9] D. M. Williams, J. Jylhävä, N. L. Pedersen and S. Hägg. 'A Frailty Index for UK Biobank Participants'. In: *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 74 (4 Mar. 2019), pp. 582–587. ISSN: 1758535X. DOI: 10.1093/gerona/gly094.

[10] S. Horvath. 'DNA methylation age of human tissues and cell types'. In: *Genome Biology* 14 (3156 2013). DOI: 10.1145/2820783.2820789. URL: https://doi.org/10.1186/gb-2013-14-10-r115.

[11] S. Horvath and K. Raj. 'DNA methylation-based biomarkers and the epigenetic clock theory of ageing'. In: *Nature Reviews Genetics* 19 (6 June 2018), pp. 371–384. ISSN: 14710064. DOI: 10.1038/s41576-018-0004-3.

[12] J. Deelen, J. Kettunen, K. Fischer, A. van der Spek, S. Trompet, G. Kastenmüller, A. Boyd, J. Zierer, E. B. van den Akker, M. Ala-Korpela, N. Amin, A. Demirkan, M. Ghanbari, D. van Heemst, M. A. Ikram, J. B. van Klinken, S. P. Mooijaart, A. Peters, V. Salomaa, N. Sattar, T. D. Spector, H. Tiemeier, A. Verhoeven, M. Waldenberger, P. Würtz, G. D. Smith, A. Metspalu, M. Perola, C. Menni, J. M. Geleijnse, F. Drenos, M. Beekman, J. W. Jukema, C. M. van Duijn and P. E. Slagboom. 'A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals'. In: *Nature Communications* 10 (1 2019), pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-019-11311-9.

[13] M. J. Peters *et al.* 'The transcriptional landscape of age in human peripheral blood'. In: *Nature Communications* 6 (Oct. 2015). ISSN: 20411723. DOI: 10.1038/ncomms9570.

[14] T. Tanaka, A. Biancotto, R. Moaddel, A. Z. Moore, M. Gonzalez-Freire, M. A. Aon, J. Candia, P. Zhang, F. Cheung, G. Fantoni, R. D. Semba and L. Ferrucci. 'Plasma proteomic signature of age in healthy humans'. In: *Aging Cell* 17 (5 2018). ISSN: 14749726. DOI: 10.1111/acel.12799.

[15] F. Galkin, P. Mamoshina, A. Aliper, E. Putin, V. Moskalev, V. N. Gladyshev and A. Zhavoronkov. 'Human Gut Microbiome Aging Clock Based on Taxonomic Profiling and Deep Learning'. In: *iScience* 23 (6 June 2020). ISSN: 25890042. DOI: 10.1016/j.isci.2020.101199.

[16] P. Würtz, A. J. Kangas, P. Soininen, D. A. Lawlor, G. D. Smith and M. Ala-Korpela. 'Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies'. In: *American Journal of Epidemiology* 186 (9 2017), pp. 1084–1096. ISSN: 14766256. DOI: 10.1093/aje/kwx016.

[17] B. Lehallier, D. Gate, N. Schaum, T. Nanasi, S. E. Lee, H. Yousef, P. M. Losada, D. Berdnik, A. Keller, J. Verghese, S. Sathyan, C. Franceschi, S. Milman, N. Barzilai and T. Wyss-Coray. 'Undulating changes in human plasma proteome profiles across the lifespan'. In: *Nature Medicine* 25 (12 Dec. 2019), pp. 1843–1850. ISSN: 1546170X. DOI: 10.1038/s41591-019-0673-2.

[18] T. Eiriksdottir, S. Ardal, B. A. Jonsson, S. H. Lund, E. V. Ivarsdottir, K. Norland, E. Ferkingstad, H. Stefansson, I. Jonsdottir, H. Holm, T. Rafnar, J. Saemundsdottir, G. L. Norddahl, G. Thorgeirsson, D. F. Gudbjartsson, P. Sulem, U. Thorsteinsdottir, K. Stefansson and M. O. Ulfarsson. 'Predicting the probability of death using proteomics'. In: *Communications Biology* 4 (1 2021), p. 758. ISSN:

2399-3642. DOI: 10.1038/s42003-021-02289-6. URL: https://doi.org/10.1038/s42003-021-02289-6.

[19] A. T. Lu, A. Quach, J. G. Wilson, A. P. Reiner, A. Aviv, K. Raj, L. Hou, A. A. Baccarelli, Y. Li, J. D. Stewart, E. A. Whitsel, T. L. Assimes, L. Ferrucci and S. Horvath. 'DNA methylation GrimAge strongly predicts lifespan and healthspan'. In: *Aging (Albany NY)* 11 (2 2019), pp. 303–327. ISSN: 19454589. DOI: 10.18632/aging.101684.

[20] C. McCrory, G. Fiorito, B. Hernandez, S. Polidoro, A. M. O'Halloran, A. Hever, C. N. Cheallaigh, A. T. Lu, S. Horvath, P. Vineis and R. A. Kenny. 'GrimAge Outperforms Other Epigenetic Clocks in the Prediction of Age-Related Clinical Phenotypes and All-Cause Mortality'. In: *Journals of Gerontology - Series A Biological Sciences and Medical Sciences* 76 (5 May 2021), pp. 741–749. ISSN: 1758535X. DOI: 10.1093/gerona/glaa286.

[21] L. M. Kuiper, H. A. Polinder-Bos, D. Bizzarri, D. Vojinovic, C. L. Vallerga, M. Beekman, M. E. T. Dollé, M. Ghanbari, T. Voortman, M. J. T. Reinders, W. M. M. Verschuren, P. E. Slagboom, E. B. van den Akker and J. B. J. van Meurs. 'Epigenetic and Metabolomic Biomarkers for Biological Age: A Comparative Analysis of Mortality and Frailty Risk'. In: *The Journals of Gerontology: Series A* 78 (10 June 2023), pp. 1753–1762. ISSN: 1758-535X. DOI: 10.1093/gerona/glad137.

[22] D. E. d. C. Leme and C. De Oliveira. 'Machine learning models to predict future frailty in community-dwelling middle-aged and older adults: the ELSA cohort study'. In: *The Journals of Gerontology: Series A* 78.11 (2023), pp. 2176–2184.

[23] L. Yan, E. Ren, C. Guo, Y. Peng, H. Chen and W. Li. 'Development and validation of a predictive model for frailty risk in older adults with cardiovascular-metabolic comorbidities'. In: *Frontiers in Public Health* 13 (2025), p. 1561845.

[24] Y. Wu, M. Jia, C. Xiang and Y. Fang. 'Latent trajectories of frailty and risk prediction models among geriatric community dwellers: an interpretable machine learning perspective'. In: *BMC geriatrics* 22.1 (2022), p. 900.

[25] S. Sathyan, E. Ayers, T. Gao, S. Milman, N. Barzilai and J. Verghese. 'Plasma proteomic profile of frailty'. In: *Aging Cell* 19.9 (2020), e13193. DOI: https://doi.org/10.1111/acel.13193. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/acel.13193.

[26] X. Jia, W. Gao, J. Zhang, Y. Zhao, L. Zhang, X. Cao, Y. Wu, L. Ma, L. Chen, L. Sun, H. Guo, C. Zhang, J. Jylhävä, Z. Hu, H. Hagelin, S. Hägg, E. O. Hoogendijk and Z. Liu. 'Plasma Proteomic Signature of Frailty in 50,506 Adults'. In: *medRxiv* (2025). DOI: 10.1101/2025.06.13.25329572. URL: https://www.medrxiv.org/content/early/2025/06/14/2025.06.13.25329572.

[27] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman and R. Collins. 'UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age'. In: *PLOS Medicine* 12 (3 Mar. 2015), e1001779. DOI: 10.1371/journal.pmed.1001779.

[28] B. B. Sun, J. Chiou, M. Traylor, C. Benner, Y.-H. Hsu, T. G. Richardson, P. Surendran, A. Mahajan, C. Robins, S. G. Vasquez-Grinnell, L. Hou, E. M. Kvikstad, O. S. Burren, J. Davitte, K. L. Ferber, C. E. Gillies, Å. K. Hedman, S. Hu, T. Lin, R. Mikkilineni, R. K. Pendergrass, C. Pickering, B. Prins, D. Baird, C.-Y. Chen, L. D. Ward, A. M. Deaton, S. Welsh, C. M. Willis, N. Lehner, M. Arnold, M. A. Wörheide, K. Suhre, G. Kastenmüller, A. Sethi, M. Cule, A. Raj, H. M. Kang, L. Burkitt-Gray, E. Melamud, M. H. Black, E. B. Fauman, J. M. M. Howson, H. M. Kang, M. I. McCarthy, P. Nioi, S. Petrovski, R. A. Scott, E. N. Smith, S. Szalma, D. M. Waterworth, L. J. Mitnaul, J. Szustakowski, B. W. Gibson, M. R. Miller, C. D. Whelan, Alnylam Human Genetics, AstraZeneca Genomics Initiative, Biogen Biobank Team, Bristol Myers Squibb, Genentech Human Genetics, GlaxoSmithKline Genomic Sciences, Pfizer Integrative Biology, Population Analytics of Janssen Data Sciences and Regeneron Genetics Center. 'Plasma proteomic associations with genetics and health in the UK Biobank'. In: *Nature* 622.7982 (Oct. 2023), pp. 329–338. DOI: 10.1038/s41586-023-06592-6.

[29] A. B. Mitnitski, A. J. Mogilner and K. Rockwood. 'Accumulation of deficits as a proxy measure of aging.' In: *TheScientificWorldJournal* 1 (2001), pp. 323–336. ISSN: 1537744X. DOI: 10.1100/tsw.2001.58.

[30] M. A. Ikram, B. C. Kieboom, W. P. Brouwer, G. Brusselle, L. Chaker, M. Ghanbari, A. Goedegebure, M. K. Ikram, M. Kavousi, R. J. de Knegt, A. I. Luik, J. van Meurs, L. M. Pardo, F. Rivadeneira, F. J. van Rooij, M. W. Vernooij, T. Voortman and N. Terzikhan. 'The Rotterdam Study. Design update and major findings between 2020 and 2024'. In: *European Journal of Epidemiology* 39 (2 2024). ISSN: 15737284. DOI: 10.1007/s10654-023-01094-1.

[31] M. Schoenmaker, A. J. de Craen, P. H. de Meijer, M. Beekman, G. J. Blauw, P. E. Slagboom and R. G. Westendorp. 'Evidence of genetic enrichment for exceptional survival using a family approach: The Leiden Longevity Study'. In: *European Journal of Human Genetics* 14 (1 2006), pp. 79–84. ISSN: 10184813. DOI: 10.1038/sj.ejhg.5201508.

[32] A. Fry, T. J. Littlejohns, C. Sudlow, N. Doherty, L. Adamska, T. Sprosen, R. Collins and N. E. Allen. 'Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population'. In: *American Journal of Epidemiology* 186 (9 Nov. 2017), pp. 1026–1034. ISSN: 0002-9262. DOI: 10.1093/aje/kwx246.

[33] J. L. Atkins, J. Jylhävä, N. L. Pedersen, P. K. Magnusson, Y. Lu, Y. Wang, S. Hägg, D. Melzer, D. M. Williams and L. C. Pilling. 'A genome-wide association study of the frailty index highlights brain pathways in ageing'. In: *Aging cell* 20.9 (2021), e13459.

[34] J. K. Mak, C. Qin, M. Krüger, A. Kuukka, S. Hägg, J. Lin and J. Jylhävä. 'Large-scale genome-wide analyses with proteomics integration reveal novel loci and biological insights into frailty'. In: *Nature Aging* 5.8 (2025), pp. 1589–1600.

[35] D. B. Rosoff, L. A. Mavromatis, A. S. Bell, J. Wagner, J. Jung, R. E. Marioni, G. Davey Smith, S. Horvath and F. W. Lohoff. 'Multivariate genome-wide analysis of aging-related traits identifies novel loci and new drug targets for healthy aging'. In: *Nature aging* 3.8 (2023), pp. 1020–1035.

[36] Y. Ye, R. B. Noche, N. Szejko, C. P. Both, J. N. Acosta, A. C. Leasure, S. C. Brown, K. N. Sheth, T. M. Gill, H. Zhao *et al.* 'A genome-wide association study of frailty identifies significant genetic correlation with neuropsychiatric, cardiovascular, and inflammation pathways'. In: *Geroscience* 45.4 (2023), pp. 2511–2523.

[37] M. A. Argentieri, S. Xiao, D. Bennett, L. Winchester, A. J. Nevado-Holgado, U. Ghose, A. Albukhari, P. Yao, M. Mazidi, J. Lv, I. Millwood, H. Fry, R. S. Rodosthenous, J. Partanen, Z. Zheng, M. Kurki, M. J. Daly, A. Palotie, C. J. Adams, L. Li, R. Clarke, N. Amin, Z. Chen and C. M. van Duijn. 'Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations'. In: *Nature Medicine* 30 (9 2024), pp. 2450–2460. ISSN: 1546-170X. DOI: 10.1038/s41591-024-03164-7.

[38] D. A. Gadd, R. F. Hillary, Z. Kuncheva, T. Mangelis, Y. Cheng, M. Dissanayake, R. Admanit, J. Gagnon, T. Lin, K. L. Ferber, H. Runz, C. N. Foley, R. E. Marioni, B. B. Sun and B. B. Team. 'Blood protein assessment of leading incident diseases and mortality in the UK Biobank'. In: *Nature Aging* 4.7 (July 2024), pp. 939–948. ISSN: 2662-8465. DOI: 10.1038/s43587-024-00655-7.

[39] C.-L. Kuo, Z. Chen, P. Liu, L. C. Pilling, J. L. Atkins, R. H. Fortinsky, G. A. Kuchel and B. S. Diniz. 'Proteomic aging clock (PAC) predicts age-related outcomes in middle-aged and older adults'. In: *Aging Cell* 23.8 (2024), e14195. DOI: https://doi.org/10.1111/acel.14195.

[40] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.-B. Fan, Y. Gao, R. Deconde, M. CHen, I. Rajapakse, S. Friend, T. Ideker and K. Zhang. 'Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates'. In: *Molecular Cell* 49 (2 2013), pp. 359–367. DOI: 10.1016/j.molcel.2012.10.016.Genome-wide.

[41] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk and G. Kasneci. 'Deep Neural Networks and Tabular Data: A Survey'. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.6 (2024), pp. 7499–7519. DOI: 10.1109/TNNLS.2022.3229161.

[42] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister and F. Hutter. 'Accurate predictions on small data with a tabular foundation model'. In: *Nature* 637.8045 (Jan. 2025), pp. 319–326. ISSN: 1476-4687. DOI: 10.1038/s41586-024-08328-6. URL: https://doi.org/10.1038/s41586-024-08328-6.

[43] L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, J. Gottdiener, T. Seeman, R. Tracy, W. J. Kop, G. Burke and M. A. McBurnie. 'Frailty in Older Adults: Evidence for a Phenotype'. In: *The Journals of Gerontology: Series A* 56.3 (Mar. 2001), pp. M146–M157. ISSN: 1079-5006. DOI: 10.1093/gerona/56.3.M146. URL: https://doi.org/10.1093/gerona/56.3.M146.

[44] O. Theou, T. D. Brothers, A. Mitnitski and K. Rockwood. 'Operationalization of Frailty Using Eight Commonly Used Scales and Comparison of Their Ability to Predict All-Cause Mortality'. In: *Journal of the American Geriatrics Society* 61.9 (2013), pp. 1537–1551. DOI: https://doi.org/10.1111/jgs.12420. URL: https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/jgs.12420.

[45] K. Rockwood, M. Andrew and A. Mitnitski. 'A Comparison of Two Approaches to Measuring Frailty in Elderly People'. In: *The Journals of Gerontology: Series A* 62.7 (July 2007), pp. 738–743. ISSN: 1079-5006. DOI: 10.1093/gerona/62.7.738. URL: https://doi.org/10.1093/gerona/62.7.738.

[46] J. Blodgett, O. Theou, S. Kirkland, P. Andreou and K. Rockwood. 'Frailty in NHANES: Comparing the frailty index and phenotype'. In: *Archives of Gerontology and Geriatrics* 60.3 (2015), pp. 464–470. ISSN: 0167-4943. DOI: https://doi.org/10.1016/j.archger.2015.01.016. URL: https://www.sciencedirect.com/science/article/pii/S0167494315000242.

[47] M. Ezzati and A. D. Lopez. 'Estimates of global mortality attributable to smoking in 2000'. In: *The Lancet* 362.9387 (Sept. 2003), pp. 847–852. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(03)14338-3. URL: https://doi.org/10.1016/S0140-6736(03)14338-3.

[48] M. B. Reitsma *et al.* 'Spatial, temporal, and demographic patterns in prevalence of smoking tobacco use and attributable disease burden in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019'. In: *The Lancet* 397.10292 (June 2021), pp. 2337–2360. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(21)01169-7. URL: https://doi.org/10.1016/S0140-6736(21)01169-7.

[49] J. Lin, N. Mars, Y. Fu, P. Ripatti, F. study, T. Kiiskinen, T. Tukiainen, S. Ripatti and M. Pirinen. 'Integration of biomarker polygenic risk score improves prediction of coronary heart disease'. In: *Basic to Translational Science* 8.12 (2023), pp. 1489–1499.

[50] D. Wissel, D. Rowson and V. Boeva. 'Systematic comparison of multi-omics survival models reveals a widespread lack of noise resistance'. In: *Cell Reports Methods* 3 (4 2023). ISSN: 26672375. DOI: 10.1016/j.crmeth.2023.100461.

[51] Y. Li, T. Herold, U. Mansmann and R. Hornung. 'Does combining numerous data types in multi-omics data improve or hinder performance in survival prediction? Insights from a large-scale benchmark study'. In: *BMC Medical Informatics and Decision Making* 24 (1 2024), p. 244. ISSN: 1472-6947. DOI: 10.1186/s12911-024-02642-9. URL: https://doi.org/10.1186/s12911-024-02642-9.

[52] D. Bizzarri, M. J. T. Reinders, M. Beekman, P. E. Slagboom, E. B. van den Akker and on behalf of the BbmriNl. 'Technical Report: A Comprehensive Comparison between Different Quantification Versions of Nightingale Health's 1H-NMR Metabolomics Platform'. In: *Metabolites* 13.12 (2023). ISSN: 2218-1989. DOI: 10.3390/metabo13121181. URL: https://www.mdpi.com/2218-1989/13/12/1181.

[53] J. D. Schoufour, N. S. Erler, J. C. K.-d. Jong, J. C. K.-d. Jong, T. Voortman, G. Ziere, J. Lindemans, C. C. Klaver, H. Tiemeier, B. Stricker, A. M. Ikram, J. S. Laven, G. G. Brusselle, F. Rivadeneira and O. H. Franco. 'Design of a frailty index among community living middle-aged and older people: The Rotterdam study'. In: *Maturitas* 97 (2017), pp. 14–20. ISSN: 18734111. DOI: 10.1016/j.maturitas.2016.12.002.

[54] C. A. Rietveld *et al.* 'GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment'. In: *Science* 340 (6139 June 2013). doi: 10.1126/science.1235488, pp. 1467–1471. DOI: 10.1126/science.1235488. URL: https://doi.org/10.1126/science.1235488.

[55] M. J. Leening, M. Kavousi, J. Heeringa, F. J. V. Rooij, J. V.-V. Heemst, J. W. Deckers, F. U. Mattace-Raso, G. Ziere, A. Hofman, B. H. Stricker and J. C. Witteman. 'Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study'. In: *European Journal of Epidemiology* 27 (3 2012), pp. 173–185. ISSN: 03932990. DOI: 10.1007/s10654-012-9668-8.

[56] J. H. Friedman, T. Hastie and R. Tibshirani. 'Regularization Paths for Generalized Linear Models via Coordinate Descent'. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. DOI: 10.18637/jss.v033.i01. URL: https://www.jstatsoft.org/index.php/jss/article/view/v033i01.

[57] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang and Y. Kluger. 'DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network'. In: *BMC Medical Research Methodology* 18.1 (Feb. 2018), p. 24. ISSN: 1471-2288. DOI: 10.1186/s12874-018-0482-1. URL: https://doi.org/10.1186/s12874-018-0482-1.

[58] Y. Benjamini and Y. Hochberg. 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1 1995), pp. 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[59] D. Bizzarri, M. J. Reinders, M. Beekman, P. E. Slagboom and E. B. V. D. Akker. 'MiMIR: R-shiny application to infer risk factors and endpoints from Nightingale Health's 1H-NMR metabolomics data'. In: *Bioinformatics* 38 (15 Aug. 2022), pp. 3847–3849. ISSN: 14602059. DOI: 10.1093/bioinformatics/btac388.

[60] D. Pelegí-Sisó, P. de Prado, J. Ronkainen, M. Bustamante and J. R. González. 'methylclock: a Bioconductor package to estimate DNA methylation age'. In: *Bioinformatics* 37 (12 June 2021), pp. 1759–1760. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btaa825. URL: https://doi.org/10.1093/bioinformatics/btaa825.

[61] A. E. Teschendorff and S. Horvath. 'Epigenetic ageing clocks: statistical methods and emerging computational challenges'. In: *Nature Reviews Genetics* 26.5 (May 2025), pp. 350–368. ISSN: 1471-0064. DOI: 10.1038/s41576-024-00807-w. URL: https://doi.org/10.1038/s41576-024-00807-w.

[62] S. D. Searle, A. Mitnitski, E. A. Gahbauer, T. M. Gill and K. Rockwood. 'A standard procedure for creating a frailty index'. In: *BMC Geriatrics* 8.1 (Sept. 2008), p. 24. ISSN: 1471-2318. DOI: 10.1186/1471-2318-8-24. URL: https://doi.org/10.1186/1471-2318-8-24.

[63] E. F. Lock, K. A. Hoadley, J. S. Marron and A. B. Nobel. 'Joint and individual variation explained (JIVE) for integrated analysis of multiple data types'. In: *Annals of Applied Statistics* 7 (1 2013). ISSN: 19326157. DOI: 10.1214/12-AOAS597.

[64] Q. Feng, M. Jiang, J. Hannig and J. Marron. 'Angle-based joint and individual variation explained'. In: *Journal of Multivariate Analysis* 166 (2018), pp. 241–265. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2018.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0047259X1730204X.

7.4. Methods

7.4.1. Study Population

UK BioBank

The UK Biobank is a prospective cohort study consisting of 500,000 participants recruited in the United Kingdom between 2006 and 2010. Invitations were sent to approximately 9.2 million individuals aged 40–69 years who lived within 25 miles (40 km) of one of 22 assessment centers across England, Wales, and Scotland, with 5.5% taking part in the baseline assessment [27]. We limited our UKB sample to participants with Olink Explore data available at baseline and who were randomly selected from the broader UKB population (n = 40,696). The data in the current study were accessed under application number 67864.

The UK Biobank received approval from the National Information Governance Board for Health and Social Care and the National Health Service North West Multicentre Research Ethics Committee.

Rotterdam Study

The Rotterdam Study (RS) is a prospective population-based cohort based in Ommoord, a suburb of Rotterdam, the Netherlands. The RS began in 1990 with 7,983 participants aged 55 and older and later expanded with additional cohorts: RS-II (in 2000, 3,011 participants aged 55 ≤), RS-III (in 2006, 3.932 participants aged 45 ≤), and RS-IV (in 2016, 3,005 participants aged 40 ≤). Participants underwent home interviews and standardized health assessments at baseline and follow-ups every 3 to 5 years, conducted by trained research staff in a dedicated research facility [30]. For this study, we measured the Olink Explore Cardiometabolic panel in RS-III participants who attended the research facility for the second visit, prioritizing those with the greatest overlap with metabolomics, gut microbiome, and DNA methylation data measured(n = 996), for which participants had previously been randomly selected. Despite this prioritization, not all selected participants had complete data across all molecular platforms.

The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and Sport (Population Screening Act WBO, license number 1071272-159521-PG). The Rotterdam Study Personal Registration Data collection is filed with the Erasmus MC Data Protection Officer under registration number EMC1712001. All participants provided written informed consent to participate in the study and to have their information obtained from their treating physicians.

Leiden Longevity Study

The Leiden Longevity Study (LLS), conducted between 2002 and 2006, includes 420 Dutch Caucasian families with at least two long-lived full siblings (≥89 years for men, ≥91 years for women) who were willing to participate. The study enrolled 991 siblings and cousins (first generation), as well as 1,671 of their children and 744 of their children's partners (second generation) [31]. For this study, we included a

subset of 500 second-generation participants and their partners who attended the third measurement of whom NMR-metabolomics was previously measured, and in whom the Olink Explore Cardiometabolic panel was assessed.

The Leiden Longevity Study protocol was approved by the Medical Ethical Committee of the Leiden University Medical Center before the start of the study (P09.140) at August 5^{th}, 2009. The first participant was enrolled at August 24^{th}, 2009. In accordance with the Declaration of Helsinki, the Leiden Longevity Study obtained informed consent from all participants prior to their entering the study.

7.4.2. Data preprocessing

Proteomic profiling

The Olink Explore assay (Olink, Uppsala, Sweden), based on the proximity extension assay technique, was used to quantify plasma levels of 2,923 proteins in the UK Biobank and 384 proteins in the Rotterdam Study and Leiden Longevity Study. The Olink Explore 3072 platform integrates eight Olink panels—Cardiometabolic, Cardiometabolic II, Inflammation, Inflammation II, Neurology, Neurology II, Oncology, and Oncology II. Based on data availability at the time of analysis and our overarching aim to maximize biomarker generalizability, we used the non-II panels for biomarker development (total of 1463 proteins before quality control). The II-panels were only included in the calculation of previously published aging biomarkers [37, 39] that incorporated proteins from these panels. In the Rotterdam Study and Leiden Longevity Study, only the Cardiometabolic panel was measured. Previous proteomic studies in the UK Biobank have described the profiling conducted in that cohort [28].

In the Rotterdam Study and Leiden Longevity Study, proteomic analysis on EDTA plasma was performed by the Genomics Core Facility (www.genomicserasmusmc.nl) in Rotterdam following the manufacturer's protocol. Data are presented as NPX (Normalized Protein eXpression) values, Olink's relative protein quantification unit on a log2 scale. NPX values are calculated from matched counts using Next-Generation Sequencing (NGS) as the readout, providing a standardized measure of protein abundance. To enhance the comparability of NPX values across cohorts, all protein measurements were Z-transformed.

Quality Control

Quality control for the UKB proteomics data was done by UKB [28]. For the Rotterdam Study and Leiden Longevity Study proteomics data, the quality control followed the standard manufacturer's procedure. In short, three internal controls are added to each sample, the Incubation control, the Extension Control, and the Amplification control. The Extension Control is used for the generation of the NPX values. The Incubation Control and the Amplification Control are used to monitor the quality of assay performance, as well as the quality of individual samples. Three external controls are included in each run, Plate Control (healthy pooled plasma), Sample Control (healthy pooled plasma), and Negative Control. The Plate Control is used for

data normalization, the Sample Control is used to assess potential variation between runs and plates, and the Negative Control is used to calculate the Limit of Detection for each assay and to assess potential contamination of assays. To pass the quality control, a threshold of an average of 500 reads per combination of sample and protein was used. Additionally, for each protein per plate, the deviation of the median of the Negative Controls must be ≤ 5 standard deviations. One RS-participant failed quality control and was excluded from all analyses.

In the UK Biobank, there were after quality control by the cohort 52,704 participants with olink protein information available. We excluded participants with >2% missingness on their proteomic information, and participants with proteomic information from multiple visits. This resulted in 40,696 particiants whose proteomics data was used. Subsequently, we excluded proteins with >2% missing values, resulting in the removal of 36 proteins, leaving 1428 proteins for analysis. Further missing proteins were imputed using K-nearest-neighbor imputation, using the KNNImputer function of the SKLearn package, with k = 5.

The same procedure was followed in both external cohorts. In both RS and LLS, no participants were excluded due to missingness. However, three proteins that were not included in either ProtFI or ProtMort, namely BMP6, EPHX2, and PGLYRP1, failed the quality control in both studies. These proteins have been imputed by the proteins with the highest correlations to these proteins in the UK Biobank, respectively PDGFA, GSTA1, and RETN (see supplementary table 7.5 for a full list of highest correlating proteins in the UK Biobank).

Profiling other molecular layers
Metabolomic profiling of EDTA plasma was performed using the Nightingale platform. In the UK Biobank, the Nightingale 2020 assay was measured, whereas the Rotterdam Study and the Leiden Longevity Study used the 2016 assay, which was subsequently re-quantified to align with the Nightingale 2020 assay [52]. Additionally, DNA methylation patterns were measured in the Rotterdam Study, with further details provided elsewhere [21].

Data splits
In the UK Biobank, we created a 70/20/10 split for all participants, representing the train/test/validation split (n = 28,487/8,138/4,071) for all protein-based biomarkers. We also created a similar split for the subgroup of participants who had both proteomics and metabolomics information available (n = 15,060/4,321/2,053 for train/test/validation splits). We sampled the multi-omic set in such a way that each of the splits for the multi-omics dataset contains only participants of its respective split in the protein-only dataset (i.e. the 15,060 samples in the multi-omics training set are all also within the 28,487 samples of the proteomics-only training set); See Supplementary Figure 7.3 for details. We did not stratify on sex, as preliminary experiments (as well as earlier work) [37] show no significant differences between sexes (Supplementary Table 7.3).

7.4.3. Ascertainment of covariates and outcomes

Ascertainment of primary outcomes

The Rockwood frailty index (FI), an accumulation measure of age-related deficits [29], was calculated using the previously established methods [9, 53]. The FI consisted of 49 deficits in the UK Biobank and 38 in the Rotterdam Study. The FI was calculated if a person had information on at least 20 deficits before multi-chain imputation as the number of deficits/total number of deficits in that cohort.

All-cause mortality was defined as participants who died from any cause during the total follow-up period, which concluded on the 30^{th} of November 2022 in the UK Biobank, the 5^{th} of July 2024 in the Rotterdam Study, and 8^{th} of April 2024 in the Leiden Longevity Study.

Ascertainment of covariates

Chronological age was calculated as the time between birth and blood sampling. In the UK Biobank, the exact date of birth was not available to researchers; therefore, we used the age at assessment as provided by the UK Biobank. Socioeconomic status was defined as the highest attained education following the UNESCO classification [54]. Current smoking status was defined as a binary variable (0 = non-smoker, 1 = current smoker) indicating active tobacco use. In all cohorts, weight and length were measured at the research center. Subsequently, body mass index (BMI) was calculated using the established formula weight(kg)/[length(m)]^2. In the Rotterdam Study and UK Biobank, information on cell counts was also available and measured using the percentages of monocytes and lymphocytes.

Ascertainment of secondary outcomes

Having high blood pressure (no = 0, yes = 1) was defined in the UK Biobank by answering "High blood pressure" as one of the answers to the touch screen question: "Has a doctor ever told you that you have had any of the following conditions? (You can select more than one answer)". In the Rotterdam Study and Leiden Longevity Study, high blood pressure was defined as a resting blood pressure exceeding 140/90 mmHg or the use of blood pressure lowering medication. Prevalent cardiovascular disease was defined as a diagnosis prior to blood sampling with a disease with ICD-10 code I20-I25, I48, I50, I60-64, I70, I73, or I48 in the UK Biobank; in the Leiden Longevity Study as I20, I21, or I60–I69, and the Rotterdam Study classification has been described more in depth elsewhere [55]. Incident cardiovascular disease was defined as a first diagnosis with any of these ICD-10 codes after blood sampling and before the end of follow-up: 31^{st} November 2020 in the UK Biobank, 1^{st} January 2020 in the Rotterdam Study, and 1^{st} November 2021 in the Leiden Longevity Study. Participants with prevalent cardiovascular disease were excluded from the incident cardiovascular disease analyses. Maximum hand grip strength was defined as the strength measured using a Jamar hand dynamometer across the two arms in the UK Biobank, the dominant hand in the Leiden Longevity Study, and the non-dominant hand in the Rotterdam Study. Self-reported poor health (no = 0, yes = 1) in the UK Biobank was defined as answering "Poor" or "Fair" to the question "In general

how would you rate your overall health?" in the UK Biobank. When a person in the Rotterdam study answered "worse" on the question "How do you rate your health at this moment, compared to your peers" they were seen as having self-rated poor health. In the Leiden Longevity study, answering "Poor" or "Very poor" to "How would you rate your health in general?" was determined as self rated poor health.

7.4.4. Biomarker models

Linear

ElasticNet regression models were used to construct the linear models. This approach combines Lasso and Ridge regression, allowing for both variable selection and regularization. The loss function that is minimized is:

$$L_{en} = L_{model} + \alpha(\beta * ||w||_1 + (1-\beta) * ||w||_2) \tag{7.1}$$

where α and β are regularization hyperparameters used to weigh the L_1 and L_2 loss, and L_{model} depends on whether the model is trained on the frailty index (L^{ft}_{model}) or all-cause mortality (L^{mort}_{model}). In the case of frailty index prediction, L_{model} constitutes of the mean squared error loss:

$$L^{ft}_{model} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{7.2}$$

with N the training set size, y_i the frailty index value for participant i, and $\hat{y}_i$ the estimated frailty index, which is a linear combination of the input: $\hat{y}_i = \sum_{j=1}^{m} w_j x_j$ with m the amount of proteins, w_j the scalar pertaining to protein j, and x_j being the protein value of protein j. In the case of mortality prediction, L_{model} is defined by the cox regression loss:

$$L^{mort}_{model} = \prod_{i \in E_i = 1} \frac{exp(\hat{h})}{\sum_{j \in \{i:T_i \geq t\}} exp(\hat{h})} \tag{7.3}$$

with $\{E_i = 1\}$ the set of participants with an event, and $\{i : T_i \geq t\}$ the set of participants still at risk at time t. $\hat{h}$ is the estimated hazard, defined by the linear combination of proteins as defined by the ElasticNet model.

In the training set, the linear models for the frailty index were built using sklearn's ElasticNetCV v1.3.2. function and the mortality models were developed using sksurv's CoxnetSurvivalAnalysis v0.22.2 with age at censoring as time variable and a binary variable indicating whether the participant was dead at the end of follow-up as event variable.

For both frailty index and mortality prediction, hyperparameter optimization was conducted using 5-fold cross-validation on the training set. A grid search over the β (ranging from 0.1 to 1 in steps of 0.01) was performed. For each β, a list of α values

was automatically generated and evaluated to identify the optimal combination of β and α. For both prediction outcomes, the α values were logarithmically spaced between a data-driven maximum α_{max} and a minimum α_{min}, where α_{max} is defined as the smallest value for which all coefficients are zero, based on the input data [56]: $\alpha_{max} = \frac{max(|\langle X,y \rangle|)}{N\beta}$. We consistently applied default settings for α_{min} and α_{max}. For the frailty models (using ElasticNetCV), α_{min} corresponded to $\alpha_{max} * 10^{-3}$, whereas for the mortality models (using CoxnetSurivalAnalysis), the default α_min corresponded to $\alpha_{max} * 10^{-4}$. Subsequently, the performance of the resulting biomarker models was assessed using mean squared error for frailty index models and concordance index for mortality models. The best model was selected based on lowest error or highest concordance, respectively.

To avoid data leakage, this optimization was nested within the training set (so not touching the test or validation set). The test set was used for the selection of the final models based on performance metrics, while the validation set remained untouched during all modeling decisions. After model selection on the test set, final models were retrained on the combined train and test set (90%) using the previously selected hyperparameters in the nested training on the train set, and evaluated on the held-out validation set (10%) to report final performance and biomarker values.

Table 7.3.: Hyperparameters for the ElasticNet models

	Frailty index		All-cause mortality	
	α	β	α	β
Allprot	$2.5 * 10^{-4}$	0.99	$1.6 * 10^{-3}$	0.99
CMB	$1.5 * 10^{-4}$	0.99	$8.4 * 10^{-3}$	0.1
FFS	$2.2 * 10^{-5}$	0.99	$7.6 * 10^{-4}$	0.69

Deep Learning

The deep learning models consist of a sequence of alternating linear layers with dropout and reLU nonlinearities. Due to the difference in input dimensionality, slightly different architectures were used for the different protein sets used; See Supplementary Table 7.4 for architecture details. All architectures end in a single output neuron, which represents the estimated frailty or log-risk function when trained on the frailty index or mortality, respectively.

The same architectures were used regardless of whether the model was trained on the frailty index or mortality, except for having one extra input neuron when training on frailty index to give age as an input, to match the ElasticNet input. Training on the frailty index uses a regularized mean squared error loss:

$$L_{ft} = \frac{1}{N}\sum_{i \in N}(\hat{y}_i - y_i)^2 + \gamma||W||_2 \tag{7.4}$$

with $\hat{y}_i$ and y_i being the estimated and true frailty index scores of participant i, and N

being the number of training samples. γ is the regularization parameter, with $||W||_2$ the l_2-norm of the model parameters (i.e., weights and biases).

For training a FNN on mortality, following the approach from DeepSurv [57], the standard cox partial likelihood loss was transformed into the average negative log partial likelihood:

$$L_{mort} = -\frac{1}{N_{E=1}} \sum_{i \in \{E_i=1\}} (\hat{h} - log(\sum_{j:\{i:T_i \geq t\}} e^{\hat{h}}) + \gamma ||W||_2) \tag{7.5}$$

with $N_{E=1}$ the number of participants with an observed event, $\hat{h}$ the estimated negative log partial likelihood (i.e., the output of the neural network), $\{i : T_i \geq t\}$ being all the participants still at risk at time T_i , and $\gamma||W||_2$ being the same regularization as in equation 7.4. All models were implemented using pyTorch.

We used a gridsearch approach to optimize the batch size, learning rate and regularization (γ in equations 7.4 and 7.5) parameters. For the models trained on frailty the gridsearch spanned from 8 to 32, 0.001 to $1.0 * 10^{-5}$, and 0.01 to $1.0 * 10^{-4}$ for batch size, learning rate and γ, respectively. For the mortality models, the gridsearch was conducted from 1000 to 20000, 0.001 to $1 * 10^{-6}$, and 0.001 to $1.0 * 10^{-5}$ for batch size, learning rate and γ, respectively. Final model parameters can be found in Tables 7.3 and 7.4.

While ElasticNet models were optimized using cross-validation, this approach was infeasible for the deep learning models due to the computational complexity. Therefore, the deep learning models were optimized by picking the hyperparameter combination with the highest result on the test (20% hold-out) set (leaving the validation set untouched).

Afterwards, our 70/20/10 data split was transformed into a 90/10 data split, retaining the same 10% hold-out validation set and merging the train and test sets. The final models were trained and evaluated on the validation set with the previously determined model parameters. 10 different seeds for parameter initialization were run, and the best model was selected based on either R^2 score or C-index (for the frailty index and all-cause mortality, respectively) in the 90 % training set, which was then used for association analysis in the validation set.

Table 7.4.: Model hyperparameters for the Feedforward Neural Networks

	Frailty index			All-cause mortality		
	Batch size	Learning rate	γ	Batch size	Learning rate	γ
Allprot	16	$5.0 * 10^{-6}$	$5.0 * 10^{-3}$	1000	$5.0 * 10^{-6}$	$1.0 * 10$
CMB	22	$1.0 * 10^{-5}$	$1.0 * 10^{-3}$	2000	$5.0 * 10^{-5}$	$5.0 * 10$
FFS	17	$1.0 * 10^{-4}$	$1.0 * 10^{-4}$	4060	$6.3 * 10^{-4}$	$1.0 * 10$

7.4.5. Forward Feature Selection

For implementing the forward feature selection procedure, we used the SequentialFeatureSelector function from SKLearn v1.3.2. In short, all features are used to independently create a univariate model to estimate the frailty index (or all-cause mortality). Next, these models are tested using the R2 score (or C-index). Using a 5-fold nested cross-validation setup on the training set, the protein with the best score metric is selected as the first (most important) protein. In the subsequent rounds, all linear combinations with the previously selected protein(s) are tested using linear regression, and the best scoring proteins are selected. This process repeats until the scoring metric does not improve more than the tolerance parameter, which was set at 0.001. Note that this procedure only uses the training set, leaving the test and validation set untouched. This resulted in 27 proteins and age selected for the frailty index and 20 for mortality; see Supplementary Table 7.2 for the specific proteins.

7.4.6. Statistical model comparison

To compare the performance of different modeling choices for the aging biomarkers, we applied bootstrap resampling (100 iterations) after completing hyperparameter optimization (on the train (EN) or train and test (FNN) data). This procedure was designed solely to quantify the variability of goodness-of-fit estimates (e.g. R^2 or C-index), not to further optimize the biomarkers. In each iteration, the hyperparameters selected on the original training set were reused. We drew a bootstrap resample from the 90% combined training and test set (sampling participants with replacement) and evaluated model performance on the fixed 10% validation set. Next, to determine whether goodness-of-fit metrics differed significantly between sets, we first assessed the normality of differences using the Shapiro-Wilk test (shapiro function, SciPy v1.12.0). Since all p-values were greater than 0.05, indicating that the differences followed a normal distribution, we proceeded with a paired Z-test assuming unequal variance (ztest function, StatsModels). P-values were false discovery rate adjusted using the Benjamini-Hochberg procedure [58]. A false discovery rate p-value threshold of <0.05 was used to define statistical significance.

7.4.7. Calculation of previously developed aging biomarkers

We calculated PAC following the dedicated R-script [39]. The ProteinMortality Score and MetaboMortality Score from Gadd et al. [38] were calculated by multiplying the included proteins and metabolites by their respective weights and calculating a sum score. MetaboHealth was calculated using the MiMIR package [59]. In the Rotterdam Study, we calculated DNAm Hannum, DNAm Horvath using the methylclock R-package [60] and DNAm GrimAge using the Python scripts provided by the researchers who developed these biomarkers.

7.4.8. Association Analysis

In order to determine the component of each biomarker independent of age, we created a linear regression predicting chronological age using the biomarker score

for each biomarker [61] on the validation set. Then, we determined the residuals of each of these linear regressions, which indicate whether an individual is biologically 'older' or 'younger' than the population average for their chronological age. After calculating the residuals, to enhance comparison among aging biomarkers, we applied a Z-transformation to these residuals and used this Z-transformed residual in the subsequent analyses with end-points.

Linear regression analysis using the OLS function from statsmodels v0.14.1 with an added constant was used to assess the associations of the aging biomarkers with continuous outcomes. In case of BMI and smoking, the aging biomarkers were used as the dependent variables in this analysis, as lifestyle factors are hypothesized to influence biological aging, and aging biomarkers are intended to reflect that. Associations of aging biomarkers with binary outcome measures were assessed with logistic regression analyses using the Logit function from statsmodels v0.14.1 with an added constant. Additionally, associations of aging biomarkers with incident outcomes were assessed using Cox-Proportional Hazard models using the CoxPHFitter function from lifelines v0.29.0.

In all analyses, the first model included only sex as a covariate. In the second model, socioeconomic status was additionally added as a covariate. In the third model, next to the covariates from the previous models, BMI and smoking were added as a covariate. In the Rotterdam Study and UK Biobank, we performed analyses in a fourth model and additionally adjusted the analyses for cell counts. Next, we performed the analyses adjusting for the fourth, or in the Leiden Longevity Study third, model with additionally either GDF15 or MetaboHealth to determine whether the associations were independent of these previously established biomarkers. Z-tests were used to assess whether effect sizes between aging biomarkers differed statistically significantly. In a post-hoc sensitivity analysis, we adjusted the all-cause mortality models for cross-sectional frailty index scores to evaluate whether the selected biomarker, ProtFI, contributed prognostic information beyond that provided by frailty alone.

7.5. Supplementary information

7.5.1. Frailty index deficits

UK Biobank

Table 7.1.: The Rockwood frailty index in the UK Biobank, as developed by Williams et al. [9], following the procedure of Searle et al. [62]. The Rockwood frailty index in the UK Biobank consist of 49 items, coded according to the table above. The frailty score is calculated by summing all 49 codes and dividing by the total number of items.

Type of deficit	Item	Trait	Coding in Frailty Index
Sensory	1	Glaucoma *	Categorized 0/1
	2	Cataracts *	Categorized 0/1
	3	Hearing difficulty	Categorized 0/1 (combined yes/deaf groups as 1)
Cranial	4	Migraine *	Categorized 0/1
	5	Dental problems	Categorized 0/1 for none vs. any
Mental wellbeing	6	Self-rated health	0 – excellent; 0.25 – good; 0.5 - fair, 1 – poor
	7	Fatigue: frequency of tiredness/ lethargy in last two weeks	0 - not at all, 0.25 - several days, 0.5 - more than half, 1 - nearly every day
	8	Sleep: experience of sleeplessness/insomnia	0 - never/rarely, 0.5 - sometimes, 1 - usually
	9	Depressed feelings: frequency in last two weeks	0 – not at all, 0.5 – several days, 0.75 - more than half, 1 – nearly every day
	10	Self-described nervous personality	Categorized 0/1
	11	Severe anxiety/ panic attacks *	Categorized 0/1
	12	Common to feel loneliness	Categorized 0/1
	13	Sense of misery (ever/never)	Categorized 0/1
Infirmity	14	Infirmity: long-standing illness or disability	Categorized 0/1
	15	Falls in last year	0 - no falls, 0.5 - one fall, 1 - more than one
	16	Fractures/broken bones in last five years	Categorized 0/1
Cardiometabolic	17	Diabetes *	Categorized 0/1
	18	Myocardial infarction *	Categorized 0/1
	19	Angina *	Categorized 0/1
	20	Stroke *	Categorized 0/1
	21	High blood pressure *	Categorized 0/1
	22	Hypothyroidism *	Categorized 0/1
	23	Deep-vein thrombosis *	Categorized 0/1

	24	High cholesterol *	Categorized 0/1
Respiratory	25	Breathing: wheeze in last year	Categorized 0/1
	26	Pneumonia *	Categorized 0/1
	27	Chronic bronchitis/ emphysema *	Categorized 0/1
	28	Asthma *	Categorized 0/1
Musculoskeletal	29	Rheumatoid arthritis *	Categorized 0/1
	30	Osteoarthritis *	Categorized 0/1
	31	Gout *	Categorized 0/1
	32	Osteoporosis *	Categorized 0/1
Immunological	33	Hayfever, allergic rhinitis or eczema *	Categorized 0/1
	34	Psoriasis *	Categorized 0/1
Cancer	35	Any cancer diagnosis *	Categorized 0/1
	36	Multiple cancers diagnosed (number reported)	0 - no cancer or single cancer, 1 - multiple cancers
Pain	37	Chest pain	Categorized 0/1
	38	Head and/or neck pain	Categorized 0/1
	39	Back pain	Categorized 0/1
	40	Stomach/abdominal pain	Categorized 0/1
	41	Hip pain	Categorized 0/1
	42	Knee pain	Categorized 0/1
	43	Whole-body pain	Categorized 0/1
	44	Facial pain	Categorized 0/1
	45	Sciatica *	Categorized 0/1
Gastrointestinal	46	Gastric reflux *	Categorized 0/1
	47	Hiatus hernia *	Categorized 0/1
	48	Gall stones *	Categorized 0/1
	49	Diverticulitis *	Categorized 0/1

7.5.2. Multi-omics

We explored several methods for integrating metabolomics and proteomics for biomarker training: concatenate, principal component analysis (PCA) and then concatenate, combining with MetaboHealth and Angle-Based Joint and Individual Variation Explained (AJIVE). For the concatenate method, we simply concatenated all proteins and metabolites together. For the PCA and then concatenate method, we first took separate PCAs for the protein and metabolomics data, then concatenated the first 10 or 20 principal components. When combining with MetaboHealth, we added the MetaboHealth score[12] to all of the protein features. Finally, AJIVE is a method for identifying shared and separate components from two data modalities [63] [64]. Using AJIVE, we identified the shared and individual components of the metabolomics and proteomics data, and used those to train our biomarkers.
Supplementary Table 7.2 shows that none of the aforementioned methods increase either the R^2 score or the C-index when trained on the frailty index or mortality,

respectively, compared to the baseline using only the cardiometabolic proteins (protein-only). However, this protein-only model was trained on a larger set of participants, due to reduced metabolomics availability. When reducing the training set size for the protein-only model, some deterioration in performance can be seen, which can partially be compensated for by adding metabolites at the input (through concatenation). However, adding more complex multi-omics approaches did not seem to improve performance further.

Table 7.2.: Goodness-of-fit metrics on various multi-omics integration methods.

Dataset / method	ElasticNet		Feedforward Neural Networks	
	Frailty index (R2)	All-cause mortality (C)	Frailty index (R2)	All-cause mortality (C)
Using only cardiometabolic proteins	0.24	0.71	0.25	0.71
Cardiometabolic proteins in population subset with metabolomics	0.23	0.69	0.20	0.68
Cardiometabolic proteins and NMR-metabolites concatenated	0.24	0.69	0.23	0.69
Using AJIVE	0.17	0.61	0.14	0.62
PCA then concatenate, using 10 principal components	0.17	0.61	0.16	0.64
PCA then concatenate, using 20 principal components	0.19	0.67	0.17	0.66
Cardiometabolic proteins and MetaboHealth, concatenated	0.23	0.69	0.23	0.68

7.5.3. Full comparison between allprot, CMB and FFS

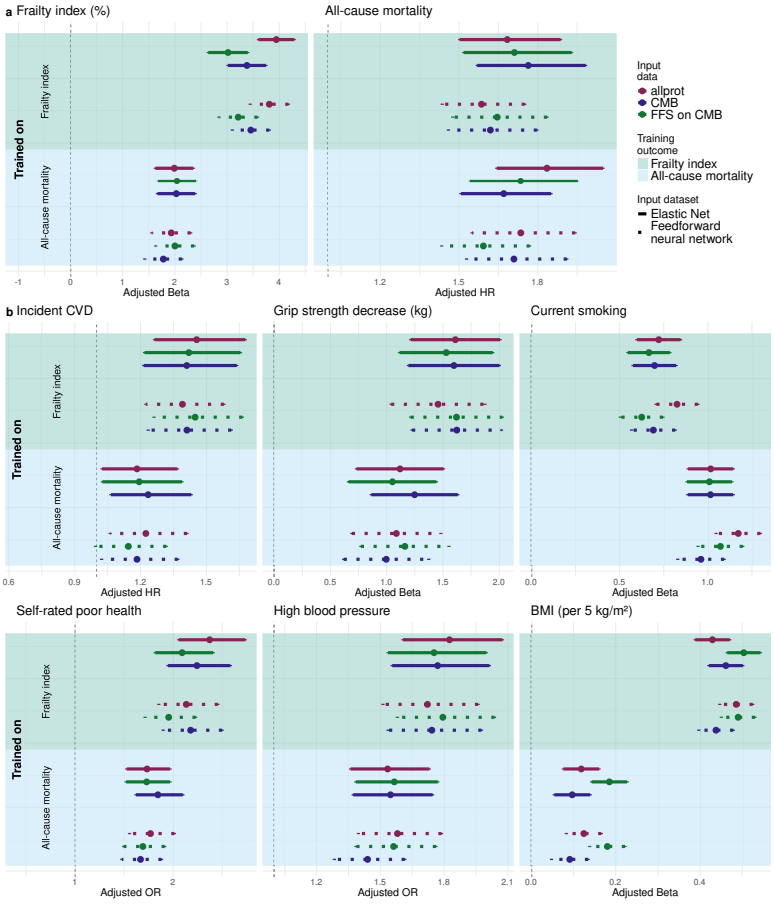


Figure 7.1.: Full comparison on the training endpoints between biomarkers trained on the full set of 1428 proteins (allprot), the subset defined by the cardiometabolic panel (CMB), and the subset of CMB defined by the forward feature selection (FFS) procedure.

7.5.4. External validation of biomarkers

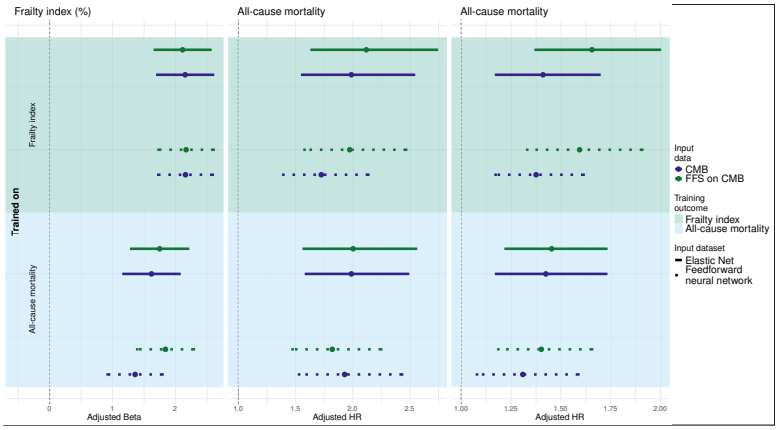


Figure 7.2.: External validation of constructed biomarkers on the Rotterdam Study (left and middle column), and the Leiden Longevity Study (right column)

7.5.5. Data splits

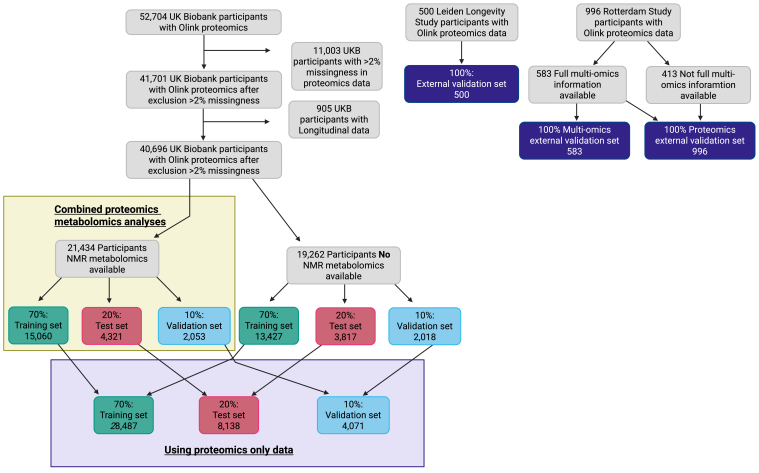


Figure 7.3.: Flowchart of data split creation on the UK Biobank, Leiden Longevity Study and the Rotterdam Study

7.5.6. Sex-specific analysis

In order to evaluate stratification by sex, we applied three approaches: a model with no adjustment for sex, a model adjusted for sex by giving it as an input covariate, and two models stratified by sex (one for men, one for women). We evaluated both the ElasticNet and deep neural network approaches, both trained to predict mortality. All these models were trained on the complete set of 1428 proteins (allprot). Supplementary Table 7.3 shows that there is no improvement in model performance when stratifying by sex, or by using sex as an input covariate. There even seems to be a slight decrease in performance, which could be attributed to the decrease in training set size of the stratified models. Therefore, in order to maximize performance, we decided not to stratify by sex in further analysis.

Table 7.3.: Concordance indexes on the four protein panel (allprot) dataset for models trained to predict mortality using various adjustments for sex.

	ElasticNet	Deep Neural Network
No adjustment	0.73	0.73
Sex as input covariate	0.73	0.73
Trained on men only	0.72	0.70
Trained on women only	0.71	0.72

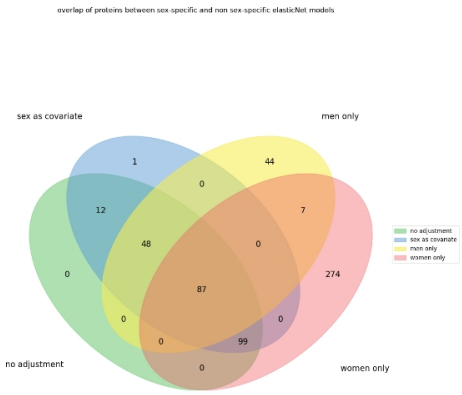


Figure 7.4.: Venn diagram showing the overlap between proteins selected by the different stratification approaches.

7.5.7. Neural network architectures

Table 7.4.: Neural network architectures

	Layer 1	Layer 2	Layer 3	Layer 4
Allprot	Linear: 1428*x 750 Dropout: p = 0.2 ReLU	Linear: 750 x 400 Dropout: p = 0.1 ReLU	Linear: 400 x 100 Dropout: p = 0.1 ReLU	Linear: 100 x 1
CMB	Linear: 344*x 200 Dropout: p = 0.2 ReLU	Linear: 200 x 100 Dropout: 0.1 ReLU	Linear: 100 x 10 ReLU	Linear: 10 x 1
FFS	Linear: 20,28**x 15 ReLU	Linear: 15 x 10 ReLU	Linear: 10 x 5 ReLU	Linear: 5 x 1

* input dimension is 1 higher when trained on the frailty index, to give age as an input
** input dimension is 20 for mortality-trained, and 28 for frailty index-trained models.

7.5.8. Additional tables

Tables 7.5 7.6, 7.7 and 7.8 can be found by scanning the following QR code:

8

Multitask learning approaches for aging biomarkers

Proteomics-based aging biomarkers have been a promising development in characterizing aging. Frailty and mortality are both meaningful endpoints for training such aging biomarkers, yet biomarkers trained on either endpoint differ in which aspects of aging they emphasize. Furthermore, frailty is a compound score, consisting of multiple groups of deficits which might have different relations with the input proteins, potentially hampering prediction from such proteins. In this chapter, we propose two biomarkers based on a multitask learning approach: One which is trained on frailty and mortality simultaneously (fmNN), and one which is trained to predict the eleven deficit categories that make up the frailty score(fgNN). Although these models do not increase in association with aging related health outcomes, these associations do seem more robust. Furthermore, our results on fgNN give insight in the driving factors behind frailty prediction from proteins.

8.1. Introduction

Frailty has emerged as a clinically meaningful phenotype for capturing heterogeneity in biological aging, as it integrates deficits across multiple physiological systems and is strongly associated with adverse health outcomes [1]. At the same time, mortality remains a key endpoint in aging research, reflecting the cumulative impact of pathological processes over the course of life [2].

Protein biomarkers derived from blood measurements offer a promising avenue to quantify frailty in a biologically grounded and minimally invasive manner [3–5]. However, while frailty-focused biomarkers tend to better reflect multidimensional functional decline, they may be less strongly associated with mortality and closely related outcomes such as smoking history [3]. Meanwhile, biomarkers optimized for mortality prediction often show strong associations with survival-related outcomes, but are less sensitive to earlier or more subtle manifestations of biological aging captured by frailty. This trade-off suggests that a unified modeling approach capable of capturing both frailty and mortality could leverage complementary information from each outcome. Such a model may better reflect the shared and distinct biological processes underlying functional decline and survival risk, and consequently yield a more robust and generalizable biomarker of aging.

An additional challenge arises from the structure of commonly used frailty measures, such as the Rockwood frailty index [6]. This index is a composite score derived from heterogeneous groups of deficits, including physical functioning, comorbidities, cognitive and mental health assessments, and self-reported health indicators. The biological pathways reflected in circulating protein levels are unlikely to relate uniformly to all of these components. Aggregating them into a single target may obscure meaningful signal specific to particular frailty domains. Modeling these domains separately, while accounting for their shared biological basis, may therefore improve predictive performance and enhance interpretability.

Multitask learning [7] provides a natural framework to address these challenges. In multitask models, multiple related outcomes are learned simultaneously using shared representations, with task-specific outputs. By jointly predicting frailty and mortality the model is encouraged to learn a latent embedding of protein abundances that captures biological features relevant across both end points. Alternatively, by predicting each frailty domain separately, the model could capture factors that are independent between these domains, thereby increasing the performance of the resulting frailty prediction.

In this work, we investigate multitask learning approaches for predicting frailty and mortality from blood-based protein measurements. We hypothesize that jointly modeling these outcomes will yield a latent representation that more faithfully captures underlying biological aging processes, resulting in improved performance compared to single-task models, both in training target prediction, as well as in associations with aging related health outcomes in a general population. Furthermore, we introduce a

multitask approach to predicting frailty, in which each group of deficits is predicted separately. We hypothesize that this approach could improve the frailty prediction, ultimately leading to a more meaningful aging biomarker.

8.2. Methods

8.2.1. Data

We use proteomics assay data from the UK Biobank. Specifically, we selected participants for who Olink Explore data was available at baseline. We then excluded participants with >2% missingness and participants with proteomics from multiple visits, resulting in 40,696 participants. Then, we excluded proteins with >2% missingness, after which we imputed the remaining missing values using the KNNImputer from SKLearn with k = 5. Finally, we selected only the proteins from the cardiometabolic panel, resulting in 344 proteins. We split the data into a train/validation/test split using a 70/20/10 ratio, resulting in three data splits of 28,487, 8,138 and 4,071 participants for the train, validation and test splits, respectively.

8.2.2. Predicting Frailty and Mortality Simultaneously

We introduce a multitask neural network that predicts frailty and mortality simultaneously, termed fmNN. fmNN consists of three hidden layers with dropout and reLU activation functions (table 7.4), and has two outputs, one whose value represents the estimated frailty ($\hat{y}$), and one whose value represents the mortality log-risk function ($\hat{h}$). The full loss function is defined as:

$$L_{ft_mort} = \beta * L_{ft} + (1-\beta) * L_{mort} + \gamma||W||_2 \tag{8.1}$$

where β is a weighing parameter that is used to emphasize either the frailty or the mortality loss, L_{ft} and L_{mort}, respectively. $\gamma||W||_2$ denotes the L2 regularization loss used to increase generalization.

The frailty loss is defined as the mean squared error between the predicted and actual frailty:

$$L_{ft} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8.2}$$

with N the batch size, y_i the actual frailty and $\hat{y}_i$ the predicted frailty. The mortality loss is defined as the average negative log partial likelihood introduced by deepSurv [8]:

$$L_{mort} = -\frac{1}{N_{E=1}} \sum_{i \in \{E_i=1\}} (\hat{h} - log(\sum_{j:\{i:T_i \geq t\}} e^{\hat{h}})) \tag{8.3}$$

with $N_{E=1}$ the number of participants who deceased before censoring, $\hat{h}$ the estimated negative log partial likelihood, which is the output of the neural network, and $\{i : T_i \geq t\}$ being all the participants still at risk at time T_i.

8.2.3. Predicting frailty groups

Williams et. al. categorized the frailty deficits for the Rockwood frailty index into 11 groups of related deficits [1], see table 8.1 for details. We created a model with 11 outputs, each representing the estimation of one of these groups (see table 8.2 for architecture details), which we termed the frailty groups neural network (fgNN). Each group is represented as the mean of all deficits in that group. The full loss is defined as the average loss over each group, which is the mean squared error between the estimated and true average for that group:

$$L_{fg} = \frac{1}{11}\frac{1}{N}\sum_{j=1}^{11}\sum_{i=1}^{N}(y_i^j - \hat{y}_i^j)^2 \tag{8.4}$$

where y_i^j is the mean of all deficits in group j of sample i, and with $\hat{y}_i^j$ the model output for this group.

Table 8.1.: Overview of frailty groups, and which specific deficits are part of which group

Group Name	# of deficits	Deficit descriptions
Sensory	3	Glaucoma; Cataracts; Hearing difficulty.
Cranial	2	Migraine; Dental problems.
Mental wellbeing	8	Self-rated health; Fatigue; Sleeplessness. Depressed feelings; Nervousness; Severe anxiety; Loneliness; Misery.
Infirmity	3	Infirmity; Falls; Fractures.
Cardiometabolic	8	Diabetes; Myocardial infarction; Angina. Stroke; High blood pressure; Hypothyroidism. Deep-vein trombosis; High cholesterol.
Respiratory	4	Breathing; Pneumonia; Chronic bronchitis. Asthma.
Musculoskeletal	4	Rheumatoid arthritis; Osteoarthritis. Gout; Osteoporosis.
Immunological	2	Hayfever, allergic rhinitis or eczema; Psoriasis.
Cancer	2	Any cancer diagnosis; Multiple cancers diagnosed.
Pain	9	Chest pain; Head and/or neck pain; Back pain. Stomach/abdominal pain; Hip pain; Knee pain; Whole-body pain; Facial pain; Sciatica.
Gastrointestinal	4	Gastric reflux; Hiatus hernia; Gall stones; Diverticulitis.

Table 8.2.: Architecture overview of the two multitask models, either created to predict frailty and mortality (fmNN), or predict the separate frailty groups (fgNN) simultaneously.

	Layer 1	Layer 2	Layer 3	Layer 4
fmNN	Linear: 344 x 200, Dropout: p = 0.2, ReLU	Linear: 200 x 100, Dropout: p = 0.1, ReLU	Linear: 100 x 10 ReLU	Linear: 10 x 2
fgNN	Linear: 344 x 250, Dropout: p = 0.2, ReLU	Linear: 250 x 150, Dropout: p = 0.1, ReLU	Linear: 150 x 100, ReLU	Linear: 100 x 11

8.2.4. Hyperparameter optimization

For both models, we use a grid search to optimize for the learning rate, batch size and l2 regularization parameter γ. Additionally, we add β to the grid search for fmNN. See tables 8.3 and 8.4 for final grid search parameters (minima, maxima and step sizes).

Table 8.3.: grid search settings for fmNN.

	learning rate	batch size	γ	β
minimum	$1 * 10^{-6}$	8	$1 * 10^{-4}$	0.01
maximum	$1 * 10^{-3}$	1000	$1 * 10^{-2}$	0.1
# of steps	5 (logarithmic)	5 (linear)	5 (logarithmic)	2 * 5 (linear between 0 and 0.01, and between 0.01 and 0.1)
selected parameters	$6 * 10^{-6}$	8	$1 * 10^{-3}$	0.055

Table 8.4.: grid search settings for fgNN.

	learning rate	batch size	γ
minimum	$1 * 10^{-5}$	8	$1 * 10^{-4}$
maximum	$1 * 10^{-3}$	256	0.18
# of steps	5 (logarithmic)	5 (linear)	5 (logarithmic)
selected parameters	$3.2 * 10^{-4}$	8	$1 * 10^{-4}$

8.2.5. Goodness-of-fit evaluation

For the fmNN model, we assess frailty prediction using the R^2 score [9], and mortality prediction using the C-index [10]. We expect a tradeoff between frailty and mortality prediction performance, based on the value of β in equation 8.1. This tradeoff results in a Pareto front of models whose performance of either outcome cannot be increased without losing performance for the other outcome. For the fgNN model,

we determine the frailty estimation from the estimations of the 11 groups. Then, we use the R^2 score to assess the goodness-of-fit of the final frailty estimation.

8.2.6. Clinical evaluations

To explore whether these multitask models contain biologically relevant information, we associate the predicted outcomes with various clinical parameters. We explored associations with frailty, mortality, BMI, incident cardiovascular disease, high blood pressure, grip strength decrease, smoking status, and self rated poor health. This is done by creating a regression predicting one of various aging related health outcomes from the model output, and then investigating the beta value corresponding to the model output. For the fmNN model, we performed independent analyses for both the frailty and the mortality output. For the fgNN model, we use the reconstructed frailty estimation for the association analyses.

Before performing association analysis, we determined the component of each model that was independent of chronological age [11]. To do this, we ran a linear regression predicting chronological age using the model output, correcting for age, sex, social economic status, smoking, BMI and cell counts as covariates. We then performed a Z-transformation on the residuals of this regression, and used the result for the subsequent association analyses.

Associations with incident outcomes (mortality and incident cardiovascular disease) were performed using cox proportional hazard models. Binary outcomes (high blood pressure and self rated poor health) were assessed using logistic regression. All other endpoints were analyzed using linear regression. Finally, for BMI and smoking status, the transformed model residuals were used as the dependent variables.

8.2.7. Baseline methods

For comparison of performance metrics, we use neural networks with a single output as baseline. These models have the same architecture as fmNN, except that the last layer reduces to a single neuron instead of two. They are trained using the same losses defined by L_{mort} and L_{ft} as defined in equation 8.3 and 8.2 respectively, with an added L2 regularization. For the association analysis, we compare with ProtFI and ProtMort [12]. These models are based on elasticNet models which are trained on a selection of proteins from the cardiometabolic panel which was used for fgNN and fmNN. See chapter 7 for more details.

8.3. Results

We introduce two multitask models, both trained on the Olink cardiometabolic (CMB) platform: One to predict frailty and mortality simultaneously (fmNN) an one to predict frailty groups separately (fgNN).

8.3.1. Predicting frailty and mortality simultaneously leads to a performance tradeoff between tasks

For fmNN, we compare an optimal set of hyperparameter configurations based on the frailty and mortality performance, measured by the R^2 score and C-index, respectively. This set of configurations results in a pareto front of models, whose performance can not be improved on both metrics simultaneously. Figure 8.1 shows that this front is dependent on β, albeit with outliers, with a higher β leading to a better prediction for mortality at the cost of the frailty prediction, and vice versa. However, there is no model that outperforms the best single-task models for both tasks simultaneously. Nevertheless, the best performing models for both frailty and mortality do outperform their respective baselines.

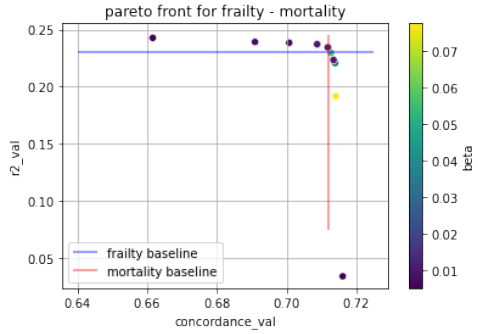


Figure 8.1.: Pareto front of best performing frailty-mortality multitask models. The horizontal and vertical lines show baseline performances of single task neural networks aimed at predicting frailty and mortality, respectively. Each dot is a different model instance with its own hyperparameters, colored by its value for β. The model that was selected for further analysis is in green (β = 0.05).

8.3.2. Estimating frailty groups separately does not increase performance

We use fgNN to compute the 11 frailty groups from table 8.1 simultaneously. Then, we calculate the estimated frailty index from these outputs, and assess the error compared to the actual frailty. Figure 8.2 shows that computing the frailty components separately does not increase the goodness-of-fit of the resulting frailty estimation. Furthermore, the reconstructed frailty seems mostly driven by the prediction of the cardiometabolic group, which aligns with the fact that the input proteins are selected for their relation with the cardiometabolic system.

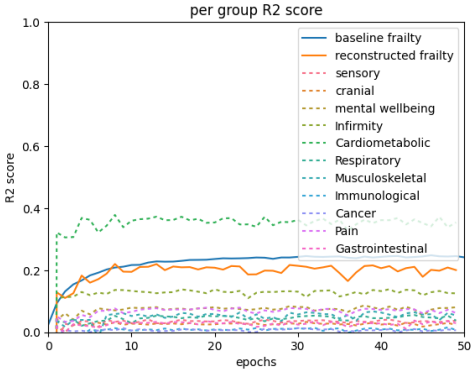


Figure 8.2.: The R^2 score of each output of fgNN, corresponding to a single frailty group, respresented by the dotted lines. Baseline frailty (solid blue) is the output of the baseline model described in section 8.2.7. Reconstructed frailty (solid orange) is the assessment of the estimated frailty from the eleven outputs of fgNN.

8.3.3. Multitask models can give more stable associations with health related outcomes

After hyperparameter optimization, we train a model and explore its associations with various health related outcomes. Specifically, we investigate associations with frailty, mortality, BMI, incident cardiovascular disease, high blood pressure, grip strength decrease, smoking status, and self rated poor health. For fmNN, we ran separate analyses for each of the two outputs (i.e. one linear regression model for the frailty output (fmNN_frailty), and one for the mortality output (fmNN_mortality)). For fgNN, we calculate the estimated frailty from the model outputs, and use that as input to the association analysis.

For the frailty-mortality model, we selected the model with moderate goodness-of-fit performance for both tasks (the model in light green in figure 8.1, with β = 0.05) for association analysis. Furthermore, we ran association analysis using the best performing frailty groups-model based on R^2-score of the reconstructed frailty measure.

Figure 8.3 shows the comparison of both multitask models with ProtFI and ProtMort, the biomarkers introduced in the previous chapter. The fmNN model mostly follows their single task counterparts, with the mortality prediction resembling protMort, while the frailty predictions resembles protFI. Furthermore, the fgNN model shows similar associations as protFI, in some cases showing slightly higher associations (e.g. mortality and self-rated poor health), with other endpoints giving somewhat lower associations (e.g. incident high blood pressure, smoking). Comparing the multitask approaches to one another, fmNN closely resembles fgNN_frailty, though

with consistently lower associations with the exception of mortality. While there is no clear increase in association for either multitask model, the confidence intervals do seem to be consistently smaller for both multitask approaches compared to the single task models.

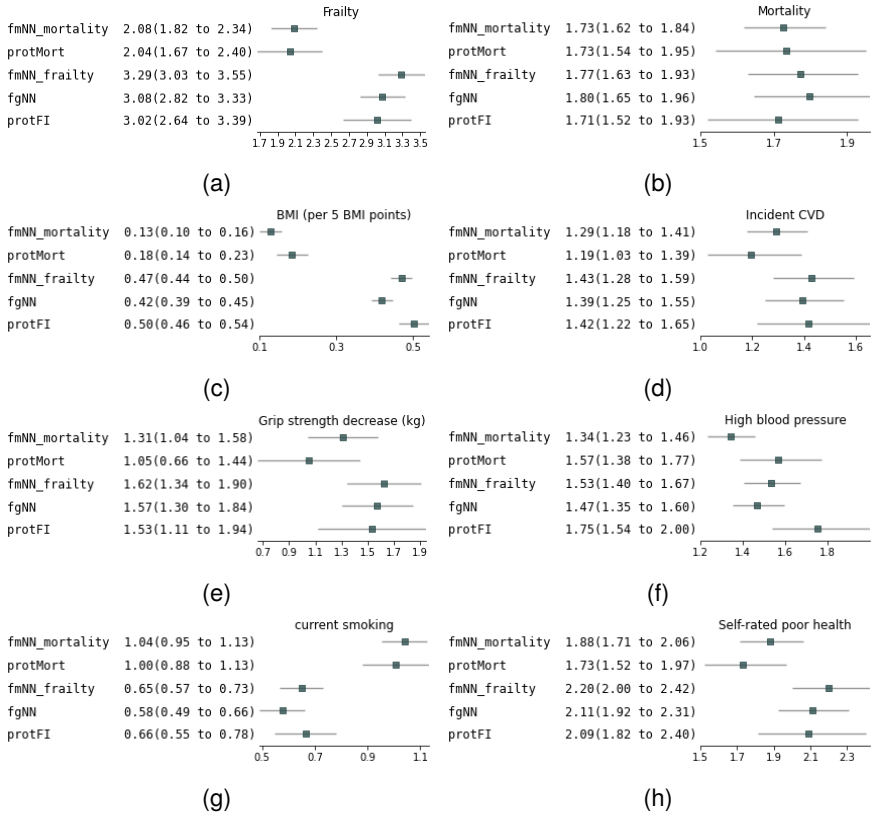


Figure 8.3.: Association analyses comparing the multitask models (fmNN, fgNN) with the earlier developed biomarkers, protFI and protMort. fmNN has two outputs, one for mortality prediction (fmNN_mortality) and for frailty prediction (fmNN_frailty), which are analyzed separately.

8.3.4. Discussion and Conclusion

In this study, we investigated the use of multitask learning to develop protein-based biomarkers of aging by jointly modeling frailty and mortality, as well as by predicting distinct frailty domains that were subsequently aggregated into a composite frailty estimate. While fmNN was able to outperform single-task models, the settings required for the performance gain on one task came at a cost for the other task, and no setting in which fmNN outperformed both single-task models was found. Moreover, predict-

ing frailty groups separately did not increase performance compared to predicting frailty directly, although it did uncover that the cardiometabolic group is the main contributor to the frailty prediction for this model. Finally, while we did not observe improvements in terms of associations with aging-related health outcomes, we did see consistently smaller confidence intervals in these associations, suggesting that these multitask models could be a more robust alternative to the classical single-task models. Nevertheless, we did not observe a consistent increase in associations.

These findings align with broader observations that increased model complexity does not necessarily translate into improved performance for proteomics-based aging biomarkers [3]. In particular, deep learning models optimized for a single target did not outperform simpler elastic net models in this setting, suggesting that the structure of the input data may impose fundamental limitations. Circulating proteomic measurements constitute high-dimensional but largely unstructured tabular data, which deep learning architectures struggle to leverage effectively [13]. As a result, the potential benefits of learning a shared representation for multiple tasks may be attenuated.

Several additional factors may have contributed to the lack of performance gains from multitask learning. First, although frailty and mortality are biologically related, they capture distinct aspects of aging, and the extent to which their underlying protein signatures overlap may be limited. Jointly modeling such outcomes may therefore introduce competing optimization objectives rather than reinforcing shared signal. Second, the multitask architectures explored here were relatively generic and may not have been optimally adapted to the heterogeneity of the frailty domains or the differing noise characteristics of the prediction targets. More advanced task balancing strategies, hierarchical task formulations, or explicit separation of shared and task-specific representations could potentially yield different results.

Finally, the fact that the frailty prediction of fgNN was mostly driven by the cardiometabolic group suggests a possible limitation in frailty prediction from the selected proteins. In general, proteomics carries information on the transient cellular state, which can be broadly informative for predicting cardiometabolic endpoints, perhaps more so than deficits regarding e.g. mental health or cancer. This effect might be exacerbated by the fact that the input proteins were all part of the cardiometabolic panel only, suggesting that they have been selected for their connection to the cardiometabolic system. We decided on using only the cardiometabolic platform based on earlier results, showing that associations with health outcomes did not deteriorate when using these proteins only compared to using all proteins. However, when employing a multitask model, using a wider range of proteins might result in better prediction for frailty categories other than the cardiometabolic group. Nevertheless, applying a multitask approach such as fgNN can help in identifying which frailty categories can be better predicted, which can be insightful for followup analyses.

Bibliography

[1] D. M. Williams, J. Jylhävä, N. L. Pedersen and S. Hägg. 'A frailty index for UK Biobank participants'. In: *The Journals of Gerontology: Series A* 74.4 (2019), pp. 582–587.

[2] M. Moqri, C. Herzog, J. R. Poganik, K. Ying, J. N. Justice, D. W. Belsky, A. T. Higgins-Chen, B. H. Chen, A. A. Cohen, G. Fuellen *et al.* 'Validation of biomarkers of aging'. In: *Nature medicine* 30.2 (2024), pp. 360–372.

[3] S. Garst, L. Kuiper, E. van den Akker, N. v. d. Berg, M. Ghanbari, S. Mooijaart, M. Beekman, M. Reinders, E. Slagboom and J. van Meurs. 'ProtFI, an efficient frailty-trained proteomics-based biomarker of aging, robustly predicts age-related decline'. In: *medRxiv* (2025), pp. 2025–09.

[4] S. Sathyan, E. Ayers, T. Gao, S. Milman, N. Barzilai and J. Verghese. 'Plasma proteomic profile of frailty'. In: *Aging Cell* 19.9 (2020), e13193. DOI: https://doi.org/10.1111/acel.13193. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/acel.13193.

[5] X. Jia, W. Gao, J. Zhang, Y. Zhao, L. Zhang, X. Cao, Y. Wu, L. Ma, L. Chen, L. Sun, H. Guo, C. Zhang, J. Jylhävä, Z. Hu, H. Hagelin, S. Hägg, E. O. Hoogendijk and Z. Liu. 'Plasma Proteomic Signature of Frailty in 50,506 Adults'. In: *medRxiv* (2025). DOI: 10.1101/2025.06.13.25329572. URL: https://www.medrxiv.org/content/early/2025/06/14/2025.06.13.25329572.

[6] E. Dent, P. Kowal and E. O. Hoogendijk. 'Frailty measurement in research and clinical practice: A review'. In: *European Journal of Internal Medicine* 31 (2016), pp. 3–10. ISSN: 0953-6205. DOI: https://doi.org/10.1016/j.ejim.2016.03.007. URL: https://www.sciencedirect.com/science/article/pii/S0953620516300279.

[7] R. Caruana. 'Multitask learning'. In: *Machine learning* 28.1 (1997), pp. 41–75.

[8] J. L. Katzman, U. Shaham, A. Cloninger, J. Bates, T. Jiang and Y. Kluger. 'DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network'. In: *BMC Medical Research Methodology* 18.1 (Feb. 2018), p. 24. ISSN: 1471-2288. DOI: 10.1186/s12874-018-0482-1. URL: https://doi.org/10.1186/s12874-018-0482-1.

[9] N. Draper. *Applied regression analysis*. McGraw-Hill. Inc, 1998.

[10] F. E. Harrell Jr, K. L. Lee, R. M. Califf, D. B. Pryor and R. A. Rosati. ‘Regression modelling strategies for improved prognostic prediction’. In: *Statistics in medicine* 3.2 (1984), pp. 143–152.

[11] A. E. Teschendorff and S. Horvath. ‘Epigenetic ageing clocks: statistical methods and emerging computational challenges’. In: *Nature Reviews Genetics* 26.5 (May 2025), pp. 350–368. ISSN: 1471-0064. DOI: 10.1038/s41576-024-00807-w. URL: https://doi.org/10.1038/s41576-024-00807-w.

[12] S. Garst, J. Dekker and M. Reinders. ‘A comprehensive experimental comparison between federated and centralized learning’. In: *Database* 2025 (2025), baaf016.

[13] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk and G. Kasneci. ‘Deep Neural Networks and Tabular Data: A Survey’. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.6 (2024), pp. 7499–7519. DOI: 10.1109/TNNLS.2022.3229161.

9

Multi-cohort federated learning shows synergy in mortality prediction for MRI-based and metabolomics-based age scores

While biological age scores have been shown to characterize aging by estimating chronological age based on physiological biomarkers, interactions between different age scores are largely unknown. To study this, large-scale multi-modal data are crucial. However, such data are scarce as population-based cohorts are generally restricted in sharing their data. Here, we employ federated learning to study the relationship between the two types of biological age scores: BrainAge based on brain MRI and MetaboAge based on metabolites. Using three large population-based cohorts, we trained a federated deep learning model to estimate BrainAge and compared its performance to models trained in a single cohort. The federated BrainAge model yielded significantly lower error for age prediction across the cohorts than locally trained models. Harmonizing the age interval between cohorts further improved BrainAge accuracy. Subsequently, we compared BrainAge and MetaboAge by performing association analysis and survival analysis for dementia and mortality prediction to further characterize both scores. The association analysis showed a weak association between BrainAge and MetaboAge, while the survival analysis indicated complementary predictive values for the mortality risk of the two scores. Federated learning has shown to be a valuable technique for enabling the use of research cohorts that are restricted in data sharing. We conclude that BrainAge and MetaboAge act synergetically for prediction of time to all-cause mortality and both aging scores capture different aspects of the aging process.

P. Mateus, S. Garst, J. Yu et al. *'Multi-Cohort Federated Learning Shows Synergy in Mortality Prediction for MRI-Based and Metabolomics-Based Age Scores'. In:* Journal of Healthcare Informatics Research *9.4 (2025), pp. 606–628*

9.1. Introduction

Understanding health in the context of aging is challenging, as aging encompasses various functional and structural changes in the body, including alterations in brain structure [1] and body metabolism [2]. As people age, heterogeneity among individuals increases, as some individuals may have greater health changes than what is common for their age. As a result, as aging processes start to become more prevalent, usually starting around the age of 40 [3], chronological age becomes less indicative of health. To address this issue, previous research introduced the concept of biological age, employing biomarkers based on physiological measurements [4], [5], [6]. Such biological age scores may help to understand health in the context of aging, reflecting different components of aging that can progress at a different pace between individuals, and can provide a reference for identifying pathological changes.

The biological age estimation methods consist of regression models optimized to predict chronological age from biomarkers in healthy aging individuals. These scores have been proposed based on various biomarkers using vastly different data modalities. In the field of neuroimaging, brain structure quantified with magnetic resonance imaging (MRI) was used to identify biological age predictors [7] [8] [9] [10] [11] (i.e., BrainAge). BrainAge has been shown to predict mortality [12], various age-related diseases - such as dementia [9], Alzheimer's Disease (AD) and schizophrenia [11], and diabetes type 2 [13] - and non-aging related diseases such as HIV [8]. In the field of metabolomics, several biological age scores have been proposed based on blood-based metabolomics [14], [15], [16] (i.e., MetaboAge). MetaboAge has been associated with cardiometabolic related outcomes, such as diabetes and heart failure, as well as more general aging-related phenotypes, such as decline in instrumental activities of daily living and all-cause mortality [14]. In contrast to MetaboAge, MetaboHealth was also proposed, which predicts time to (all-cause) mortality instead of age [2]. While the predictive value of single brain-based and metabolomics-based age scores has been well studied, the relationships between these different biological age scores are largely unknown. Gaining a better understanding of this relationship may give insight on their added value and on how to optimally combine them to improve their predictive value.

For studying biological age scores, it is crucial that the models generalize effectively to unseen data from diverse sources. Therefore, large-scale data from multiple studies and institutes are required [17]. Additionally, current studies have proposed to use deep neural networks (i.e., [9]), which can capture complex patterns in high-dimensional data, demands access to extensive datasets. The volume and diversity of data necessary for training is not usually owned by a single institution and, therefore, multi-cohort collaborations are essential. However, privacy and safety concerns make it difficult, often impossible, to centrally collect data from multiple cohorts and make it available to train these models.

In recent years, federated learning [18] has emerged as an approach to use sensitive data to train machine learning models while protecting privacy. Rather than training the model in a single institution (known as centralized learning), federated learning

works by separately training at each institution's local computing nodes and only transferring aggregate statistics, like model parameters, between institutions. A central server initiates the model parameters, aggregates the parameters sent back from each node after one or multiple epochs of local training on their local data, and then sends the aggregated parameters to each node. This routine is repeated until the model converges. As a result, it produces an optimized global model with knowledge of diverse local studies, which is trained over several distinct data collections without exchanging the data.

In this study, we apply federated learning over three population-based cohorts to study the relationship between the two types of biological age scores, BrainAge based on brain MRI [9] and MetaboAge based on metabolites [14]. We also include MetaboHealth [2] for comparation. The main contributions of our work are as follows:

1. We provide insight into the relation between the two types of biological age scores. Our results suggest the association between them is primarily driven by age, and they provide complementary information for time to all-cause mortality. However, only BrainAge is predictive of dementia.
2. We implement a federated learning infrastructure connecting three large-scale population cohorts and train a federated deep learning model for BrainAge prediction. To this end, pre-processing the imaging data in all cohorts employing the same pipeline was essential. We observe increased computational time, mainly driven by the desynchronization of the resource availability between cohorts.
3. We compare federated and locally trained deep learning models. The results emphasize the advantage of federated learning in developing models with better generalizability, benefiting cohorts without sufficient data to train a model. In addition, we show that optimizing a federated model may require further attention than a local model due to the heterogeneity between the cohorts' data.

9.2. Methods

9.2.1. Data Preparation

Study population

We included participants from three cohort studies that take part in the Netherlands Consortium of Dementia Cohorts (NCDC): the Rotterdam Study (RS) [19], the Maastricht Study (TMS) [20], and the Leiden Longevity Study (LLS) [21]. These three cohorts include imaging and blood sample data necessary for our analysis.

The Rotterdam Study is a prospective population-based study targeting causes and consequences of age-related diseases among 14,926 community-dwelling subjects aged 45 years and over [19].

The Maastricht Study is a prospective population-based study with focus on the etiology of type 2 diabetes of 10,000 individuals. It comprises individuals aged between 40 and 75 years from the southern regions of the Netherlands [20].

The Leiden Longevity study includes 421 Caucasian families, each comprising long-lived siblings, along with their offspring and the spouses of the offspring. Families meeting the criteria for inclusion had a minimum of two long-lived siblings who were alive and willing to participate. Males were considered long-lived if 89 years or older and females 91 years or older [21].

In addition, we used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`) in the preparation of our federated learning infrastructure (detailed in the Supplementary Material).

Data selection and splitting

The two main modalities used are T1-weighted MRI brain data [9] and metabolomic data from the Nightingale metabolomics platform measured on blood draws [14].

The data selection flowchart is shown in Figure 9.1. We included participants from the studies who had at least complete data of age, sex and brain MRI scans. For RS, 5,409 participants were included from the RS-I, RS-II and RS-III cohorts (2,509 participants had both MRI scans and blood samples available acquired within 7 years from each other). For TMS, 5,055 participants were included, of whom 2,419 with both MRI and blood; for LLS, 364 participants were included (all with both MRI and blood). For participants with metabolomic data, only the MRI scan taken closest to the blood sample was included for each participant, and additional covariates were selected if available: diabetes mellitus (DM), i.e. diabetes type 1 or 2 diagnosis, body mass index (BMI) and education category (EC) corresponding to their blood sampling time. Considering blood samples and MRI scans were taken at different times in RS and TMS, we used the interval years (lag time) between blood sampling time and MRI scanning time as an additional covariate.

The data was first split into two parts, based on the availability of metabolomic data. Participants with metabolomic data and absolute lag time < 7 years (between blood samples and MRI scans) were used for testing the BrainAge model, performing the correlation analysis and the survival analysis (after excluding those with missing values in the covariates). In TMS and RS, all other participants were randomly split into training (80%) and validation (20%) set for training the BrainAge model. LLS was used as external validation for assessing BrainAge performance on unseen cohorts. In TMS and RS, to maximize the number of training images, thereby introducing a natural type of data augmentation, we used all available longitudinal MRI scans for each participant. Table 9.1 shows the number of MRI scans used in the analyses.

Image processing

FreeSurfer version 6.0 [22] was used to segment supratentorial gray matter (GM) based on the T1-weighted brain MRI images [23]. GM density maps were computed based on an optimized voxel-based morphometry (VBM) protocol using the FSLVBM pipeline [24] [25] . First, all GM maps were nonlinearly registered to the standard Montreal Neurological Institute GM probability template (ICBM 152 Nonlinear atlases

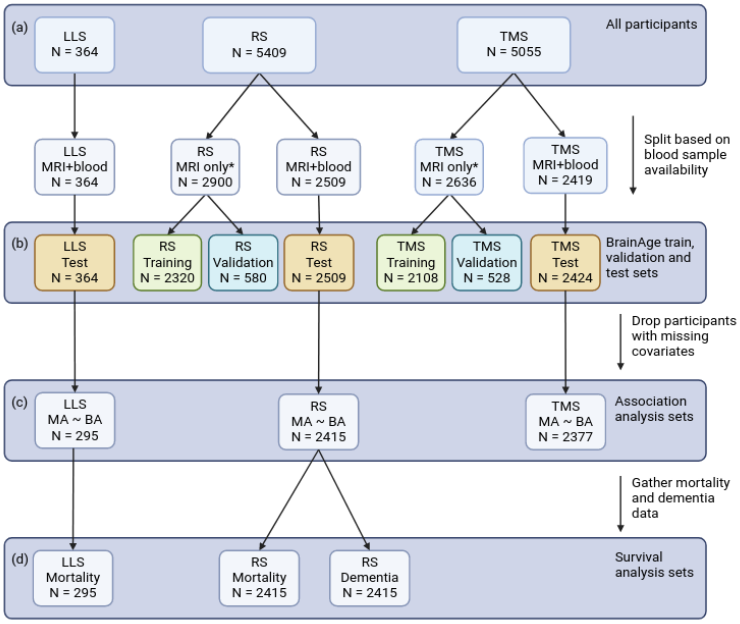


Figure 9.1.: Data selection flowchart for all analyses. (a): The total number of participants included in the three cohorts: Leiden Longevity Study (LLS), Rotterdam Study (RS), and The Maastricht Study (TMS). (b): We split subsets based on availability of blood samples for the BrainAge training. (c): The participants have both MRI scans and blood samples are used as test set for the BrainAge model and are used for association analysing. Covariates considered are: age, sex, lag time, body mass index, diabetes mellitus and education category.(d) We run the survival analysis on a subset of cohorts because of the availability of mortality and dementia incidence data.

version 2009) with a 1 × 1 × 1 mm voxel resolution. Second, a spatial modulation procedure was used to compensate for differences in absolute GM volume due to the registration. This is achieved by multiplying voxel density values by the Jacobian determinants estimated during registration. The matrix size of the modulated GM density maps was 196 x 232 x 188. As smoothing is a subgroup of possible mathematical operations which the network filters in the convolutional layer can represent, we did not apply smoothing on the VBM results. We performed a quality control based on the proportion (5%) of outlier voxels and an additional manual check to exclude the outliers. Finally, we applied cropping and padding on the images to cut proper 0 edges, and masked the images with a k-Nearest-Neighbor-classifier segmented GM mask [26]. The matrix size of the final images was 160 x 192 x 144.

Table 9.1.: Data split in numbers of MRI scans for the three cohorts

	The Rotterdam Study (RS)	The Maastricht Study (TMS)	The Leiden Longevity Study (LLS)
Total number of scans	8318	5055	364
Training and validation	5809	2636	-
Test	2509	2419	364

Includes multiple scans at different time points for some participants
Includes healthy participants with MRI scans ONLY or with
blood draw and MRI scan lag time > 7 years

Image acquisition details from each cohort are described in Supplementary Material 2.1.

9.2.2. Federated Learning Infrastructure

For the federated learning infrastructure, we adopted the Vantage6 Personal Health Train (PHT) framework [27]. Vantage6 is a dockerized solution for federated learning, which comes with an access control system. A Vantage6 system consists of a central PHT server node and a set of distributed PHT station nodes. Each station node is located behind the institute's firewall and its control connection is regulated by the central server node, which uses a private key to determine which nodes can connect. Federated learning algorithms are implemented as Docker images and authorized by each station node.

The Vantage6 framework addresses the main requirements to establish the federated network. However, additional adaptations were necessary to train the deep learning model, such as establishing a connection to local high-performance clusters and guaranteeing data interoperability between the nodes. For this, we created a technical solution by extending the Vantage6 software and harmonizing the data in each station node using a data model [28] for clinical data and the open-source Extensible Neuroimaging Archive Toolkit [29] (XNAT) for imaging data. Detailed technical descriptions can be found in sections 1.1 and 1.3 of the Supplementary Material. Additionally, we established a governance protocol to address data security and privacy preservation, as described in section 1.2 of the Supplementary Material.

9.2.3. Deep Learning for BrainAge prediction

We used a 3D CNN model architecture proposed by [9] to train a BrainAge model. This network takes as input the GM density maps obtained from the MRI scans and outputs a predicted age. The architecture consists of four convolutional blocks, used to extract valuable image features, followed by a fully connected layer that concatenates information on the participant's sex. We used the mean squared error (MSE) as the loss function to train the model and optimized the model parameters

based on the model with the lowest MSE on the validation set. The model's accuracy was evaluated using the mean absolute error (MAE) on the test set. Both metrics measure the difference between model output and the participant's chronological age. To evaluate the associated uncertainty with each model, we performed bootstrapping with resampling (1000 resamples) on the test set to calculate the 95% confidence interval. Additionally, to better estimate the model's performance, we performed a 3-fold cross-validation.

Age-bias correction

As observed in previous studies [30], BrainAge models are prone to overestimate the age of younger participants and underestimate the age of older participants. Since this behavior can impact subsequent analysis, an age-bias correction is normally applied using a linear regression model. In our study, we calculated three age-bias correction models based on [31]'s approach, one for each training set separately (TMS and RS) and one for the federated approach. Additionally, we evaluated the generalizability of these models by assessing the performance in the cohorts' test sets.

Federated Training

We trained the BrainAge model using federated averaging (FedAvg) [18]. Initially, the deep learning model weights are randomly initialised and distributed to the participating cohorts. For every cohort, the model is individually trained for several epochs on their data, starting from the shared parameters. Next, the local model is sent back to the central server. Here, the model parameters are aggregated and shared with the training cohorts. This cycle continues until reaching the convergence criteria.

Implementation Details

We trained the deep learning model using Tensorflow [32] (version 2.8.0) and Python (version 3.8). We did not employ data augmentation and manually tuned the hyperparameters based on the original model. Additional details are provided in Section 2.3 of the Supplementary Material. We trained the network for 20 rounds, three epochs each round and used a batch size of eight. The model's weights were initialized using the default Tensorflow method, the Xavier initialization [33]. The Adam algorithm was employed to train the model. Furthermore, we employed Docker [34] to containerize the scripts developed and provide the exact environment used in our experiments within a Docker image. A complete description of the libraries employed and the respective versions is provided in the public repository. Training and testing were performed at each cohort GPU cluster, specifically: an NVidia A40 GPU with 48GB and an NVidia RTX 2080 Ti GPU with 11GB for RS, a Tesla V100 with 32GB of RAM for TMS, and a TitanXp with 12GB for LLS. Finally, we followed

the checklist for artificial intelligence in medical imaging (CLAIM) [35], assessment provided in the Supplementary Material, to promote the reproducibility of our work.

9.2.4. MetaboAge prediction

We applied the trained model from [14] to determine MetaboAge scores for the three cohorts. MetaboAge is a linear model based on a selection of 56 metabolites, as measured by the high-throughput proton nuclear magnetic resonance (^{1}H-NMR) metabolomics measurement platform Nightingale [36]. The original model was trained and tested on a total of 18,716 blood samples, originating from 26 Dutch biobanks with ages ranging from 18 to 85 years.

9.2.5. MetaboHealth prediction

MetaboHealth [2] is a Cox proportional hazards model trained to predict all-cause mortality. MetaboHealth uses 14 metabolites from the high-throughput proton nuclear magnetic resonance (^{1}H-NMR) metabolomics measurement platform Nightingale [36], similar to MetaboAge. These were selected using a forward-backward process, that identified the metabolites with the lowest correlation with each other while being the most predictive for age at death. The original model was trained on 44.168 samples from 12 cohorts, with ages ranging from 18 to 110 years.

9.2.6. The relation between biological age scores

Association Analysis

We explored the association between BrainAge and the MetaboAge through linear regression. We also used MetaboHealth in the association analysis with BrainAge as comparation, since MetaboHealth is not explicitly trained on age and as such is not expected to have a high correlation with BrainAge that is driven by age. We used MetaboAge or MetaboHealth as the response variable, and BrainAge as the explanatory variable:

$$Y = \beta_b * BrainAge + \sum_{x_i \in X} \beta_i * x_i \tag{9.1}$$

with Y being MetaboAge or MetaboHealth, and with X being the set of covariates. BainAge and MetaboAge were not corrected for age bias since we included age as covariate in the association analysis. We firstly trained a model with no covariates. In addition, to correct for confounders, three sets of covariates were considered:

1. Adjustment for age: $X = \{Age\}$
2. Additional adjustment for sex, DM and lag time:
   $X = \{Age, Sex, DM, Lag\ Time\}$

3. Additional adjustment for BMI and EC:
   $X = \{Age, Sex, DM, Lag\ Time, BMI, EC1, EC3\}$

Since age seemed to have a large effect on these models in our experiment, we also ran both 2 and 3 without age, finally resulting in 6 sets of covariates in total. All variables were normalized when training the models.

The linear regression was also performed using the federated approach. Each round t, the PHT server sends out global beta values ($\beta_g = \{\beta_b, \beta_1, .., \beta_i\}$) to all cohorts. These then create a local update of the beta values (β_l) using one iteration of gradient descent:

$$\beta_l^t = \beta_g^t - \eta * \Delta L \tag{9.2}$$

with L being the mean squared error loss function. η was set at 0.1 for all models. Then, each cohort sends back their own β_l to the PHT server, which creates a new value for β_g using a weighted average:

$$\beta_g^{t+1} = \frac{1}{\sum_{j=0}^{J} n^j} \sum_{j=0}^{J} n^j * \beta_{l,j}^t \tag{9.3}$$

with $\beta_{l,j}^t$ being the local beta values coming from cohort j at round t. This iterative process continued until MAE did not change anymore. We repeated this process 10 times, and chose the model with the lowest MAE. The beta values of the linear regression model were used for measuring the association of the involved covariates.

In addition, we compared the federated model to the closed-form solution using meta-analytical framework HASE [37].

Survival analysis

To assess the complementary value of biological age scores in estimating the vulnerability of individuals, we performed survival analyses using Cox proportional hazards models based on BrainAge and MetaboAge. We used the difference between participants' age score and their chronological age, BrainAge Gap (BAG: BrainAge - Age) and MetaboAge Gap (MAG: MetaboAge - Age) as input to the survival analyses:

$$\lambda(t|MAG, BAG, X) = \lambda_0(t) * exp(\beta_1 * MAG + \beta_2 * BAG + \sum_{x_i \in X} \beta_{x_i} x_i) \tag{9.4}$$

with X being the set of covariates adjusted for. Two sets of covariates were used: 1) age only ($X = \{Age\}$) and 2) the full set of covariates used in the association analysis
($X = \{Age, Sex, DM, Lag\ Time, BMI, EC_1, EC_3\}$). BainAge and MetaboAge were not corrected for age bias since we included age as covariate in the survival analysis.

We compared the survival curves on the 1^{st} and 3^{rd} quartiles of BAG and MAG. Intuitively, individuals at this first cutoff point have a lower than average age gap, indicating that they have aged less than expected. Conversely, at the second cutoff, the age score is relatively higher than average, indicating accelerated aging. Combining the two scores, this creates four cutoff points:

- BAG 1^{st} quartile, MAG 1^{st} quartile : decelerated aging according to both BrainAge and MetaboAge
- BAG 1^{st} quartile, MAG $3rd$ quartile : decelerated aging according to BrainAge, accelerated aging according to MetaboAge
- BAG 3^{rd} quartile, MAG 1^{st} quartile : accelerated aging according to BrainAge, decelerated aging according to MetaboAge
- BAG 3^{rd} quartile, MAG 3^{rd} quartile: accelerated aging according to both BrainAge and MetaboAge

Separate survival curves were created by applying the trained model from equation 9.4. We further examined using MetaboHealth instead of MAG, and the separate and pairwise effects of BAG, MAG, and MetaboHealth on the risks of dementia and mortality, adjusting for the full set of covariates.

The survival analyses were run locally in RS with mortality and dementia as outcomes, and in LLS on mortality only as this cohort had only few dementia cases at the latest follow-up ($N = 3$); TMS had no long-term mortality or dementia data available.

9.3. Results

9.3.1. Population characteristics

The subject characteristics are shown in Table 9.2. No demographic bias was found between participants with MRI scans only and participants with both MRI scans and blood samples available within each cohort. However, there are some biases between the cohorts. TMS has younger participants, while the RS has elder participants and a larger proportion of females. In addition, TMS has a larger proportion of diabetes cases than the other two cohorts. Such cohort level difference is quite common in a multi-center study and we expect our federated learning based analysis is able to mitigate this bias.

9.3.2. BrainAge and MetaboAge models

BrainAge model

Using the federated learning infrastructure, we ran the training and testing of our federated BrainAge model at the three cohorts locations. We used TMS and RS as training cohorts, with LLS held out as external testing cohort. As shown in Table 9.3,

Table 9.2.: Population characteristics summary of 3 cohorts included in the study

	The Rotterdam Study (RS)	The Maastricht Study (TMS)	The Leiden Longevity Study (LLS)
participants with MRI scans ONLY	2,900	2,636	0
female, n (%)	1,583 (54.6)	1,331 (50.5)	-
age at MRI, mean (sd)	67.0 (10.5)	59.6 (9.0)	-
age at MRI, range	45 - 100	40 - 75	-
participants with MRI scans AND blood samples	2,509	2,419	364
female, n (%)	1,433 (57.1)	1,187 (49.0)	190 (52.2)
age at MRI, mean (sd)	67.5 (9.5)	60.3 (8.4)	65.5 (6.6)
age at MRI, range	46 - 96	40 - 75	45 - 84
participants with complete covariates eligible for association analysis	2,415	2,377	295
lag time (years) between blood sample and MRI scan, mean (sd)	-1.39 (3.22)	2.2 (1.3)	0
BMI, mean (sd)	27.3 (3.8)	26.6 (4.2)	25.3 (3.3)
educational level low/medium/high, %	46.3/29.7/24.0	30.4/28.9/40.7	55.2/8.6/36.2
diagnosis of diabetes, n (%)	221 (8.8)	538 (22.2)	21 (5.8)
diagnosis of dementia , n (%)	154 (6.4)	-	3 (1.0)
follow-up years of dementia, mean (sd)	6.9 (2.9)	-	13.0 (2.5)
mortality, n (%)	662 (27.4)	-	48 (16.3)
follow-up years of mortality, mean (sd)	10.0 (2.9)	-	13.0 (2.5)

Additionally includes participants with blood sampling and MRI scanning lag time > 7 years, and excludes scans with dementia diagnose or stroke.
8 missing values for dementia in RS and 1 missing value in LLS.

the federated BrainAge model demonstrated its capability to predict the chronological age across different cohorts with mean absolute errors (MAE) of 5.59 years in the TMS test set, 4.36 years in the RS test set, and 4.60 years in the external test set (LLS). To better interpret such federated BrainAge model performance, we also ran local models (trained at TMS or RS separately). Our federated model showed a better performance than local models that were tested on data from a different cohort. To highlight, the local model trained on RS data (training MAE = 2.48, test MAE = 4.21) achieved lower performance on TMS (test MAE = 6.45) and LLS (test MAE = 5.66) than the federated model. Similarly, the local model trained on TMS data (training MAE = 3.10, test MAE = 4.72) achieved lower performance on RS (test MAE= 7.29)

and LLS (test MAE = 5.63) than the federated model, suggesting the local models were unable to maintain performance when tested with data from a different cohort. Furthermore, the performances for the federated model in each cohort and subset were more similar than those of the local models, indicating less overfitting. These observations are supported by the results obtained with the 3-fold cross-validation approach (Supplementary Table 4), indicating similar MAE estimates for the three models.

We observed that the federated BrainAge model yielded a higher MAE for TMS cohort than for the other two cohorts. We suspected that this higher MAE was due to the TMS cohort being relatively younger than the overall training set (mean age TMS: 59.6 years; mean age RS: 67.0 years). To further validate this, we trained the federated model and two local models with a sub-selection of the participants in the training set with the same age range between 53 and 75 years in both RS and TMS (70% of the training set, mean age of 61.8 years). With this new training set with similar age ranges, we observed that the model converged to a solution with smaller MAE differences in all three cohorts (TMS test MAE of 4.97 vs 4.87 in RS and 4.21 in LLS).

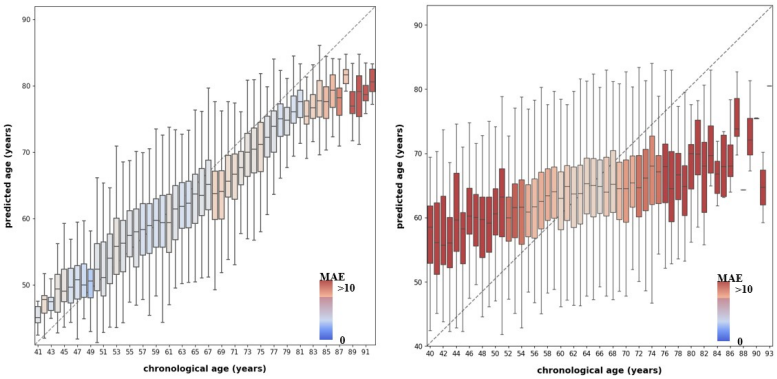
Table 9.3.: MAE [95% confidence interval] of BrainAge models trained locally (at TMS and RS separately) and in a federated way (using both TMS and RS). Models were trained and tested in The Maastricht Study (TMS) and the Rotterdam Study (RS), using the Leiden Longevity Study (LLS) as the external test cohort.

		Local models		Federated model
		TMS	RS	TMS & RS
TMS	Training	3.10 [2.99, 3.23]	-	4.75 [4.61, 4.90]
TMS	Validation	4.59 [4.30, 4.87]	-	5.56 [5.23, 5.89]
TMS	Testing	4.72 [4.59, 4.87]	6.45 [6.22, 6.71]	5.59 [5.44, 5.76]
TMS	Testing	4.67 [4.58, 4.73]	6.55 [6.39, 6.71]	4.97 [4.84, 5.12]
RS	Training	-	2.48 [2.42, 2.54]	4.34 [4.25, 4.43]
RS	Validation	-	2.50 [2.37, 2.62]	4.87 [4.66, 5.06]
RS	Testing	7.29 [7.11, 7.44]	4.21 [4.09, 4.34]	4.36 [4.21, 4.48]
RS	Testing	7.00 [6.22, 7.62]	5.10 [4.95, 5.25]	4.87 [4.74, 5.01]
LLS	Testing	5.63 [5.24, 6.05]	5.66 [5.26, 6.08]	4.60 [4.25, 4.95]
LLS	Testing	5.82 [5.10, 6.25]	5.89 [5.51, 6.30]	4.21 [3.85, 4.60]

Testing results of the model trained with a sub-selection of the training participants with age between 53 and 75 years

Additionally, we observed a tendency of the model to overestimate the age of younger subjects and underestimate the age of older subjects, as shown in Figure 9.2a. Although this tendency was apparent in all cohorts, differences existed in the age interval where it occurred between the cohorts. We therefore applied an age-bias cor-

rection to the BrainAge model (see methods section 9.2.3 for details). Supplementary Table 5 shows that bias correction on the federated model resulted in considerable improvements for the RS (MAE before correction: 4.36, MAE after correction: 3.33) and LLS (MAE before correction: 4.60, MAE after correction 3.62) but little for the TMS (MAE before correction: 5.59, MAE after correction: 5.51). Moreover, evaluating the bias correction with data from a single cohort, with either the TMS or the RS training set, displayed considerable improvements in the corresponding cohort (see Supplementary figure 2, first two rows) but did not benefit external cohorts.



(a) (b)

Figure 9.2.: Chronological age vs. predicted age for (a) Federated BrainAge and (b) MetaboAge for the complete test set. The color indicates the predicted Mean Absolute Error (MAE) of participants at certain chronological age, with color red indicating an MAE > 10.

We further compared the computation time, which was 64% higher on average for federated training of the BrainAge model than that of a central training (Table 9.4). We also optimized the federated training hyperparameters, as detailed in the Supplementary Table 2. Optimized model hyperparameters were an initial learning rate of 1×10^{-3}, a learning rate decay of 1×10^{-2}, and a dropout rate in the last layer of 5×10^{-1}. In the federated architecture, selecting the model with the lowest MAE from each cohort showed optimal convergence. Performance was most similar between cohorts when models were aggregated between cohorts with equal importance weighting.

MetaboAge model

The MetaboAge models resulted in prediction performances of MAE = 7.9 for the TMS test set, MAE = 7.4 for the RS test set, and MAE = 6.4 for LLS. Overall prediction results per chronological age can be seen in Figure 9.2b.

These results show a larger age bias than those for BrainAge. They suffer more from regression to the mean and have a wider range of predictions for each age bracket.

Table 9.4.: Computation time (range) taken to train the BrainAge model centrally and in a federated way (same computational resources in both approaches). The results presented comprise the time from the 3-fold cross-validation training.

Trained	Data	Computation Time (hours)	Number of Epochs	Time / Epoch (min)
Centrally	TMS	13.70 (12.93, 14.53)	100	8.22
Centrally	RS	16.83 (16.35, 17.23)	100	10.10
Federated	TMS & RS	15.22 (13.23, 16.56)	60	15.22

9.3.3. Relation between the biological age scores

Association analysis

The association analysis results between BrainAge and MetaboAge, BrainAge and MetaboHealth are shown in Table 9.5. A more detailed table including p-values and standard errors can be found in Supplementary Table 7. For MetaboAge, the model with BrainAge as the only predictor (M1) resulted in a small but significant association (beta = 0.16, SE = 0.014, P = $4.3 * 10^{-32}$) (Table 9.5 (a)). Then, adding age as a covariate (M2) showed a strong effect on the observed relationship between BrainAge and MetaboAge (beta = -0.08, SE = 0.022, P = $6 * 10^{-5}$), indicating that the information both scores provide is chronological age. This effect of age on the relation between the two age scores was consistent when including sex, DM and lag time (M3-M4), as well as when further adding BMI and EC (M5-M6). Covariates other than age did not show a strong influence on the relation between BrainAge and MetaboAge. For MetaboHealth, the association between BrainAge and MetaboHealth without covariates was similar to that of MetaboAge (Table 9.5 (b)). However, this correlation persisted after adjusting for age, suggesting that BrainAge and MetaboHealth share common information beyond chronological age.

Table 9.5.: Beta values for various levels of covariates for estimating MetaboAge (a) and MetaboHealth (b).

(a)

Model ID	BrainAge	Age	Sex	DM[1]	Lag Time	BMI[2]	EC1[3]	EC3[3]	Error (MAE)
M1	0.16**	-	-	-	-	-	-	-	0.77
M2	-0.08*	0.32**	-	-	-	-	-	-	0.75
M3	0.25**	-	-0.17**	0.16*	0.08*	-	-	-	0.73
M4	-0.01$^{=}$	0.39**	-0.14**	0.19*	0.14**	-	-	-	0.74
M5	0.22**	-	-0.27**	0.08*	0.01$^{=}$	-0.03*	0.16*	0.11*	0.76
M6	-0.06*	0.38**	-0.23**	0.08*	0.04*	-0.03*	0.07 *	0.11*	0.74

(b)

Model ID	BrainAge	Age	Sex	DM[1]	Lag Time	BMI[2]	EC1[3]	EC3[3]	Error (MAE)
M1	0.13**	-	-	-	-	-	-	-	0.76
M2	0.11*	0.03$^{=}$	-	-	-	-	-	-	0.76
M3	0.13**	-	-0.04*	0.70**	0.02$^{=}$	-	-	-	0.76
M4	0.10*	0.02$^{=}$	-0.04*	0.71**	0.04*	-	-	-	0.75
M5	0.10**	-	-0.01$^{=}$	0.68**	0.00$^{=}$	0.15**	-0.07*	-0.26**	0.75
M6	0.09*	0.06*	0.06*	0.66**	0.03*	0.10**	-0.10*	-0.27**	0.74

DM = Diabetes Mellitus, i.e. diabetes (type 1 or 2) diagnosis.
BMI = Body Mass Index
EC1-3 = Education Category, mapped to low/medium/high based on years of education.
One-hot encoded relative to the medium level.
$P \leq 0.05$
$P \leq 5 * 10^{-10}$
$P > 0.05$

Survival analysis

To investigate whether BrainAge and MetaboAge have complementary information about an individuals' health, we performed a survival analysing assessing there predictive performance fore the time to mortality and dementia prediction.

Figure 9.3 shows the survival curves for four groups of individuals that are either young-appearing or old-appearing as quantified by the biological age scores. Group are formed with different combinations of BrainAge and Metaboage scores, having either a relatively low or high BrainAge Gap (BAG: BrainAge-Age) and either a relatively low or high MetaboAge Gap (MAG: MetaboAge-Age). We inspected the survival of different combinations of these scores, in which a clear separation can be observed. Young-appearing individuals on both biomarkers (MAG 1st quartile; BAG 1st quartile) showed the highest survival rate, while old-appearing individuals (MAG 3rd quartile, BAG 3rd quartile) showed the lowest survival rate. Indiviuals scoring different on both markers had an intermediate survival rate. As suggested from the hazard ratios of BAG and MAG, this effect was more pronounced in RS than in LLS, that both the BAG and MAG have more significant effects on the survival probability in RS. For dementia prediction (Supplementary Figure 4), we found that only BAG was significantly associated with the time to dementia diagnosis, while MAG did not differentiate participants with dementia, independent of covariates. Using MetaboHealth instead of MAG in the survival analyses (Supplementary Figure 5) resulted in similar observations, though with a larger hazard ratio for MetaboHealth, likely due to the fact that MetaboHealth was trained to predict mortality.

Supplementary Table 9 and 10 show the separate and pairwise effects of BAG, MAG, and MetaboHealth on the risks of dementia and mortality, adjusting for the full set of covariates. The combined analysis including both MAG/MetaboHealth and BAG showed minimal changes in their coefficients and significance compared to the separate models. This pattern was consistent across both dementia and mortality outcomes, suggesting that BAG and MAG contribute independent information to the prediction of these risks. MetaboHealth, similar to MetaboAge, was not associated with dementia risk. However, it showed a much stronger association with mortality, even after adjusting for BAG. When included together in the same model, MetaboHealth attenuated the significance of MAG, indicating potential overlap in the biological information they capture regarding mortality risk.

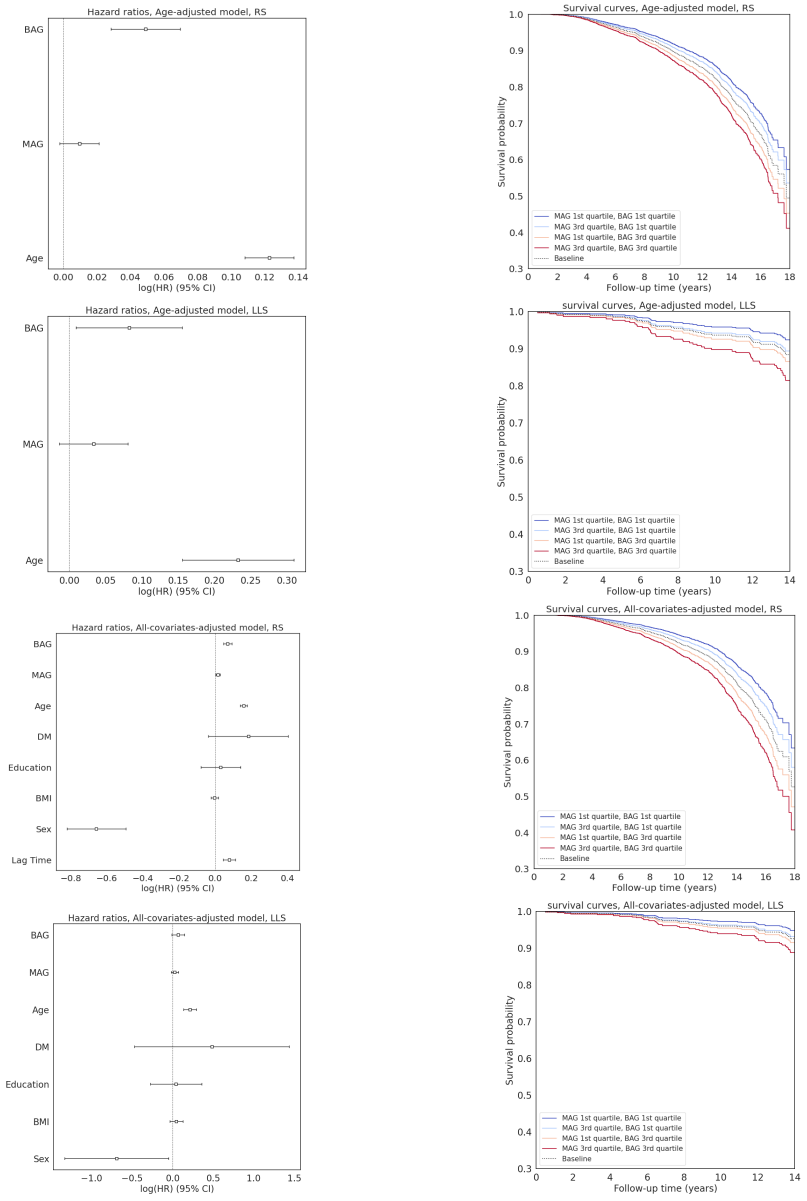


Figure 9.3.: Survival analysis for mortality prediction in the Rotterdam Study (RS) and the Leiden Longevity Study (LLS) using Cox Proportional Hazard models. Figure a-d show the results of the age-adjusted models (a-b in RS, c-d in LLS), taking only BrainAge Gap (BAG), MetaboAge Gap (MAG) and age into account. Figures e-h show the results for the all covariates-adjusted models (e-f in RS, g-h in LLS), additionally adjusting for diabetes mellitus, education category, body mass index, sex and lag time. Note that the lag time in LLS is 0 for all participants.

9.4. Discussion and Conclusion

This study demonstrated a federated BrainAge model on three large-scale population based cohorts that outperformed local models trained on only one cohort. Our federated analysis results additionally suggest that BrainAge and MetaboAge carry non-overlapping information with regard to time to all-cause mortality.

Regarding MetaboAge results, the test performance of MetaboAge prediction was comparable to the original study with a median error of 7.3 [14]. Regarding BrainAge results, the performance of the federated BrainAge model was similar to other deep learning and 3D CNN-based models reported in the literature (MAE between 4-5 years) [38] [39]. These suggest the validity of our models. The federated model yielded significantly lower error (MAE) for age prediction based on Brain MRI across cohorts than the locally trained models, showing that the federated model has a better generalizability to external data. Such generalizability is a major concern for data methods in current medical practice [40]. Although data privacy rules prevented comparing a centralized model (trained on pooled data) with the federated model, both performed similarly when trained on a publicly available dataset (Section 2.2 of the Supplementary Material). Federated learning can enable cohorts with insufficient data to train an accurate model leveraging other cohorts' datasets to still get accurate model predictions. However, this is only possible if federated models generalize well to unseen cohorts.

Age differences between cohorts had an impact on the results. We observed that the BrainAge model performance showed smaller MAE differences between cohorts when restricting participant selection to an equal age interval on all cohorts. Besides, we observed that both BrainAge and MetaboAge tend to overpredict the age of younger subjects and underpredict the age of older subjects, which is a known problem for biological age scores [41] based on chronological age. Part of the performance gap between cohorts can be attributed to this effect. We showed that a bias correction can help to decrease this tendency for our federated BrainAge model, however its effectiveness varies, especially for external test cohorts. Additional methods focused on both sample-level and age-level bias may improve the correction effectiveness and the reliability of BrainAge [42]. As for explainability, Wang et al. [9] analyzed the model's attention maps which highlighted the importance of the amygdala and hippocampus regions in predicting brain age, especially with increasing chronological age.

Regarding the federated association analysis, we found a low association between BrainAge and MetaboAge, which was drastically reduced after adjusting for age, indicating that the main association between the two biological age scores is their common correlation with age. It is known that different aspects of aging, such as cardiovascular and cognitive decline, can occur at different paces between individuals [3]. We therefore hypothesize that the underlying reason for the low association between BrainAge and MetaboAge is that they characterize different aging aspects, while being less informative for the aspect for which the other score is most informative. The low association between BrainAge and MetaboAge suggests complementarity of both scores, which is supported by the survival analyses. In contrast, association

between MetaboHealth and BrainAge did remain when adjusted for age. This could be related to earlier observations that MetaboHealth is (like BrainAge, and unlike MetaboAge) associated with cognitive decline [43].

Aside from age, the other covariates for which we corrected had only a minor impact on the association between BrainAge and MetaboAge. These tested covariates included education level, which has a known connection with lifestyle [44] [45]. Due to low coverage, we did not adjust for other lifestyle factors such as physical activity or medication use, which could direct future research upon improving the data collection. Future study of those additional lifestyle factors will help us to better understand the observed relation between BrainAge and MetaboAge.

The survival analysis showed that individuals who scored high (indicating accelerated aging) on both BrainAge and MetaboAge had a lower survival rate than those that scored low on one or both of the scores. This suggests that BrainAge and MetaboAge have complementary information about an individual's mortality. While only BrainAge was predictive for dementia, which also indicates the differing roles of the two age scores and the crucial role of brain-specific aging processes in the individuals' onset of dementia.

In the survival analysis results there are specific differences between cohorts that could be explained by study design differences between RS and LLS. Mortality was lower and BrainAge was less informative in LLS compared to RS. While LLS had inclusion criteria favouring healthy and long-living individuals, RS aims to include a general population and is less selective. We therefore hypothesize that LLS included participants with a relatively low BrainAge.

One of the limitations of our analysis was the relatively limited amount of events in our survival analysis. Survival analysis on dementia was only possible in RS due to the lack of dementia cases in LLS. The amount of cases for mortality was 27.4% and 16.3% in RS and LLS, respectively. Performing a survival analysis with more cases could strengthen our results. Another limitation is the diversity of the populations. Although RS and TMS are population-based studies, they mostly include participants from western European descent, limiting the applicability of our findings to other populations. Finally, due to the inability to share data, our federated BrainAge model could not be compared to a centralized model trained on the same collection of data.

Federated learning and federated analysis enable the use of data for which collaboration was not possible before, thereby increasing the pool of data for research. However, using a federated infrastructure for real-life data also comes with several challenges. First, harmonized data pre-processing across cohorts is essential [46] [47]. We took account of this by harmonizing all data and reprocessing all imaging data with the same image analysis pipeline. Second, the distributed processing environment may provide challenges both for the model optimization itself as well as for the optimization time needed. Regarding optimization, data heterogeneity between cohorts can lead to either overfitting in a single cohort or fluctuation in convergence between different cohorts when training the BrainAge model in a federated setting. By altering parameters such as lowering the local amount of epochs and increasing the dropout rate, we were able to decrease overfitting on the largest cohort (see

supplementary results 2.3 for details). Furthermore, we experienced that the time needed for model optimization was relatively high as compute resource availability was not synchronized between cohorts. Finally, Vantage6 was unable to interact directly with the cohort's high-performance compute platform required to train the BrainAge model. Therefore, we created a technical solution by extending the station node Docker image to allow establishing a connection to local HPC platforms.

For the choice of federated learning platform, we considered two aspects. First, regarding suitability for deployment, Vantage6 provides a user authentication system and a whitelisting system for algorithms that allow secure implementation. Second, regarding convenience of algorithm development, Vantage6 runs all algorithms in docker containers providing flexibility on what algorithms to use not limited to federated learning but also allowing for example federated association analysis. Nevertheless, as federated learning platforms are still emerging and under development, the choice for the most optimal platform might change over time.

In summary, this study used federated learning to train and validate a BrainAge model across three cohorts and federated analysis to perform association and survival analysis of BrainAge and MetaboAge/MetaboHealth. Our results highlight that federated learning is a promising technique for cases in which data sharing is not possible. In addition, we conclude that BrainAge and MetaboAge act synergetically for prediction of time to all-cause mortality. We consider combining biological age scores based on different data modalities an interesting future research direction, as a combined age score will provide a more complete information for understanding health and may have a higher predictive value for identifying pathological changes in individuals.

Bibliography

[1] E. Vinke, M. de Groot, V. Venkatraghavan, S. Klein, W. Niessen, M. Ikram and M. Vernooij. 'Trajectories of imaging markers in brain aging: the Rotterdam Study'. In: *Neurobiol Aging.* (2018). DOI: 10.1016/j.neurobiolaging.2018.07.001.

[2] J. Deelen, J. Kettunen, K. Fischer, A. van der Spek, S. Trompet, G. Kastenmüller, A. Boyd, J. Zierer, E. van den Akker, M. Ala-Korpela, N. Amin, A. Demirkan, M. Ghanbari, D. van Heemst, M. Ikram, J. van Klinken, S. Mooijaart, A. Peters, V. Salomaa, N. Sattar, T. Spector, H. Tiemeier, A. Verhoeven, M. Waldenberger, P. Würtz, G. Davey Smith, A. Metspalu, M. Perola, C. Menni, J. Geleijnse, F. Drenos, M. Beekman, J. Jukema, C. van Duijn and P. Slagboom. 'A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals.' In: *Nature Communications* 10 (Aug. 2019). DOI: https://doi.org/10.1038/s41467-019-11311-9.

[3] Q. D. Nguyen, E. M. Moodie, M.-F. Forget, P. Desmarais, M. R. Keezer and C. Wolfson. 'Health Heterogeneity in Older Adults: Exploration in the Canadian Longitudinal Study on Aging'. In: *Journal of the American Geriatrics Society* 69.3 (2021), pp. 678–687. DOI: https://doi.org/10.1111/jgs.16919. eprint: https://agsjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/jgs.16919. URL: https://agsjournals.onlinelibrary.wiley.com/doi/abs/10.1111/jgs.16919.

[4] J. Jylhävä, N. L. Pedersen and S. Hägg. 'Biological Age Predictors'. In: (2017). DOI: http://dx.doi.org/10.1016/j.ebiom.2017.03.046.

[5] H. S.-H. Oh, J. Rutledge, D. Nachun, R. Pálovics, P. M.-L. Olamide Abiose and, D. Channappa, D. Y. Urey, K. Kim, Y. J. Sung, L. Wang, J. Timsina, D. Western, M. Liu, P. Kohlfeld, J. Budde, E. N. Wilson, Y. Guen, T. M. Maurer, M. Haney, A. C. Yang, Z. He, M. D. Greicius, K. I. Andreasson, S. Sathyan, E. F. Weiss, S. Milman, N. Barzilai, C. Cruchaga, A. D. Wagner, E. Mormino, B. Lehallier, V. W. Henderson, F. M. Longo, S. B. Montgomery and T. Wyss-Coray. 'Organ aging signatures in the plasma proteome track health and disease. ' In: *Nature* 624 (2023), pp. 164–172. DOI: 10.1038/s41586-023-06802-1. URL: https://www.nature.com/articles/s41586-023-06802-1#citeas.

[6] J. Rutledge, H. Oh and T. Wyss-Coray. 'Measuring biological age using omics data'. In: *Nat Rev Genet* 23 (2022), pp. 715–727. DOI: 10.1038/s41576-022-00511-7. URL: www.nature.com/nrg.

[7] 'Ten years of brainage as a neuroimaging biomarker of brain aging: What insights have we gained?' In: *Frontiers in Neurology* 10.JUL (2019). ISSN: 16642295. DOI: 10.3389/fneur.2019.00789.

[8] J. H. Cole, R. P. K. Poudel, D. Tsagkrasoulis, M. W. A. Caan, C. Steves, T. D. Spector and G. Montana. 'Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker'. In: (2017). DOI: 10.1016/j.neuroimage.2017.07.059.

[9] J. Wang, M. J. Knol, A. Tiulpin, F. Dubost, M. De Bruijne, M. W. Vernooij, H. H. Adams, M. A. Ikram, W. J. Niessen and G. V. Roshchupkin. 'Gray matter age prediction as a biomarker for risk of dementia'. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.42 (Oct. 2019), pp. 21213–21218. ISSN: 10916490. DOI: 10.1073/PNAS.1902376116/-/DCSUPPLEMENTAL.

[10] B. A. Jonsson, G. Bjornsdottir, T. E. Thorgeirsson, L. M. Ellingsen, G. B. Walters, D. F. Gudbjartsson, H. Stefansson, K. Stefansson and M. O. Ulfarsson. 'Brain age prediction using deep learning uncovers associated sequence variants'. In: (). DOI: 10.1038/s41467-019-13163-9.

[11] V. M. Bashyam, G. Erus, J. Doshi, M. Habes, I. M. Nasrallah, M. Truelove-Hill, D. Srinivasan, L. Mamourian, R. Pomponio, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, S. C. Johnson, J. Fripp, N. Koutsouleris, T. D. Satterthwaite, D. Wolf, R. E. Gur, R. C. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. Resnick, R. Nick Bryan, D. A. Wolk, H. Shou and C. Davatzikos. 'MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide'. In: (). ISSN: 2312-2324. DOI: 10.1093/brain/awaa160.

[12] J. Cole, S. Ritchie, M. Bastin, M. Valdés Hernández, S. Muñoz Maniega, N. Royle, J. Corley, A. Pattie, S. Harris, Q. Zhang, N. Wray, P. Redmond, R. Marioni, J. Starr, S. Cox, J. Wardlaw, D. Sharp and I. Deary. 'Brain age predicts mortality'. In: *Molecular Psychiatry* 23 (2018), pp. 1385–1392. DOI: 10.1038/mp.2017.62.

[13] K. Franke, C. Gaser, B. Manor, V. Novak, P. H. Reddy, N. Koutsouleris and S. Brain. 'Advanced BrainAGE in older adults with type 2 diabetes mellitus'. In: (2013). DOI: 10.3389/fnagi.2013.00090.

[14] E. B. Van Den Akker, S. Trompet, H. J. J. Barkey Wolf, M. Beekman, ; H. Eka, D. Suchiman, J. Deelen, ; Folkert, W. Asselbergs, ; Bbmri-Nl, E. Boersma, ; Davy, P. M. Elders, ; J. M. Geleijnse, ; M. Arfan Ikram, M. Kloppenburg, H. Mei, I. Meulenbelt, S. P. Mooijaart, ; Rob, G. H. H. Nelissen, M. G. Netea, B. W. J. H. Penninx, M. Slofstra, ; Coen, D. A. Stehouwer, M. A. Swertz, C. E. Teunissen, G. M. Terwindt, ; Leen, M. 't Hart, ; Arn, M. J. M. Van Den Maagdenberg, ; Pim Van Der Harst, C. C. Van Der Horst, ; Carla, J. H. Van Der Kallen,

M. M. J. Van Greevenbroek, ; W. Erwin Van Spil, C. Wijmenga, A. Zhernakova, A. H. Zwinderman, N. Sattar, ; J. Wouter Jukema, C. M. Van Duijn, ; Dorret, I. Boomsma, M. J. T. Reinders and ; P. E. Slagboom. 'Metabolic Age Based on the BBMRI-NL H-NMR Metabolomics Repository as Biomarker of Age-related Disease'. In: *Circ Genom Precis Med* 13 (2020), p. 2610. DOI: 10.1161/CIRCGEN.119.002610. URL: https://www.ahajournals.org/doi/suppl/10.1161/CIRCGEN.119.002610..

[15] J. Hertel, N. Friedrich, K. Wittfeld, M. Pietzner, K. Budde, S. Van Der Auwera, T. Lohmann, A. Teumer, H. V"o, M. Nauck and H. J"o Rgen Grabe. 'Measuring Biological Age via Metabonomics: The Metabolic Age Score'. In: (2015). DOI: 10.1021/acs.jproteome.5b00561.

[16] O. Robinson, M. Chadeau Hyam, I. Karaman, R. Climaco Pinto, M. Ala-Korpela, E. Handakas, G. Fiorito, H. Gao, A. Heard, M.-R. Jarvelin, M. Lewis, R. Pazoki, S. Polidoro, I. Tzoulaki, M. Wielscher, P. Elliott, P. Vineis and C. Oliver Robinson. 'Determinants of accelerated metabolomic and epigenetic aging in a UK cohort'. In: *Aging Cell* 19 (2020). DOI: 10.1111/acel.13149.

[17] S. I. Huisman, A. T. J. Van Der Boog, F. Cialdella, J. J. C. Verhoeff and S. David. 'Quantifying the post-radiation accelerated brain aging rate in glioma patients with deep learning'. In: (2022). DOI: 10.1016/j.radonc.2022.08.002.

[18] H. B. McMahan, D. R. Eider Moore, S. Hampson and B. A. y Arcas. 'Communication-Efficient Learning of Deep Networks from Decentralized Data'. In: *International Conference on Artificial Intelligence and Statistics (AISTATS)* 54 (2017), p. 10.

[19] M. A. Ikram, B. C. Kieboom, W. P. Brouwer, G. Brusselle, L. Chaker, M. Ghanbari, A. Goedegebure, M. K. Ikram, M. Kavousi, R. J. de Knegt, A. I. Luik, J. van Meurs, L. M. Pardo, F. Rivadeneira, F. J. van Rooij, M. W. Vernooij, T. Voortman and N. Terzikhan. 'The Rotterdam Study. Design update and major findings between 2020 and 2024'. In: *European journal opf epidemiology* 39 (2024), pp. 183–206.

[20] M. T. Schram, S. J. Sep, C. J. Van Der Kallen, P. C. Dagnelie, A. Koster, N. Schaper, R. M. Henry and C. D. Stehouwer. 'The Maastricht Study: an extensive phenotyping study on determinants of type 2 diabetes, its complications and its comorbidities'. In: *European journal of epidemiology* 29 (2014), pp. 439–451.

[21] M. Schoenmaker, A. J. de Craen, P. H. de Meijer, M. Beekman, G. J. Blauw, P. E. Slagboom and R. G. Westendorp. 'Evidence of genetic enrichment for exceptional survival using a family approach: the Leiden Longevity Study'. In: *European Journal of Human Genetics* 14.1 (2006), pp. 79–84.

[22] B. Fischl. 'FreeSurfer'. In: *Neuroimage* 62.2 (2012), pp. 774–781.

[23] B. Fischl, D. H. Salat, A. J. Van Der Kouwe, N. Makris, F. Ségonne, B. T. Quinn and A. M. Dale. 'Sequence-independent segmentation of magnetic resonance images'. In: *Neuroimage* 23 (2004), S69–S84.

[24] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston and R. S. Frackowiak. 'A voxel-based morphometric study of ageing in 465 normal adult human brains'. In: *Neuroimage* 14.1 (2001), pp. 21–36.

[25] G. V. Roshchupkin, H. H. Adams, S. J. van der Lee, M. W. Vernooij, C. M. van Duijn, A. G. Uitterlinden, A. van der Lugt, A. Hofman, W. J. Niessen and M. A. Ikram. 'Fine-mapping the effects of Alzheimer's disease risk loci on brain morphology'. In: *Neurobiology of aging* 48 (2016), pp. 204–211.

[26] H. A. Vrooman, C. A. Cocosco, F. van der Lijn, R. Stokking, M. A. Ikram, M. W. Vernooij, M. M. Breteler and W. J. Niessen. 'Multi-spectral brain tissue segmentation using automatically trained k-Nearest-Neighbor classification'. In: *Neuroimage* 37.1 (2007), pp. 71–81.

[27] A. Moncada-Torres, F. Martin, M. Sieswerda, J. Van Soest and G. Geleijnse. 'VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange'. In: *AMIA annual symposium proceedings*. Vol. 2020. American Medical Informatics Association. 2020, p. 870.

[28] P. Mateus, J. Moonen, M. Beran, E. Jaarsma, S. M. van der Landen, J. Heuvelink, M. Birhanu, A. G. Harms, E. Bron, F. J. Wolters, D. Cats, H. Mei, J. Oomens, W. Jansen, M. T. Schram, A. Dekker and I. Bermejo. 'Data harmonization and federated learning for multi-cohort dementia research using the OMOP common data model: A Netherlands Consortium of Dementia Cohorts case study'. In: *Journal of Biomedical Informatics* (2024), p. 104661. ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2024.104661.

[29] D. S. Marcus, T. R. Olsen, M. Ramaratnam and R. L. Buckner. 'The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data'. In: *Neuroinformatics* 5 (2007), pp. 11–33.

[30] A.-M. G. de Lange, M. Anatürk, J. Rokicki, L. K. Han, K. Franke, D. Alnæs, K. P. Ebmeier, B. Draganski, T. Kaufmann, L. T. Westlye *et al.* 'Mind the gap: Performance metric evaluation in brain-age prediction'. In: *Human Brain Mapping* 43.10 (2022), pp. 3113–3129.

[31] A.-M. G. De Lange, M. Anatürk, S. Suri, T. Kaufmann, J. H. Cole, L. Griffanti, E. Zsoldos, D. E. Jensen, N. Filippini, A. Singh-Manoux *et al.* 'Multimodal brain-age prediction and cardiovascular risk: The Whitehall II MRI sub-study'. In: *NeuroImage* 222 (2020), p. 117292.

[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.* 'Tensorflow: Large-scale machine learning on heterogeneous distributed systems'. In: *arXiv preprint arXiv:1603.04467* (2016).

[33] X. Glorot and Y. Bengio. 'Understanding the difficulty of training deep feedforward neural networks'. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 249–256.

[34] I. Docker. *Docker: Accelerated Container Application Development*. https://www.docker.com/. 2024.

[35] J. Mongan, L. Moy and C. Kahn. *Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2: e200029*. 2020.

[36] P. Würtz, A. J. Kangas, P. Soininen, D. A. Lawlor, G. Davey Smith and M. Ala-Korpela. 'Quantitative Serum Nuclear Magnetic Resonance Metabolomics in Large-Scale Epidemiology: A Primer on -Omic Technologies'. In: *American Journal of Epidemiology* 186.9 (May 2017), pp. 1084–1096. ISSN: 0002-9262. DOI: 10.1093/aje/kwx016.

[37] G. V. Roshchupkin, H. Adams, M. Vernooij, A. Hofman, C. Van Duijn, M. Ikram and W. Niessen. 'HASE: Framework for efficient high-dimensional association analyses'. In: *Scientific reports* 6.1 (2016), p. 36076.

[38] L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza and A. Mechelli. 'Machine learning for brain age prediction: Introduction to methods and clinical applications'. In: *EBioMedicine* 72 (2021).

[39] L. S. Kumari and R. Sundarrajan. 'A review on brain age prediction models'. In: *Brain Research* 1823 (2024), p. 148668.

[40] E. W. Steyerberg and F. E. Harrell Jr. 'Prediction models need appropriate internal, internal-external, and external validation'. In: *Journal of clinical epidemiology* 69 (2016), p. 245.

[41] A.-M. G. de Lange, M. Anatürk, J. Rokicki, L. K. M. Han, K. Franke, D. Alnæs, K. P. Ebmeier, B. Draganski, T. Kaufmann, L. T. Westlye, T. Hahn and J. H. Cole. 'Mind the gap: Performance metric evaluation in brain-age prediction'. In: *Human Brain Mapping* 43.10 (2022), pp. 3113–3129. DOI: https://doi.org/10.1002/hbm.25837.

[42] B. Zhang, S. Zhang, J. Feng and S. Zhang. 'Age-level bias correction in brain age prediction'. In: *NeuroImage: Clinical* 37 (2023), p. 103319.

[43] M. H. Zonneveld, N. A. Kuhaili, S. P. Mooijaart, P. E. Slagboom, J. W. Jukema, R. Noordam and S. Trompet. 'Increased 1H-NMR metabolomics-based health score associates with declined cognitive performance and functional independence in older adults at risk of cardiovascular disease'. In: *medRxiv* (2023). DOI: 10.1101/2023.12.21.23300037.

[44] T. Huurre, H. Aro, O. Rahkonen and E. K. and. 'Health, lifestyle, family and school factors in adolescence: predicting adult educational level'. In: *Educational Research* 48.1 (2006), pp. 41–53. DOI: 10.1080/00131880500498438.

[45] L. Koivusilta, A. Rimpelä and M. Rimpelä. 'Health related lifestyle in adolescence predicts adult educational level: a longitudinal study from Finland.' In: *Journal of Epidemiology & Community Health* 52.12 (1998), pp. 794–801. ISSN: 0143-005X. DOI: 10.1136/jech.52.12.794. eprint: https://jech.bmj.com/content/52/12/794.full.pdf. URL: https://jech.bmj.com/content/52/12/794.

[46] E. Haddad, F. Pizzagalli, A. H. Zhu, R. R. Bhatt, T. Islam, I. Ba Gari, D. Dixon, S. I. Thomopoulos, P. M. Thompson and N. Jahanshad. 'Multisite test–retest reliability and compatibility of brain metrics derived from FreeSurfer versions 7.1, 6.0, and 5.3'. In: *Human Brain Mapping* 44.4 (2023), pp. 1515–1532.

[47] T. M. Doring, T. T. Kubo, L. C. H. Cruz Jr, M. F. Juruena, J. Fainberg, R. C. Domingues and E. L. Gasparetto. 'Evaluation of hippocampal volume based on MR imaging in patients with bipolar affective disorder applying manual and automatic segmentation techniques'. In: *Journal of Magnetic Resonance Imaging* 33.3 (2011), pp. 565–572.

9.5. Addendum: Practical lessons learned from NCDC useCase 1

Besides the research questions that were answered in the previous chapter, the project was one of the first to set up a federated infrastructure across three academic medical centers within the Netherlands. This addendum describes some lessons learned from creating this federated infrastructure and algorithms, both on practical and legal aspects.

9.5.1. Implementing Federated Learning

As mentioned in the previous chapter (supplementary 1), the infrastructure distinguishes between developers and cohort owners. The developer is responsible for creating and testing the algorithms, which the cohort owner then checks and whitelists on their local node, such that the developer can run the federated analysis.

This workflow was straightforward while guaranteeing the agency of cohort owners on which algorithms would be run on their data. However, due to the collaborative nature of this setup, the development cycle was significantly slower compared to algorithm development for local data. Every development iteration would consist of creating a new version, including a new docker image and github release, sending links for both to the cohort owners by email, waiting for confirmation of each cohort owner that they checked and whitelisted the algorithm, before being able to run the analysis to see whether it worked on the federated infrastructure. While each cohort owner always responded swiftly to these requests, having such an 'outer loop' slows down the development process.

In order to alleviate this issue, a local test bed which ran three nodes with simulated mock data was created. Using this test bed helped to avoid most bugs locally before sending out the tested code to the system administrators. However, for such a test bed to be effective, the mock data has to be a good representation of the local data at each cohort. Because there was no central overview of the exact data structure at each center (this data was kept local, after all), the initial test bed itself was not accurate.

One example that comes to mind was when an error on the deployed system could not be replicated in the test bed. Luckily, the system administrator at the node where the error occurred quickly found the problem: In their cohort, patients who had multiple visits (one to have their blood drawn, and another to have their MRI scan made) would have two separate entries in the database. The fields applicable to the measurement that did not align with the one taken at that specific visit would then be left empty, appearing as NaN's in the dataset.

This example illustrates the problem of developing code without access to the data. While a local test bed helps in debugging most minor issues such as syntax errors, some assumptions on the local data structure have to be made, either implicitly or

explicitly (the fact that multiple visits could be added in separate entries was not even considered). These 'unknown unknowns' could only be found by trial and error on the federated infrastructure, which required the aforementioned slow outer loop for development.

With the power of hindsight, there are various improvements that could be made on this process. One option is that node owners also run tests on their local node before whitelisting. The test bench could be provided by the developer, such that system administrators only have to run tests before approval. Although this might be more work for the system administrators upfront, it could save time in the long run by requiring less test cycles, and therefore less whitelisting. However, this approach would scale poorly to more nodes, as now each nodes' administrator needs to test separately before giving approval, potentially extending the time a debug cycle can take.

Another solution would be to increase the quality of local test beds. By sharing heavily noised data, a developer could gain an understanding of the data structure at each cohort. Furthermore, software tools that allow for the exploration of distributed data are still emerging, and were still in a very early stage during the project itself. With better software tools, it could be possible to create a better test bed locally, decreasing the amount of bugs that have to be discovered on the deployed system. However, the problem of unknown unknowns persists, and there could always be peculiarities of the deployed system that were missed in creating the test bed.

Perhaps the best solution is a combination of the described solutions: more sophisticated federated learning software that allows for testing on the federated system. Platforms could allow for a 'test mode', in which algorithms can be sent that are not whitelisted, but will only send back error and debug logs, without a return value. However, implementing this is not straightforward. For example, how to prevent a malicious developer to copy private data into debug logs? However, if such a functionality would be implemented, it would drastically speed up the process of developing federated analyses, and further mature the federated paradigm.

9.5.2. Legal Aspects

The main potential of federated learning in healthcare research is to accelerate multi-institutional research. By ensuring data does not need to be shared, medical institutions can avoid legal processes such as setting up a data sharing agreement, which can take multiple years. However, while data sharing agreements can be avoided, some level of legal agreement is still required. These agreements need to be significantly quicker to set up in order to reach the speed-up that federated learning promises.

In the previous chapter, there was a data access agreement signed by all parties. Algorithms that are developed are accessing the data, and therefore a data access

agreement seemed like the most logical option. However, the process of creating a data access agreements is, though not as lengthy as data sharing agreements, still long and tedious. Most of the time spent waiting for the legal agreements could be used for testing on mock data, but the legal procedure was still a major part of the project's lifetime. Therefore, one can wonder how much time was really saved by using a federated learning approach.

To a large extent, the extensive legal paperwork required for this project can be attributed to a lack of precedent. We were the first federated infrastructure that was created in two out of three of the collaborating institutes, meaning that this was the first time such a system was assessed by privacy officers and legal teams of these institutes. Now that such a document is established, it can be used as a template for future collaborations, which might speed up the process.

Furthermore, no specific agreements or legislature surrounding federated learning currently exists. The data access agreement was adapted from the setting in which a *researcher* from a different institute would have to access the data. In our case, only an algorithm would be granted access to the data, and the researcher would only see the output of the algorithm, which is highly curated due to the whitelisting policy described earlier. Arguably, such a setup requires less strict legal boundaries. However, due to a lack of precedent and legislature, legal documents tailored to federated learning do not exist (yet). Such documents could speed up the legal cycle as well.

Finally, the legal landscape surrounding data access is changing. In the European Union, the European Health Data Space (EHDS) regulation [12] enforces hospitals to make their health data available to a Health Data Access Body, which is locally responsible for distributing such anonimized data for research, as well as to supply metadata to a European metadata catalog. Since the regulation is still young, implementation can still take several years, so it is currently hard to say what the implications for federated learning approaches are. Nevertheless, it indicates a shift in the legal landscape to a more open attitude towards medical data sharing for research purposes, which should aid in the acceptance of federated learning as well.

9.5.3. Concluding remarks

Taken together, in the current environment, practical considerations such as code development and legal hurdles prevent federated learning from reaching its full potential in speeding up collaborative science. However, these hurdles can be attributed to the relative newness of the federated paradigm on the one hand, with developments on the software platforms. The version of the federated platform that was used in this project (vantage6) was 1.8. At the time of writing, the most up-to-date version is 4.13. On the other hand, with legal precedents, these hurdles can be drastically reduced in the near future. Meanwhile, sharing data does not get easier; legislation such as the GDPR in Europe form a solid framework for digital privacy

protection, which does not seem to be getting less strict any time soon. This means that federated learning can still fulfill its promise of speeding up collaborative health science in the near future, given enough effort on both the software implementations and legal aspects.

10
Discussion

In this part of the thesis, we focus on two main research questions: What is the optimal approach for creating a proteomics-based ageing biomarker, and how do biomarkers which focus on different aspects of the ageing process, that are based on different data modalities, compare to each other? To answer these questions, we characterize various aspects of aging biomarkers through comparison of existing and the design of novel biomarkers of aging. We investigated the optimal approach for creating a proteomics-based ageing biomarker, for which we also explored whether frailty is an effective training target for ageing biomarkers. Furthermore, we also investigated the connection between metabolomics-based and brain MRI scan-based ageing biomarkers.

In chapter 7, we first explore various aspects of proteomics ageing biomarker design. By investigating the effect of input data, model choice and training target, we introduce novel proteomics based aging biomarkers, ProtFI and ProtMort. We then compare ProtFI and ProtMort with existing aging biomarkers, such as the Protein Aging Clock, ProteinAge and MetaboHealth. In particular, ProtFI shows strong associations with various ageing related health outcomes, indicating that frailty could be a promising training target for aging biomarkers. Furthermore, even though both ProtFI and ProtMort were trained on a minimal set of proteins (27 and 20 proteins, respectively, selected using a forward feature selection procedure), their associations with ageing related health outcomes did not lower significantly compared to biomarkers based on vastly more proteins, indicating that a minimal yet carefully selected set of proteins can retain the biological information relevant for ageing.

In chapter 8 we expand on the analyses performed in chapter 7 by introducing two multitask models fmNN and fgNN. Although neither of these models showed consistently higher associations with health outcomes compared to their single-task counterparts, these associations did seem to be more robust, indicating that using multitask models might be a promising direction for ageing biomarkers. Specifically, analyses of fgNN indicate that, for these multitask models, using more proteins might improve their results. This hypothesis is contrary to our findings in chapter 7, which highlights the need for more investigation into different modeling approaches for ageing biomarkers.

Then, in chapter 9, we compare two metabolomics-based ageing biomarkers, MetaboAge and MetaboHealth, and an MRI-based biomarker, BrainAge. We do so by using a federated learning approach, where the biomarker training and analyses were run across three cohorts without sharing any data. We first employ our federated infrastructure to retrain the BrainAge biomarker, showing that the federated version increases model performance on unseen data. Furthermore, our results show that MetaboAge and BrainAge carry non-overlapping information, indicating that these markers indeed only capture a narrow part of the ageing process, while MetaboHealth captures a broader spectrum of ageing related aspects.

10.1. Association versus causality when studying aging biomarkers

With the emergence of aging biomarkers, certain methodological development patterns have become the norm. For instance, validation tends to consist of association analyses with endpoints representing the widespread physiological decline known to be connected to the aging process, such as loss of grip strength and muscle loss, insulin resistance, cognitive decline and disease endpoints such as type 2 diabetes [1]. In fact, in chapter 7 we argue that these associations are at least equally important as goodness-of-fit metrics such as the C-index or the R^2 score.

However, it is important to note that associations of these endpoints with molecular profiles may hint at pathways contributing to such associations but by no means in a causal and mechanistic way. Besides, with the plethora of available metadata in large data banks such as the UK Biobank, it can be tempting to extend such association analyses to endpoints whose connection to aging are less well studied, in an attempt to learn about novel connections to ageing. However, exploring such connections in a data driven way should be performed with caution. There are various factors known to be highly confounding with the aging process, such as socio-economic status [2]. Furthermore, the directionality of found associations can not be determined using straightforward association analyses, although this remains a popular choice [3]. Therefore, those analyses can only serve as exploratory, and more in-depth causal methods such as Mendelian randomization [4], as well as longitudinal and experimental study design, are required for discovering more biologically robust connections.

10.2. Combining aging biomarkers

In chapter 9 we show that biomarkers trained on different modalities, aimed at characterizing different aspects of aging, carry non-overlapping information. This raises the question whether it is possible to create a single, holistic biomarker capable of incorporating information from various modalities.

Multi-modal integration can be split between early and late integration [5]. Most approaches for multi-modal aging biomarkers apply early integration: By combining data features from multiple modalities, these approaches create one single model aimed at characterizing a broader view of aging aspects. However, most recent attempts at early integration in aging biomarker research show little to no improvement over single-modality models ([6], chapter 7).

Alternatively, multi-modal aging biomarkers that utilize late integration use multiple models trained on a single data modality, with the aim to characterize a narrow part of aging, such as cognitive -or cardiovascular health decline. The predictions of these models can then be pooled in order to create a single aging score. The advantage of such an approach is that the underlying model scores can be investigated to better

interpret the combined score. However, current literature on aging biomarkers using late integration is quite sparse. More generally, it is interesting to study the best approach to aggregate models.

10.3. Deep learning strategies for aging biomarkers

Due to their ability of modeling nonlinear relations, deep learning models have the potential to create more accurate aging biomarkers. However, in chapter 7 we show that a deep learning model does not result in a better aging biomarker compared to a simpler elasticNet model, which is in line with earlier literature [7]. This raises the question what would be required to create effective deep learning-based aging biomarkers.

Using deep learning in aging biomarkers is related to the open problem in deep learning research of creating effective models for tabular data [8]. Deep learning models excel in finding patterns in for example images or text, because these data modalities have an underlying structure that the model can exploit. Since tabular data does not have such a structure, deep learning models tend to underperform compared to simpler models such as elasticNet or XGBoost [9]. While many attempts have been made at creating architectures specifically for tabular data [10], or inducing structure within the tabular data itself [11], these approaches seem to have little effect [8]. The methods inducing structure in tabular data are, however, mostly agnostic what dat is used. With -omics data, there are underlying biological mechanisms that relate molecular features to one another, both within and across -omics layers. Exploiting these biological relations to create structural data from -omics data, for example by constructing interaction networks, could unlock the potential of deep learning models for aging biomarkers. While such approaches are emerging [12], many questions, such as how to optimally construct which (combination of) -omics data for aging biomarkers, remain unanswered.

10.4. The promise of longitudinal analyses

Due to the abundance of cross-sectional -omics data, most current aging biomarkers use data from a single timepoint per participant. Since cross-sectional -omics datasets are tabular, it is understandable that deep learning models underperform on these datasets. In order to benefit from the added complexity of deep learning models, more structured -omics data would be required. Longitudinal data could serve as a means to induce temporal structure within the data which would allow deep learning architectures, such as transformers, to thrive. Furthermore, certain biological measurements, such as metabolomics and brain scans, can be susceptible to short term variability [13] [14]. By incorporating longitudinal measurements, a rate of aging can be estimated, which can be a more robust measure, leading to better model training. Furthermore, such aging rates carry more information than a single timepoint, which can result in stronger aging biomarkers. However, creating such longitudinal data requires a larger effort in data collection, which could alternatively be used for

creating larger cross-sectional databases. Finding the right balance in importance of cross-sectional versus longitudinal data collection for aging biomarker development is still an open question, and requires more investigation in aging biomarkers based on longitudinal data. Nevertheless, blood-based longitudinal biomarkers can still be limited in their biological interpretation, as blood is less informative on immune metabolic ageing compared to e.g. muscle or fat tissue. Furthermore, although short term variability could better be accounted for using longitudinal data, blood draws are usually taken in a fasted state instead of a challenged state of the cell and molecular system, further limiting biological interpretation.

10.5. The importance of federated learning for aging biomarker improvement

Until recently, there was little incentive to align data collection between medical institutes. Medical data is highly private and can often not be shared, so there used to be little incentive to unify data collection standards across medical institutes. However, federated learning enables cross-institute analyses, unlocking more data-intensive modeling approaches such as described in sections 10.3 and 10.4.

However, a unified data standard is required for efficient federated algorithm development. The FAIR data standard [15] would address a lot of practical challenges in federated learning, such as some of the challenges described in addendum 9.5. However, transforming existing data into FAIR databases is a massive undertaking, which raises the question whether creating FAIR data speeds up the federated learning process at all. Finding more efficient ways of FAIR-ifying data, or developing data standards specifically for federated learning purposes (i.e. harmonized, both semantically and technically, easy to explore while remaining private, etc.), are developments required for enabling efficient federated learning.

10.6. Clinical translation of aging biomarkers

Aging biomarkers have the potential to detect accelerated aging early on, theoretically leading to more effective decision making in clinical populations and health checks to support prevention in population health [1]. However, despite the vast variety of available aging biomarkers, very few are properly clinically tested or used [3]. The input/output information determines the potential use of an aging biomarker. Therefore, the epidemiological data used in this thesis would mainly benefit prevention in population health. This requires a biomarker to present a number of characteristics: The molecular assay should be affordable, scalable, actionable and ideally not only depend on a blood draw at general practitioners, but also be accessible through finger pricks at other points of care. The overall use in prevention of expensive molecular assays (as is the case for proteomics at the moment) may therefore be limited. For assistance in decision making in clinical situations, more molecular data from patient populations is required, such that existing biomarkers can be tested, if

not finetuned, for their efficacy in the clinic.

Most aging biomarkers scores depend on the training population, whereas an individualized score is required for use in the clinic or in population health. Furthermore, most -omics measurement platforms have not been clinically validated, which means that the error margin on these measurements is too high to be used within the clinic [16]. An example on how to transform an aging biomarker into a potential health check in middle age along these lines concerns the MetaboHealth score [17]. MetaboHealth was personalized for individuals from Caucasian descent, clinically validated, and actionable: MetaboHealth has shown to respond beneficially to lifestyle changes, hence potentially useful for prevention in population health (based on personal communications; manuscript in preparation). While the proteomics-based markers introduced in chapter 7 might be even more informative on health (risks and changes), these steps would still have to be taken for translation into population health or the clinic.

10.7. Concluding Remarks

In the second part of this dissertation, we compared existing ageing biomarkers from different modalities using a federated framework, showing that these markers carry non-overlapping information. Furthermore, we structurally analyze how to design an ageing biomarker, resulting in two novel markers, which show promising associations with ageing related health outcomes.

With the current plethora of available ageing biomarkers in literature, the field would benefit from exploring the reproducibility of the reported risks which these biomarkers indicate in independent study populations. For now, these would mainly include community dwelling middle and oler aged individuals given the informative epidemiological large scale longitudinal studies. Particularly, exploration in the clinic where molecular biomarkers can assist in decision making requires various challenges to be overcome, including the collection of clinically relevant biomaterial and data at the point of decision making. With the advance of federated learning, such data collection could be expedited, ultimately accelerating clinical adoption of ageing biomarkers. Finally, it remains to be seen whether a singular ageing biomarker is capable of satisfying clinical purposes. Perhaps a suite of biomarkers, either integrated or combined post-hoc, is required to fully characterize the heterogeneous ageing process.

Bibliography

[1] M. Moqri, C. Herzog, J. R. Poganik, J. Justice, D. W. Belsky, A. Higgins-Chen, A. Moskalev, G. Fuellen, A. A. Cohen, I. Bautmans *et al.* 'Biomarkers of aging for the identification and evaluation of longevity interventions'. In: *Cell* 186.18 (2023), pp. 3758–3775.

[2] D. J. McMaughan, O. Oloruntoba and M. L. Smith. 'Socioeconomic status and access to healthcare: interrelated drivers for healthy aging'. In: *Frontiers in public health* 8 (2020), p. 231.

[3] M. Moqri, C. Herzog, J. R. Poganik, K. Ying, J. N. Justice, D. W. Belsky, A. T. Higgins-Chen, B. H. Chen, A. A. Cohen, G. Fuellen *et al.* 'Validation of biomarkers of aging'. In: *Nature medicine* 30.2 (2024), pp. 360–372.

[4] P. C. Haycock, S. Burgess, K. H. Wade, J. Bowden, C. Relton and G. Davey Smith. 'Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies'. In: *The American journal of clinical nutrition* 103.4 (2016), pp. 965–978.

[5] C. G. Snoek, M. Worring and A. W. Smeulders. 'Early versus late fusion in semantic video analysis'. In: *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 399–402.

[6] D. A. Gadd, R. F. Hillary, Z. Kuncheva, T. Mangelis, Y. Cheng, M. Dissanayake, R. Admanit, J. Gagnon, T. Lin, K. L. Ferber, H. Runz, C. N. Foley, R. E. Marioni, B. B. Sun and B. B. Team. 'Blood protein assessment of leading incident diseases and mortality in the UK Biobank'. In: *Nature Aging* 4.7 (July 2024), pp. 939–948. ISSN: 2662-8465. DOI: 10.1038/s43587-024-00655-7.

[7] M. A. Argentieri, S. Xiao, D. Bennett, L. Winchester, A. J. Nevado-Holgado, U. Ghose, A. Albukhari, P. Yao, M. Mazidi, J. Lv, I. Millwood, H. Fry, R. S. Rodosthenous, J. Partanen, Z. Zheng, M. Kurki, M. J. Daly, A. Palotie, C. J. Adams, L. Li, R. Clarke, N. Amin, Z. Chen and C. M. van Duijn. 'Proteomic aging clock predicts mortality and risk of common age-related diseases in diverse populations'. In: *Nature Medicine* 30 (9 2024), pp. 2450–2460. ISSN: 1546-170X. DOI: 10.1038/s41591-024-03164-7.

[8] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk and G. Kasneci. 'Deep Neural Networks and Tabular Data: A Survey'. In: *IEEE Transactions on Neural Networks and Learning Systems* 35.6 (2024), pp. 7499–7519. DOI: 10.1109/TNNLS.2022.3229161.

[9] R. Shwartz-Ziv and A. Armon. 'Tabular data: Deep learning is not all you need'. In: *Information Fusion* 81 (2022), pp. 84–90.

[10] S. Ö. Arik and T. Pfister. 'Tabnet: Attentive interpretable tabular learning'. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 8. 2021, pp. 6679–6687.

[11] Y. Zhu, T. Brettin, F. Xia, A. Partin, M. Shukla, H. Yoo, Y. A. Evrard, J. H. Doroshow and R. L. Stevens. 'Converting tabular data into images for deep learning with convolutional neural networks'. In: *Scientific reports* 11.1 (2021), p. 11325.

[12] Z.-P. Li, Z. Du, D.-S. Huang and A. E. Teschendorff. 'Interpretable deep learning of single-cell and epigenetic data reveals novel molecular insights in aging'. In: *Scientific Reports* 15.1 (2025), p. 5048.

[13] K. Kim, C. Mall, S. L. Taylor, S. Hitchcock, C. Zhang, H. I. Wettersten, A. D. Jones, A. Chapman and R. H. Weiss. 'Mealtime, temporal, and daily variability of the human urinary and plasma metabolomes in a tightly controlled environment'. In: *PloS one* 9.1 (2014), e86223.

[14] A. Trefler, N. Sadeghi, A. G. Thomas, C. Pierpaoli, C. I. Baker and C. Thomas. 'Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging'. In: *Neuroimage* 133 (2016), pp. 41–52.

[15] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne *et al.* 'The FAIR Guiding Principles for scientific data management and stewardship'. In: *Scientific data* 3.1 (2016), pp. 1–9.

[16] B. Ratitch, A. Trigg, M. Majumder, V. Vlajnic, N. Rethemeier and R. Nkulikiyinka. 'Clinical validation of novel digital measures: statistical methods for reliability evaluation'. In: *Digital Biomarkers* 7.1 (2023), pp. 74–91.

[17] J. Deelen, J. Kettunen, K. Fischer, A. van der Spek, S. Trompet, G. Kastenmüller, A. Boyd, J. Zierer, E. B. van den Akker, M. Ala-Korpela, N. Amin, A. Demirkan, M. Ghanbari, D. van Heemst, M. A. Ikram, J. B. van Klinken, S. P. Mooijaart, A. Peters, V. Salomaa, N. Sattar, T. D. Spector, H. Tiemeier, A. Verhoeven, M. Waldenberger, P. Würtz, G. D. Smith, A. Metspalu, M. Perola, C. Menni, J. M. Geleijnse, F. Drenos, M. Beekman, J. W. Jukema, C. M. van Duijn and P. E. Slagboom. 'A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals'. In: *Nature Communications* 10 (1 2019), pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-019-11311-9.

Epilogue

In this thesis, we showed that federated learning is robust to non-extreme label distributions, across a wide variety of model and data types. Furthermore, we introduced a novel algorithm for federated K-means clustering, which requires no additional finetuning for varying amounts of local clusters. Next, we explored various aspects of aging biomarker design. We showed that a biomarker trained to predict frailty on a minimal subset of proteins can effectively capture associations with aging. We did so by introducing and externally validating protFI and protMort, two elasticNet based aging biomarkers. Finally, we trained and compared the brain based BrainAge wiht the metabolomics-based MetaboAge and MetaboHealth aging clocks in a distributed environment, without sharing any private data. These comparisons showed that the aging clocks based on different modalities capture non-overlapping information, raising the question on how to best capture these different aspects of aging together. Federated learning promises to enable the use of more multi-cohort data from different modalities, thereby paving the way for using multi-modal or longitudinal data for more intricate aging biomarker design.

Acknowledgements

Four years might seem like a long time, but time flies when you're having fun. And having fun is a lot easier with good company, which I am grateful to say I had plenty of over the past four years.

To start off with my supervisory team: **Marcel**, if it weren't for the pleasant experience during my master thesis, I might never have considered starting a PhD in the first place. Your pragmatic approach and quick and pointy feedback have played a major role in how pleasant my PhD experience was. **Eline**, Thank you for all the biological and epidemiological insights, even though I was hesitant to come visit all the way in Leiden.

Doing federated learning is inherently collaborative, and so I have had the pleasure to work with many great people over the past four years. **Pedro, Jing, Esther, Leon, Davy & Alexander**, setting up a real federated project is a hard eneavour, but I think it might be even harder than what we experienced. Thank you for having made up such a great team. Furthermore, **Lieke**, you've set my expectations for future collaborations unrealistically high. It was a pleasure to work together. To the students I've supervised, **Krzysztof**, **Roy, Emíls, Vlad, Matei, Henry, Alex, Philip and Lohith**: You've taught me more than I could have taught you. Thank you for trusting me in those crucial stages of your studies.

I've come to appreciate all the friends I've made at the DBL lab. **Marunka**, thank you for your endless support and being the one that really keeps the group rolling. **Ruud**, thank you for all your technical support as well. I hope you're enjoying your well-deserved retirement. **Thomas, Jasmijn, Jana and Joana**: Thank you for all the feedback during and after biotalks. **Erik**, I've thoroughly enjoyed all our discussions at the coffee machine, most of which scientific. You usually give me valuable insight, even now that I have started at SURF. **Tamim**, I probably should have listened when you told me not to start a new project in the last five months of my contract, but then again, diamonds are made under pressure. **Stavros**,early on in my PhD, your taste of humor made me realize how fun my colleagues were. I wish you all the best in Amsterdam. **Amelia**, thank you for the generous invite to your wedding in Spain, it was an amazing trip. **Roy**, thank you for all the amazing stews at the Christmas potlucks. **Bianca**, good job keeping Paul in check. he needed it. **Timo**, I'm sad I haven't gotten to visit one of your concerts yet. I promise I'll visit one soon. **Lorenzo**, you made me realize there are Italians that are not picky on food, something I didn't know existed before. **Inez**, I'm still scared by how much better you got at skiing within one week. I hope to keep visiting the Alps with you and the group. **Ivan**, your care

for the people around you always stood out to me. **Bram**, the rate at which you can switch between deeply philosophical and deeply unserious conversation is astounding, and one of the reasons I enjoy your company so much. **Niek**, even though we might disagree on e-bike policy, I deeply admire your work ethic, and I have no doubt you are going to have a successful second half of your PhD, and beyond. **Kirti**, thank you for being my office mate for almost the entire four years, and my main supplier of pens and chargers whenever my messy head would forget something. **Sara**, I really appreciate our Wednesday afternoon runs, both for the exercise and the gossip. We should keep them going. **Sander**, I think you are underrated as a scientist; your input in biotalks and group discussions have always given me a new perspective or something to think about, even when it wasn't about my own work. **Stephanie**, I've always enjoyed our rugby chats, even though we might disagree on who the best team in the world is. **Colm**, I admire your incredible tenacity throughout your own PhD. Perhaps that is what makes you such a formidable BJJ fighter. **Yasin**, when I joined the group right after covid, there were little social events, something that you can take most of the credit for changing. Thank you for being the original organiser. **Mostafa**, I remember you coming up to me and explaining where we were having lunch in the first week of my contract. Thanks for pulling me into the group at that point. **Jasper**, in a world full of learning machines, I admire you sticking to your guns. I hope to watch you defend soon as well, you have worked incredibly hard for it. **Paul**, thank you for all the fruitful conversations on some of the world's most important topics: Science, politics and the most effective way to wrestle a polar bear. **Gabriel**, thank you for inviting me to your parents place on our way to Austria. Those sausages were amazing. **Daniyal**, your speech at Amelia and Chirag's wedding was one of the highlights of the evening for me. **Chengyao**, I appreciate our discussions on shared struggles with tabular data. **Gerard**, I sometimes wished you wrote scripts that took longer to complete, because I always enjoyed the chats and discussions we had in my office while you were waiting for your code to run. **Madelon**, you were an amazing office mate, and I appreciate that you still take the effort from time to time to come visit from the far east. **Alex**, I appreciate learning about China from you. **Daan**, your presentations were often as insightful as entertaining. **Robert-jan, Alejandro, Ombretta, Jonathan, Atilla, Aurora, Rickard**, thank you for all the fun times during our skiing trips, which I hope to attend more of. **Ojas**, I hope I don't scare you as much anymore as I did the first time you saw me. **Chirag**, I'm amazed by your breadth of interests, and seeming ability to be highly invested in all of them, all while working incredibly hard as well. **Yavuz, Thomas, Sander, Hesam and Myrthe**, thank you for all the fun chats during the borrels. I hope they come back soon.

Everyone that I had the pleasure of working with at YoungCB: **Ming, Lieke, Mostafa, Cathi, Architha, Vartika, Fabienne, Roula, Gabriel and Bram**, thank you for the great experiences and all the fun events we organised together.

The ones who stuck around from my studies: **Koen, Niels, Tom, Rob & Joseph**, our (half-)yearly trips make my year complete, even if it requires me relying on the muskrat facts every once in a while. Thank you for being my friends.

To all the people crazy enough to also be in the gym at 7am, among others: **Anna, Rohan, Tristan, Maria, Sonja and of course old man Frank**), thank you for making that gym such a great place, you've probably been more important to my productivity and sanity than you realize.

The people who have shared a house with me over these years: **Fin, Tom, Amadou** and above all **Arthur**, who managed to stick it out with me for the full four years and then some. Thank you for letting me rant over dinner about whatever had bothered me that day.

Last but not least: **Mam & pap**, jullie onvoorwaardelijke steun in alles wat ik maar kan verzinnen heeft me kunnen laten worden tot wie ik ben, en daar ben ik oneindig dankbaar voor.

Curriculum Vitæ

Swier Jurjen Frans Garst

08-05-1997	Born in Dirksland, the Netherlands.

Education

2022–2026	PhD computer science, Delft University of Technology, Delft
2019–2021	MSc electrical engineering, Delft University of Technology, Delft
2015–2018	BSc electrical engineering, Delft University of Technology, Delft

Professional Experience

2026 – current	**SURF, Amsterdam** Machine learning consultant

List of Publications

- S. Garst and M. Reinders. 'Federated k-means clustering'. In: *International Conference on Pattern Recognition*. Springer. 2024, pp. 107–122
- S. Garst, J. Dekker and M. Reinders. 'A comprehensive experimental comparison between federated and centralized learning'. In: *Database* 2025 (2025), baaf016
- P. Mateus, S. Garst, J. Yu, D. Cats, A. G. Harms, M. Birhanu, M. Beekman, P. E. Slagboom, M. Reinders, J. van der Grond *et al.* 'Multi-Cohort Federated Learning Shows Synergy in Mortality Prediction for MRI-Based and Metabolomics-Based Age Scores'. In: *Journal of Healthcare Informatics Research* 9.4 (2025), pp. 606–628
- S. Garst, L. Kuiper, E. van den Akker, N. v. d. Berg, M. Ghanbari, S. Mooijaart, M. Beekman, M. Reinders, E. Slagboom and J. van Meurs. 'ProtFI, an efficient frailty-trained proteomics-based biomarker of aging, robustly predicts age-related decline'. In: *medRxiv* (2025), pp. 2025–09