Bioinformatics, 2026, **42(1)**, btaf625
https://doi.org/10.1093/bioinformatics/btaf625
Advance Access Publication Date: 9 January 2026
Original Paper

OXFORD

Databases and ontologies

dGAMLSS: an exact, distributed algorithm to fit Generalized Additive Models for Location, Scale, and Shape for privacy-preserving population reference charts

Fengling Hu[1,*], Jiayi Tong[2,3], Margaret Gardner[4], Lifespan Brain Chart Consortium, Andrew A. Chen[5], Richard A.I. Bethlehem[6], Jakob Seidlitz[4,7,8,9], Hongzhe Li[10], Aaron Alexander-Bloch[4,7,8,9], Yong Chen[2,†], Russell T. Shinohara[1,11,†]

[1]Penn Statistics in Imaging and Visualization Endeavor (PennSIVE), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States
[2]Center for Health AI and Synthesis of Evidence (CHASE), Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States
[3]Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, United States
[4]Brain-Gene-Development Lab, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States
[5]Department of Public Health Sciences, Medical University of South Carolina, Charleston, SC 29425, United States
[6]Department of Psychology, University of Cambridge, Cambridge, United Kingdom
[7]Lifespan Brain Institute, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States
[8]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA 19104, United States
[9]Department of Child and Adolescent Psychiatry and Behavioral Science, The Children's Hospital of Philadelphia, Philadelphia, PA 19104, United States
[10]Center for Statistical Methods for Big Data, Department of Biostatistics, and Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States
[11]Center for Biomedical Image Computing and Analytics (CBICA), Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, United States

*Corresponding author. Penn Statistics in Imaging and Visualization Endeavor (PennSIVE), Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Dr, Philadelphia, PA 19104, United States. E-mail: fengling.hu@pennmedicine.upenn.edu.
[†]= equal contribution.

Associate Editor: Christina Kendziorski

Abstract

Motivation: There is growing interest in estimating population reference ranges across age and sex to better identify atypical clinically-relevant measurements throughout the lifespan. For this task, the World Health Organization recommends using Generalized Additive Models for Location, Scale, and Shape (GAMLSS), which can model non-linear growth trajectories under complex distributions that address the heterogeneity in human populations.
Fitting GAMLSS models requires large, generalizable sample sizes, especially for accurate estimation of extreme quantiles, but obtaining such multi-site data can be challenging due to privacy concerns and practical considerations. In settings where patient data cannot be shared, privacy-preserving distributed algorithms for federated learning can be used, but no such algorithm exists for GAMLSS.

Results: We propose distributed GAMLSS (dGAMLSS), a distributed algorithm that can fit GAMLSS models across multiple sites without sharing patient-level data. This includes specific considerations for the fitting of smooth functions at varying levels of communication efficiency. We demonstrate the effectiveness of dGAMLSS in constructing population reference charts across clinical, genomics, and neuroimaging settings and show that dGAMLSS is able to reproduce pooled reference charts and inference down to numerical differences.

Availability and implementation: An R package providing examples of the dGAMLSS algorithm, as well as functions for sharing and aggregating site-specific parameters, is available at https://github.com/hufengling/dGAMLSS.

1 Introduction

Population reference ranges have been widely used as a cornerstone in clinical medicine to quickly screen a large number of subject-specific measurements, including anthropometrics and laboratory assays, in order to identify atypical measurements that may warrant further clinical follow-up (O'Connor 1990, Jones and Barker 2008, Weir and Jan 2024). In recent decades, there has been a growing interest in modeling such reference ranges, conceptualized as population reference charts. This idea is inspired by pediatric growth charts, which highlight that different reference ranges may be applicable to individuals at different stages of life (Cole *et al.* 1995, WHO Multicentre Growth Reference Study Group 2006, Cole *et al.* 2009, Bethlehem *et al.* 2022).

In order to estimate such population reference charts, also called normative charts, the World Health Organization has

Received: 26 June 2025; **Revised:** 22 September 2025; **Accepted:** 2 November 2025

recommended the use of Generalized Additive Models for Location, Scale, and Shape (GAMLSS) as a robust and flexible framework for modeling non-linear growth trajectories (Rigby and Stasinopoulos 1996, 2005, Borghi *et al.* 2006, Stasinopoulos and Rigby 2007). GAMLSS extends generalized linear models (GLM) and generalized additive models (GAM) to allow outcome variables to follow a broad family of distributions, including skewed and heavy-tailed distributions as well as zero-inflated distributions (Hastie and Tibshirani 1986). GAMLSS also supports semi-parametric modeling of not only the mean of the outcome variable, but also the variance, skewness, and kurtosis of the outcome for any given explanatory variables. This capability allows for the prediction of individualized reference distributions and is essential in the context of reference charts, since sources of population-wide heterogeneity—domain shifts, intrinsic population differences, sampling mechanisms, and more—suggest that higher-order moments may vary with age, sex, and other demographic factors. In addition, this added flexibility requires fewer data assumptions and is thus more generalizable to novel phenotypes. Importantly, while GAMLSS models are highly flexible, the semi-parametric nature of GAMLSS-based population reference charts still allows for a high degree of model transparency, interpretability, and inference, features that are advantageous to the adoption of such reference charts in clinical settings.

Though GAMLSS models offer significant advantages in estimating population reference charts, large sample sizes may be required for fitting. This is especially true as the complexity of the model grows in terms of explanatory variables included, increased non-linearity of the smooth terms, and estimation of higher-order moments. In the context of reference charts, even larger sample sizes are necessary to accurately estimate the extreme quantiles that are essential for identifying atypical measurements, such as those below the 2.5 percentile or above the 97.5 percentile.

Ideally, if individual patient data could be easily shared across multiple institutions, reference charts fit on such data could be confidently applied across a large, representative sample from the general population. However, the ability to construct such charts is often limited by privacy and practical challenges, including governmental regulations such as the Health Insurance Portability and Accountability Act (HIPAA) or the General Data Protection Regulation (GDPR), institutional policies on sharing data across institutions, patient consent protocols for research studies, or even collaborations where investigators are interested in maintaining full control of their data. In response to the need for managing large-scale, multi-site data, there has been a rise in the establishment of distributed research networks across healthcare systems, where each healthcare system maintains control of its own data, but cooperates within the network to answer broad-ranging questions via distributed learning. Key examples include the Observational Health Data Sciences and Informatics (OHDSI), the National Patient-Centered Clinical Research Network (PCORnet), the Sentinel Initiative, and the NIH-funded Health Care Systems Research Collaboratory (Brown *et al.* 2013, Collins *et al.* 2014, Hripcsak *et al.* 2015, Platt *et al.* 2018).

In such distributed research network settings as well as smaller-scale multi-site collaborations where sharing participant-level data is impractical, it is crucial to develop privacy-preserving federated and distributed learning algorithms that only require summary-level statistics or aggregated data to fit the desired model in a distributed manner. While distributed algorithms have been proposed for various models, including generalized linear models and generalized additive models, no such algorithm has been proposed for fitting GAMLSS (Chu *et al.* 2013, Li *et al.* 2020, Kia *et al.* 2021, Pfitzner *et al.* 2021, Yin *et al.* 2021, Luo *et al.* 2022a).

To extend population reference charts to include more representative patient populations and provide more accurate reference quantile estimation for clinical biomarkers, we propose distributed GAMLSS (dGAMLSS). dGAMLSS allows for exact distributed fitting of fully-parametric and semi-parametric models for all GAMLSS family distributions. We demonstrate the applicability of dGAMLSS for building population reference charts in clinical, genomics, and neuroimaging settings for outcomes drawn from various GAMLSS family distributions (Bethlehem *et al.* 2022, Muller *et al.* 2022, Johnson *et al.* 2023). As big data from electronic health records, clinical genomics, and quantitative neuroimaging become more available, representative population reference charts based on such data may enable the development of quantitative screening biomarkers for health and disease.

2 Materials and methods

2.1 Pooled GAMLSS framework

We first describe the pooled GAMLSS framework. GAMLSS-family distributions are a generalized set of probability distributions defined by at least two and up to four parameters. The first two parameters pertain to the mean and variance, while the third and fourth parameters describe skewness and kurtosis, respectively. Such distributions can be continuous, discrete, or mixed distributions, such as zero-inflated distributions, allowing for likelihood-based modeling of a wide range of outcome variables.

In the pooled GAMLSS setting where all patient-level data is collected across sites and accessed simultaneously, pooled data across all i sites are notated using a dot, which represents the union of matrices over all levels of the index. In GAMLSS, each $n \times 1$ vector of subject-specific distribution parameters, $\boldsymbol{\theta}_{\cdot 1}, \ldots, \boldsymbol{\theta}_{\cdot p}$, are modeled by additive terms, potentially under different monotonic link functions, $g_k(\cdot)$, and differing sets of covariates (Hastie and Tibshirani 1986). Thus, for each distribution parameter $k = 1, \ldots, p$, the semi-parametric parameter-specific GAMLSS model is defined as the following GAM:

$$g_k(\boldsymbol{\theta}_{\cdot k}) = \boldsymbol{\eta}_{\cdot k} = \mathbf{X}_{\cdot k}\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} \mathbf{Z}_{\cdot k}\boldsymbol{\gamma}_k$$

and the goal is to estimate each of the $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ in order to maximize the overall likelihood of observing all subjects' outcomes given their subject-specific predicted distributions. This model generalizes the GLM to distributions outside of the exponential family and generalizes the GAM such that all parameters, including mean, variance, skewness, and kurtosis, can be modeled in terms of both fixed and smooth terms.

The Rigby and Stasinopoulos (RS) algorithm consists of two cycles, which result in maximization of the pooled likelihood with respect to the fixed effect and smoothing coefficients. The outer cycle iterates across the p GAMLSS distribution family parameters. Meanwhile, the inner cycle

performs Newton-Raphson updates on the outer-cycle parameter-wise coefficients while keeping coefficients for all other parameters constant. In a pooled setting where all data are available, each multivariate Newton-Raphson update can be analytically solved using weighted least squares on the adjusted dependent vector $\mathbf{z}_{\cdot k}$ with diagonal weight matrix $\mathbf{W}_{\cdot k}$, where the underscore represents that the matrices are pooled across all m sites. The pooled RS algorithm is described in Appendix B and Appendix C by Rigby and Stasinopoulos in 2005 (Rigby and Stasinopoulos 2005).

2.2 Distributed GAMLSS via the RS algorithm

The distributed RS algorithm adapts the pooled RS algorithm, noting that $\mathbf{z}_{\cdot k}$ and the diagonal of $\mathbf{W}_{\cdot k}$ are easily horizontally partitioned into site-specific $\mathbf{z}_{ik}$ and $\mathbf{W}_{ik}$. Thus, updates can easily be distributed across sites—as long as each site i sends matrices M_{i1} and M_{i2} to a central site, the exact pooled weighted least squares solution can be obtained with one round of communication per inner cycle iteration (Table 1). Thus, dGAMLSS is an exact solution for federated learning of GAMLSS models.

Each inner cycle is complete when all parameter-wise coefficients have converged when compared to the prior Newton-Raphson update. The outer cycle is complete when all coefficients across all parameters have converged when compared to their values at the end of the previous outer cycle. Though pooled GAMLSS defines convergence via global deviance change, dGAMLSS defines convergence using coefficient values since global deviance estimation is delayed by one communication round. A given coefficient is considered to be converged when the proportion change between its current and previous value is smaller than some value, c. Larger values of c allow for faster, but less stable, convergence, while smaller values of c result in stable convergence but require more iterations. In pooled GAMLSS, c can be made arbitrarily small since additional iterations are processed internally on a central server, which is computationally cheap compared to transferring summary statistics across sites. In the proposed dGAMLSS algorithm, we set $c = 0.05$ to be well-balanced with respect to stability and communication cost. The exact distributed RS algorithm is provided in Algorithms 1 and 2. Additional details on technical details for smooth term fitting, knot placement, and interaction terms are provided in the Supplementary Data (Eilers and Mark 1996).

2.3 Distributed inference

Once the distributed RS algorithm has converged and coefficient point estimates are obtained, estimated GAMLSS quantiles can be easily obtained for the reference population as well as any given new data. For reference population quantiles, the central site can simulate data on a grid across all covariates. These simulated covariates can be combined with coefficient point estimates to obtain predicted distributions for each set of covariates, and reference quantiles can be obtained from these predicted distributions. Covariates from new data can be directly used to estimate predicted distributions, and quantiles for outcomes from new data can be estimated.

To perform inference, one additional round of post-fitting communication is necessary. The coefficient point estimates are sent to each site, and each site returns numerical site-wise likelihood-based Hessians, $\mathbf{H}_i$, simultaneously evaluated across all parameters. The global covariance matrix, $\mathbf{\Sigma} = \mathbf{H}^{-1} = \left(\sum_i^m \mathbf{H}_i\right)^{-1}$, is calculated by inverting the sum of site-specific Hessians. The diagonal of this covariance matrix is used in GAMLSS to obtain standard errors and t-statistics for each coefficient and, therefore, perform Wald-type inference. In this round of communication, site-specific deviances for the final model are also returned and summed to obtain the global deviance of the model.

For models using fixed effect smooth terms and fixed penalty smooth terms, inference on the overall effect of the spline can be performed via a likelihood ratio test comparing the global deviance of a full model with smooth terms to the global deviance of a reduced intercept-only model, where the null distribution of the likelihood ratio test is χ^2_{df}, where df is the number of columns of the fixed effect smooth term or the EDF of the fixed penalty smooth term being tested. Note that if multiple parameters contain smooth terms, inference for a given parameter requires a reduced intercept-only model for that parameter only; other parameters must still contain their full smooth terms. To reduce overall

Table 1. Relevant notation.

Site and subject	
m	Total number of sites.
i	Index of sites, $1,\ldots,m$.
n_i	Number of subjects in site i.
n	Total number of subjects across all sites. $n = \sum_{i=1}^m n_i$
Internal GAMLSS notation	
p	Total number of GAMLSS parameters depending on distribution, $p = 2, 3$, or 4. $p = 4$ corresponds to distributions with μ, σ, ν, and τ parameters.
k	GAMLSS parameter index, $1,\ldots,p$.
$\boldsymbol{\theta}_{ik}$	Fitted GAMLSS canonical parameter for site i for parameter k. $\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \boldsymbol{\theta}_{i3}$, and $\boldsymbol{\theta}_{i4}$ correspond to μ_i, σ_i, ν_i, and τ_i, respectively.
$\boldsymbol{\eta}_{ik}$	GAMLSS linear predictor from site i for parameter k, such that $g_k(\boldsymbol{\theta}_{ik}) = \boldsymbol{\eta}_{ik}$, where $g(\cdot)$ is the distribution-dependent canonical link function for the kth parameter.
GAMLSS Newton-Raphson update matrices	
$\boldsymbol{u}_{ik}$	First derivatives of subject-wise log likelihoods with respect to $\boldsymbol{\eta}_{ik}$, dimension $n_i \times 1$.
$\boldsymbol{W}_{ik}$	Diagonal matrix of second derivatives of subject-wise log likelihoods with respect to $\boldsymbol{\eta}_{ik}$, dimension $n_i \times n_i$.
$\boldsymbol{z}_{ik}$	Adjusted dependent variable from site i for parameter k. $\boldsymbol{z}_{ik} = \boldsymbol{\eta}_{ik} + \mathbf{W}_{ik}^{-1}\boldsymbol{u}_{ik}$.
Fixed effect and smooth term coefficients	
$\boldsymbol{\beta}_k$	Pooled fixed effect coefficients for parameter k.
$\boldsymbol{\gamma}_k$	(Optional) pooled smooth term coefficients for parameter k.
c	Convergence threshold for proportion change in each coefficient.
$\boldsymbol{H}_i$	Numerical Hessian at converged coefficients from site i.
$\boldsymbol{H}$	Pooled Hessian at converged coefficients.
Site-specific data	
$\boldsymbol{Y}_i$	Outcome vector from site i, dimension $n_i \times 1$.
$\boldsymbol{X}_{ik}$	Fixed effects design matrix from site i for parameter k, dimension $n_i \times \text{length}(\beta_k)$.
$\boldsymbol{Z}_{ik}$	(Optional) smooth design matrix from site i for parameter k, dimension $n_i \times \text{length}(\gamma_k)$.
Summary statistics sent from sites	
$\boldsymbol{M}_{i1}$	Update matrix from site i for parameter k with dimension $p \times p$.
$\boldsymbol{M}_{i2}$	Update matrix from site i for parameter k with dimension $p \times 1$.
(Optional) smoothing penalty and hyperparameter	
$\boldsymbol{P}_k$	Smoothing penalty matrix for fitting penalized splines
λ_k	Smoothing penalty hyperparameter for parameter k. Higher λ_k corresponds to more smoothing.
EDF	Effective degrees of freedom for a given $\boldsymbol{P}_k$ and λ_k, inclusive of fixed effects.

Algorithm 1: Distributed RS algorithm, fixed degrees of freedom

Init: (Central) If using splines: spline basis.
Data: (Site i) $\boldsymbol{Y}_i, \boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{ip}, \boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{ip}$
Data: (Central) If using penalized splines: $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_p, \lambda_1, \ldots, \lambda_p$
Result: Point estimates for $\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\gamma}_1, \ldots, \hat{\gamma}_p$
Init: (Central) Coefficients $\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p$
while *Outer cycle: not all coefficients have converged* **do**
  for $k = 1, \ldots, p$ **do**
    while *Inner cycle: β_k and/or γ_k have not converged* **do**
      if *only fixed effects* **then**
        1. (Site i) Solve for $\boldsymbol{\eta}_{ik} = \boldsymbol{X}_{ik}\beta_k$.
        2. (Site i) At $\boldsymbol{\eta}_{ik}$, evaluate $\boldsymbol{u}_{ik}$ and $\boldsymbol{W}_{ik}$.
        3. (Site i) Evaluate $\boldsymbol{z}_{ik} = \boldsymbol{\eta}_{ik} + \boldsymbol{W}_{ik}^{-1}\boldsymbol{u}_{ik}$.
        4. (Site i) Evaluate and send: $\boldsymbol{M}_{i1} = \boldsymbol{X}_{ik}^T\boldsymbol{W}_{ik}\boldsymbol{X}_{ik}$ and $\boldsymbol{M}_{i2} = \boldsymbol{X}_{ik}^T\boldsymbol{W}_{ik}\boldsymbol{z}_{ik}$.
        5. (Central) Update $\beta_k = (\sum_{i=1}^m \boldsymbol{M}_{i1})^{-1}(\sum_{i=1}^m \boldsymbol{M}_{i2})$.
        6. (Central) Check if β_k has converged. If not, send updated β_k to each site and continue inner cycle. If yes, continue **for** loop.
      end
      if *fixed effects and fixed effects smooth* **then**
        1. Re-define $\boldsymbol{X}_{ik} = [\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]$ and $\beta_k = [\beta_k^T\ \gamma_k^T]^T$.
        2. Fit using the "only fixed effects" algorithm.
      end
      if *fixed effects and fixed penalty smooth* **then**
        1. (Site i) Solve for $\boldsymbol{\eta}_{ik} = \boldsymbol{X}_{ik}\beta_k + \boldsymbol{Z}_{ik}\gamma_k$.
        2. (Site i) At $\boldsymbol{\eta}_{ik}$, evaluate $\boldsymbol{u}_{ik}$ and $\boldsymbol{W}_{ik}$.
        3. (Site i) Evaluate $\boldsymbol{z}_{ik} = \boldsymbol{\eta}_{ik} + \boldsymbol{W}_{ik}^{-1}\boldsymbol{u}_{ik}$.
        4. (Site i) Evaluate and send: $\boldsymbol{M}_{i1} = [\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]^T\boldsymbol{W}_{ik}[\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]$ and $\boldsymbol{M}_{i2} = [\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]^T\boldsymbol{W}_{ik}\boldsymbol{z}_{ik}$.
        5. (Central) Update $[\beta_k^T\ \gamma_k^T] = (\sum_{i=1}^m \boldsymbol{M}_{i1} + \lambda_k\boldsymbol{P}_k)^{-1}(\sum_{i=1}^m \boldsymbol{M}_{i2})$.
        6. (Central) Check if $[\beta_k^T\ \gamma_k^T]$ has converged. If not, send updated $[\beta_k^T\ \gamma_k^T]$ to each site and continue inner cycle. If yes, continue **for** loop
      end
    end
  end
  1. (Central) Check if all β_k and/or γ_k have converged compared to previous outer cycle. If not, continue outer cycle. If yes, exit.
end
Inference: (Site i) Evaluate and send Hessian $\boldsymbol{H}_i$.
Inference: (Central) Evaluate $\boldsymbol{H} = \sum_{i=1}^m \boldsymbol{H}_i$.

Algorithm 2: Distributed RS algorithm, automated penalty selection

Init: (Central) If using splines: spline basis.
Data: (Site i) $\boldsymbol{Y}_i, \boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{ip}, \boldsymbol{Z}_{i1}, \ldots, \boldsymbol{Z}_{ip}$
Data: (Central) If using penalized splines: $\boldsymbol{P}_1, \ldots, \boldsymbol{P}_p, \lambda_1, \ldots, \lambda_p$
Result: Point estimates for $\hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\gamma}_1, \ldots, \hat{\gamma}_p$
Init: (Central) Coefficients $\beta_1, \ldots, \beta_p, \gamma_1, \ldots, \gamma_p$
while *Outer cycle: not all coefficients have converged* **do**
  for $k = 1, \ldots, p$ **do**
    while *Inner cycle: β_k and/or γ_k have not converged* **do**
      1. (Site i) Solve for $\boldsymbol{\eta}_{ik} = \boldsymbol{X}_{ik}\beta_k + \boldsymbol{Z}_{ik}\gamma_k$.
      2. (Site i) At $\boldsymbol{\eta}_{ik}$, evaluate $\boldsymbol{u}_{ik}$ and $\boldsymbol{W}_{ik}$.
      3. (Site i) Evaluate $\boldsymbol{z}_{ik} = \boldsymbol{\eta}_{ik} + \boldsymbol{W}_{ik}^{-1}\boldsymbol{u}_{ik}$.
      4. (Site i) Evaluate and send: $\boldsymbol{M}_{i1} = [\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]^T\boldsymbol{W}_{ik}[\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]$ and $\boldsymbol{M}_{i2} = [\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}]^T\boldsymbol{W}_{ik}\boldsymbol{z}_{ik}$.
      5. (Central) For an arbitrarily fine grid of λ_k ranging from small to large penalization, evaluate across the grid: $[\beta_k^T\ \gamma_k^T]_{\lambda_k} = (\sum_{i=1}^m \boldsymbol{M}_{i1} + \lambda_k\boldsymbol{P}_k)^{-1}(\sum_{i=1}^m \boldsymbol{M}_{i2})$.
      6. (Central) Send list of $[\beta_k^T\ \gamma_k^T]_{\lambda_k}$ to each site.
      7. (Site i) Evaluate and send $SSR_i(\lambda_k) = \|\boldsymbol{z}_{ik} - [\boldsymbol{X}_{ik}\ \boldsymbol{Z}_{ik}][\beta_k^T\ \gamma_k^T]_{\lambda_k}\|^2$ (for GCV) or site-wise deviance, $D(\lambda_k)$ (for GAIC) for each set of coefficients in the list.
      8. (Central) Select optimal λ_k across the grid that minimizes $GCV = (1 - n^{-1} * EDF)^{-2}\sum_{i=1}^m SSR_i(\lambda_k)$ or $GAIC = EDF * \log(n) + \sum_{i=1}^m D(\lambda_k)$. Update $[\beta_k^T\ \gamma_k^T]$ to the coefficients corresponding to the optimal λ_k.
      9. (Central) Check if $[\beta_k^T\ \gamma_k^T]$ has converged. If not, send updated $[\beta_k^T\ \gamma_k^T]$ to each site and continue inner cycle. If yes, continue **for** loop
    end
  end
  1. (Central) Check if all β_k and/or γ_k have converged compared to previous outer cycle. If not, continue outer cycle. If yes, exit.
end
Inference: (Site i) Evaluate and send Hessian $\boldsymbol{H}_i$.
Inference: (Central) Evaluate $\boldsymbol{H} = \sum_{i=1}^m \boldsymbol{H}_i$.

communication rounds, if likelihood ratio test inference is desired, reduced models should be fit simultaneously to full models.

For models using automated penalty smooth terms, the use of the likelihood ratio test discussed above may be liberal since it does not appropriately account for the data-driven choice of λ_k (Nychka 1988, Marra and Wood 2012, Wood 2013a,b). Exact inference is not implemented for smooth terms in this context. However, in many GAMLSS settings where the goal is to predict the quantiles of new data given their covariates, optimally-penalized smooth terms may be desirable for their accurate model fit, despite challenges with inference.

2.4 dGAMLSS applications

2.4.1 Modeling BMI in intensive care unit patients

Patients from the MIMIC-IV database were used to model the relationship between age, sex, and BMI, controlling for different intensive care unit (ICU) types. The MIMIC-IV database has been previously described (Johnson *et al.* 2023). While the MIMIC-IV database may be sub-optimal for developing reference BMI charts applicable to a non-ICU population, the EHR setting provides a compelling scenario for the demonstration of dGAMLSS.

We filtered the MIMIC-IV dataset such that patients who had extreme heights, weights, or BMIs, where extremeness was defined as less than the 1st percentile or greater than the 99th percentile, were excluded. This accounted for data entry errors, such as incorrect measurement units, accidental extra digits, and more, which would otherwise result in a small subset of patients having physiologically implausible heights, weights, or BMIs. Additionally, hospital stays where patients were transferred between ICUs were removed, such that each remaining patient was present in exactly one ICU dataset. Finally, if a patient had more than one hospital stay resulting in longitudinal measurements, BMI and admission age from the first hospital stay was retained, and the following timepoints were discarded, since dGAMLSS requires cross-sectional data.

This filtering resulted in a total of 25 112 unique patients who were admitted at ages from 18 to 100 years. 60.2% of the patients were male, while 39.8% of the patients were female. These patients were admitted to nine unique ICUs, with the following breakdown: Medical ICU (MICU; $n = 4222$), Medical/Surgical ICU (MSICU; $n = 3522$), Surgical ICU (SICU;

$n = 2584$), Trauma Surgical ICU (TSICU; $n = 2616$), Coronary Care Unit (CCU; $n = 3102$), Cardiac Vascular Intensive Care Unit (CVICU; $n = 8269$), Neuro Stepdown Unit ($n = 140$), Neuro Intermediate Care Unit ($n = 340$), Neuro Surgical ICU (NSICU, $n = 317$). To control for the relationship between admission type and BMI, patients were defined as medical, surgical, neurologic, or cardiac based on which ICU the patients were admitted to. For example, patients admitted to the MSICU were defined as both medical and surgical patients. For the purposes of this demonstration, each ICU was defined as a unique site such that patient-level data could not be communicated across ICUs.

We fit a fixed effect smooth model for BMI using the four-parameter Box-Cox power exponential (BCPE) distribution, a generalization of the t-distribution which allows for additional parametric modeling of skewness and kurtosis. Each of the four parameters is modeled using a fixed effect smooth term for age at admission and fixed effect terms for sex and admission type. For each parameter, non-orthogonalized, regular-interval B-spline design matrices are defined using the following total number of knots: $\mu = 6$, $\sigma = 5$, $\nu = 2$, and $\tau = 2$. These knots are chosen based on the degrees of freedom given by the gold standard model and are used to demonstrate that, if optimal degrees of freedom are known, dGAMLSS can successfully fit the desired model.

In this analysis, we compare the fitted dGAMLSS model to a pooled GAMLSS model using the same fixed effect spline basis in order to show that dGAMLSS accurately reproduces pooled GAMLSS output when splines are identically defined. We analyze dGAMLSS coefficient and standard error estimates in comparison to those from the pooled GAMLSS models to assess the validity of distributed inference. Additionally, we plot dGAMLSS predicted quantiles against pooled GAMLSS predicted quantiles for each observation to assess the accuracy of predicted quantiles. Finally, we construct dGAMLSS reference charts and compare them to pooled GAMLSS reference charts.

2.4.2 Microbiome relative abundance modeling

Microbiome data from the gut microbiome-metabolome dataset collection were used to model how Proteobacteria relative abundance changes across age and sex. Processing of this dataset is previously described (Muller *et al.* 2022). Of the fourteen datasets included in the collection, two datasets were excluded due to a lack of sex covariate, and one dataset was excluded due to a lack of exact age covariate. Participants outside of the control group were excluded so that reference Proteobacteria relative abundance could be modeled. When more than one microbiome sample was available for a given participant, the earliest sample was retained, and following measurements were discarded.

In the end, our analysis included 569 participants from 60 days to 83 years of age. In this sample, 54.7% were male and 45.3% were female. The eleven studies were treated as unique sites such that participant-level data could not be transferred between studies. The smallest study contained eight subjects, while the largest study contained 89 subjects.

We fit a fixed penalty smooth model for Proteobacteria relative abundance using the three-parameter zero-inflated beta distribution, which has been previously proposed for modeling bacteria phyla relative abundances (Ho *et al.* 2019). The zero-inflated beta distribution is a mixed distribution comprised of the beta and Bernoulli distribution, where the third parameter models the probability of observing zeros. Use of this distribution is supported in our dataset since a small proportion of participants (2.66%) have zero observed Proteobacteria reads.

We model the mean parameters with a fixed penalty smooth term for age as well as fixed effects for sex and dataset. We model the variance and zero-inflation parameters with a fixed penalty smooth term for age and a fixed effect for sex. The B-spline design matrix is specified to have 20 knots, allowing for a high maximum amount of flexibility. Penalties are manually selected such that the overall EDF of each smooth term is equivalent to that of the pooled GAMLSS model. While this approach to choosing penalties is not feasible in a real application of dGAMLSS, we use this approach to demonstrate the capability of dGAMLSS to fit penalized smooths when a gold-standard penalty is known.

We compare the fitted dGAMLSS model to the pooled GAMLSS model fit using penalized B-splines with default maximum likelihood penalty selection. We plot dGAMLSS predicted quantiles against pooled GAMLSS predicted quantiles for each observation to assess the accuracy of predicted quantiles. We also compare dGAMLSS reference charts with pooled GAMLSS reference charts to assess the quality of dGAMLSS microbiome charts.

2.4.3 Brain volumetric modeling

Neuroimaging data from a subset of participants in the LBCC were used to model how gray matter volume (GMV) changes across age and sex. Details on how GMV were obtained are previously described (Bethlehem *et al.* 2022). Exclusion criteria for this subset involved removal of low-quality scans and non-physiologically plausible brain volumes (Gardner *et al.* 2024). Additionally, LBCC data from the UK Biobank were excluded from this analysis due to data-sharing challenges.

Ultimately, 26 480 participants were included in this analysis, spanning ages 3.2 to 100 years. Of these participants, 48.7% were male and 52.3% were female. Participants spanned 50 studies which, for the purposes of this demonstration, were defined as unique sites such that participant-level data could not be transferred from any study. Study sample sizes ranged such that the smallest included study only contributed 5 participants, while the largest included study contributed 7889 participants.

We fit an automated penalty smooth model for GMV using the three-parameter generalized gamma distribution, chosen based on recommendations from Bethlehem *et al.* (2022). The generalized gamma distribution is a generalized version of the gamma distribution, which allows for one additional shape parameter. We model each of the mean and variance parameters with a penalized smooth term for age, as well as fixed effects for sex and study. We model the skewness parameter with only a penalized smooth term for age and a fixed effect for sex. The B-spline design matrix is specified to have 20 knots, allowing for a high maximum amount of flexibility, such that penalization will automatically regularize the effective degrees of freedom. Based on recommendations from Rigby and Stasinopoulos, we use BIC for penalty selection (Rigby and Stasinopoulos 2005).

We compare the fitted dGAMLSS model to the pooled GAMLSS model fit using penalized B-splines with default maximum likelihood penalty selection. We plot dGAMLSS predicted quantiles against pooled GAMLSS predicted

quantiles for each observation to assess the accuracy of predicted quantiles. We also compare dGAMLSS reference charts with pooled GAMLSS reference charts to assess the quality of dGAMLSS brain charts.

3 Results

Distributed GAMLSS (dGAMLSS) provides machinery for estimating semi-parametric coefficients across a broad family of GAMLSS distributions defined by up to four parameters, including mean, variance, skewness, and kurtosis, across multiple sites without sharing any patient-level data. Once coefficients are estimated, dGAMLSS allows for inference and estimation of model-based centiles, defined as quantiles multiplied by 100. Notation for dGAMLSS is defined in Table 1.

Briefly, the proposed dGAMLSS algorithm adapts the pooled GAMLSS fitting algorithm, where the terminology "pooled" refers to the standard formulation of GAMLSS designed for datasets where all data is readily available at the subject level (Rigby and Stasinopoulos 1996, 2005). This pooled GAMLSS algorithm consists of two nested cycles, which result in maximization of the likelihood—the outer cycle iterates across the distribution parameters, while the inner cycle updates each parameter's coefficients using the Newton-Raphson method, while keeping the coefficients for other parameters constant. Noting that pooled Newton-Raphson updates can be exactly reproduced as the sum of corresponding site-level matrices and substituting as appropriate, dGAMLSS achieves the desired result of privacy-preserving model fitting (Fig. 1). Further details are provided in the Methods section.

We apply dGAMLSS to estimate body mass index (BMI) charts using electronic health record (EHR) data from the Medical Information Mart for Intensive Care (MIMIC)-IV, microbiome relative abundance charts using data from the Curated Gut Microbiome Metabolome Data Resource, and brain charts using structural magnetic resonance imaging (MRI) data from the Lifespan Brain Chart Consortium (LBCC) (Muller *et al.* 2022, Bethlehem *et al.* 2022, Johnson *et al.* 2023). To demonstrate the various smooth term implementations in dGAMLSS, we use fixed effect smooth terms with regular interval knots in the BMI setting, fixed penalty smooth terms in the microbiome setting, and automated penalty smooth terms in the brain chart setting.

We first provide the fixed effect BMI example to demonstrate the accuracy of dGAMLSS compared to the pooled analysis under a model specification that is relatively straightforward to fit but may be limited in its ability to fit high amounts of wiggliness between spline knots. We include the fixed penalty microbiome example to show how this limitation can be overcome through the addition of more knots and a penalty, potentially at the cost of a greater number of communication rounds required. The fixed effect and fixed penalty models, though reliant on an assumption of known degrees of freedom that may be impractical in real dGAMLSS applications, are used to illustrate the capacity of dGAMLSS for near-exact fitting and inference compared to identical pooled GAMLSS models. These algorithms may also be useful in scenarios where spline degrees of freedom are roughly known, such as when updating outdated or single-site GAMLSS-based growth charts.

Finally, we provide the automated penalty brain chart example to illustrate how a gold-standard model can be fit

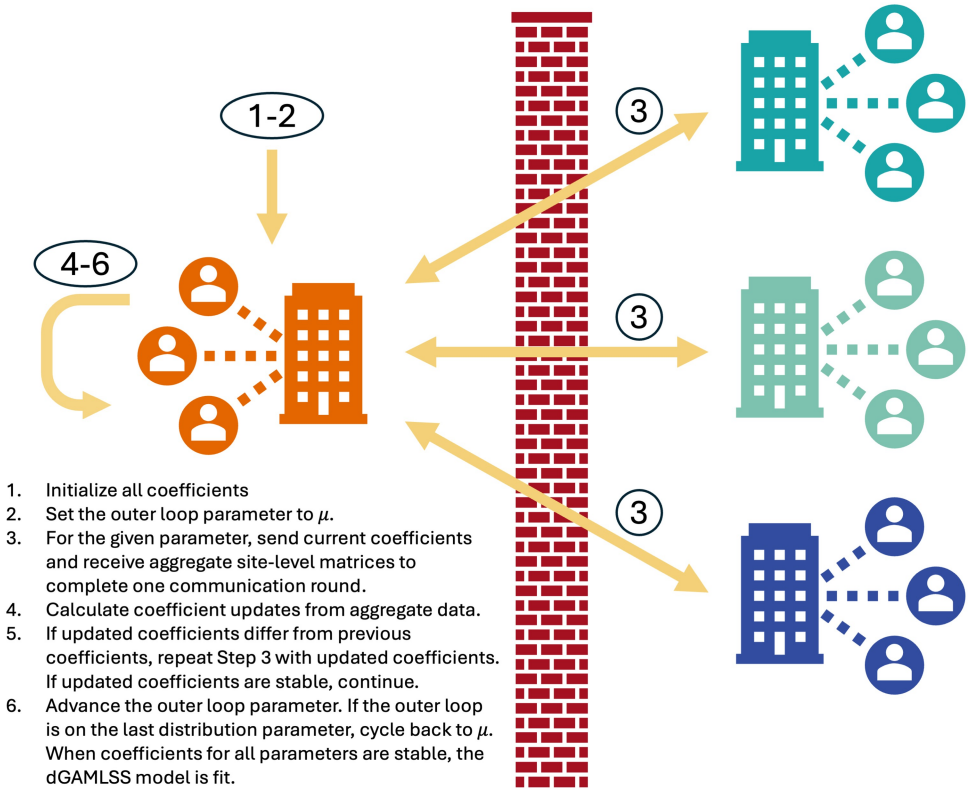

Figure 1. Overview of the dGAMLSS algorithm, which enables fitting of GAMLSS models via privacy-preserving federated learning. Fitting occurs via an outer cycle over the parameters and a nested inner cycle of Newton-Raphson updates for parameter-specific coefficients.

using automated penalty selection, with the caveat that this model introduces inferential challenges, requires sites to send additional information per communication round, and may necessitate additional communication rounds.

3.1 BMI modeling

The fixed effect dGAMLSS model, which demonstrates that dGAMLSS can near-exactly reproduce pooled GAMLSS output when a common spline basis is used, was fit on an EHR dataset of patients admitted to various intensive care units (ICUs) at Beth Israel Deaconess Medical Center. This dataset included 25 112 unique patients from ages 18 to 100 years, where 60.2% of the patients were male and 39.8% of the patients were female. These patients were admitted to nine unique ICUs which, for this example, were treated as unique sites such that patient-level data could not be communicated across ICUs.

We fit a fixed effect smooth model for BMI using the four-parameter Box-Cox power exponential (BCPE) distribution, a generalization of the t-distribution which allows for additional parametric modeling of skewness and kurtosis. The BCPE distribution was chosen based on prior BMI reference charts fit by Rigby and Stasinopoulos and the World Health Organization (Rigby and Stasinopoulos 2004, World Health Organization (WHO) 2006). Each of the four parameters is modeled using a fixed effect smooth term for age at admission and fixed effect terms for sex and admission type.

This model took one communication round to establish the spline basis, 69 communication rounds to converge, and one communication round to obtain inference. The final Bayesian Information Criteria (BIC) of the dGAMLSS model and pooled GAMLSS model were 158 575.4 and 158 575.3, respectively, indicating effectively equivalent fit.

The dGAMLSS model provided identical inference to the pooled model, with the exception of small numerical differences (Fig. 2). Large standard errors for extreme spline bases were observed for both dGAMLSS and pooled GAMLSS models due to a relative lack of observations with non-zero values for those spline bases. dGAMLSS reference BMI charts appeared qualitatively equivalent to pooled reference charts, and both reference charts appropriately reflected BMI trends over the lifespan—larger mean values and variances were observed from ages 25 to 80, while younger and older patients were observed to have comparatively lower BMIs at the median and upper centiles (Fig. 3, left). In these age ranges, skewness towards higher BMIs is also observed as well as kurtosis of the upper tails, reflecting the necessity of using a four-parameter distribution such as BCPE for reference BMI chart fitting. Finally, predicted quantiles between dGAMLSS and pooled GAMLSS fits were nearly identical, with the exception of numerical differences (Fig. 3, right). The correlation between dGAMLSS and pooled quantile predictions was >0.9999. Smooth patterns observed in the scatterplot are due to numerical differences in spline coefficients. Overall, this analysis demonstrates that, given identical models and spline bases, dGAMLSS provides an exact method for federated learning and inference in the GAMLSS setting.

3.2 Proteobacteria relative abundance modeling

The fixed penalty microbiome dGAMLSS model, which demonstrates the capacity of dGAMLSS to implement penalized splines when a gold-standard penalty is known, used a dataset composed of eleven studies. Each study was treated as a unique site such that participant-level data could not be transferred between studies. Overall, the dataset included 569 participants from ages 60 days to 83 years, where 54.7% were male and 45.3% were female.

We fit a fixed penalty smooth model for Proteobacteria relative abundance using the three-parameter zero-inflated beta distribution, which has been previously proposed for modeling bacteria phyla relative abundances (Ho *et al.* 2019). The zero-inflated beta distribution is a mixed distribution comprised of the beta and Bernoulli distribution, where the third parameter models the probability of observing zeros. Use of this distribution is supported in our dataset since a meaningful proportion of participants (2.66%) have zero observed Proteobacteria reads.

This model took two communication rounds to establish the spline basis, 117 communication rounds to converge, and one communication round to obtain centiles. The final BIC of the dGAMLSS model and pooled model were highly similar, at −2360.2 and −2365.2, respectively, suggesting minor numerical differences. The estimated degrees of freedom (EDF) of smooth terms for both models were 4.750 for the μ term, 4.168 for the σ term, and 3.368 for the ν term.

dGAMLSS reference Proteobacteria relative abundance charts appeared qualitatively similar to pooled reference charts (Fig. 4, left). In these charts, from ages 0 to 40, we observe that around half of all individuals have minimal Proteobacteria relative abundances. However, significant upward skew is observed, as the upper centiles seem to have significantly higher Proteobacteria presence in their microbiomes. For ages above 40, this trend is even more evident, with higher Proteobacteria relative abundances being observed in the upper centiles, while centiles at and below the median remained relatively similar to before. Across ages, females tended to have slightly higher Proteobacteria proportions than males, suggesting more variation of Proteobacteria abundance in females may be normal.

Finally, quantile predictions between fixed penalty dGAMLSS and pooled GAMLSS fits were nearly identical (Fig. 5, right)—most differences between predicted quantiles were <0.02. The correlation between dGAMLSS and pooled quantile predictions was 0.9994. Overall, this analysis demonstrates that fixed penalty dGAMLSS achieved an effectively identical fit when compared to the pooled model.

3.3 Gray matter volume modeling

The automated penalty dGAMLSS model demonstrates how data-driven penalty selection can be implemented by simultaneously sending site-level matrices for multiple candidate penalties during each communication round. This model was fit on a neuroimaging dataset of 26 480 participants from ages 3.2 to 100 years where 48.7% were male and 52.3% were female. Participants spanned 50 studies, which were defined as unique sites such that participant-level data could not be transferred from any study.

We fit an automated penalty smooth model for GMV using the three-parameter generalized gamma distribution, chosen based on a prior gold-standard analysis (Bethlehem *et al.* 2022). The generalized gamma distribution is a generalized version of the gamma distribution, allowing for one additional shape parameter. We model the mean and variance parameters with a penalized smooth term for age as well as fixed effects for sex and study. We model the skewness parameter with a penalized smooth term for age and a fixed

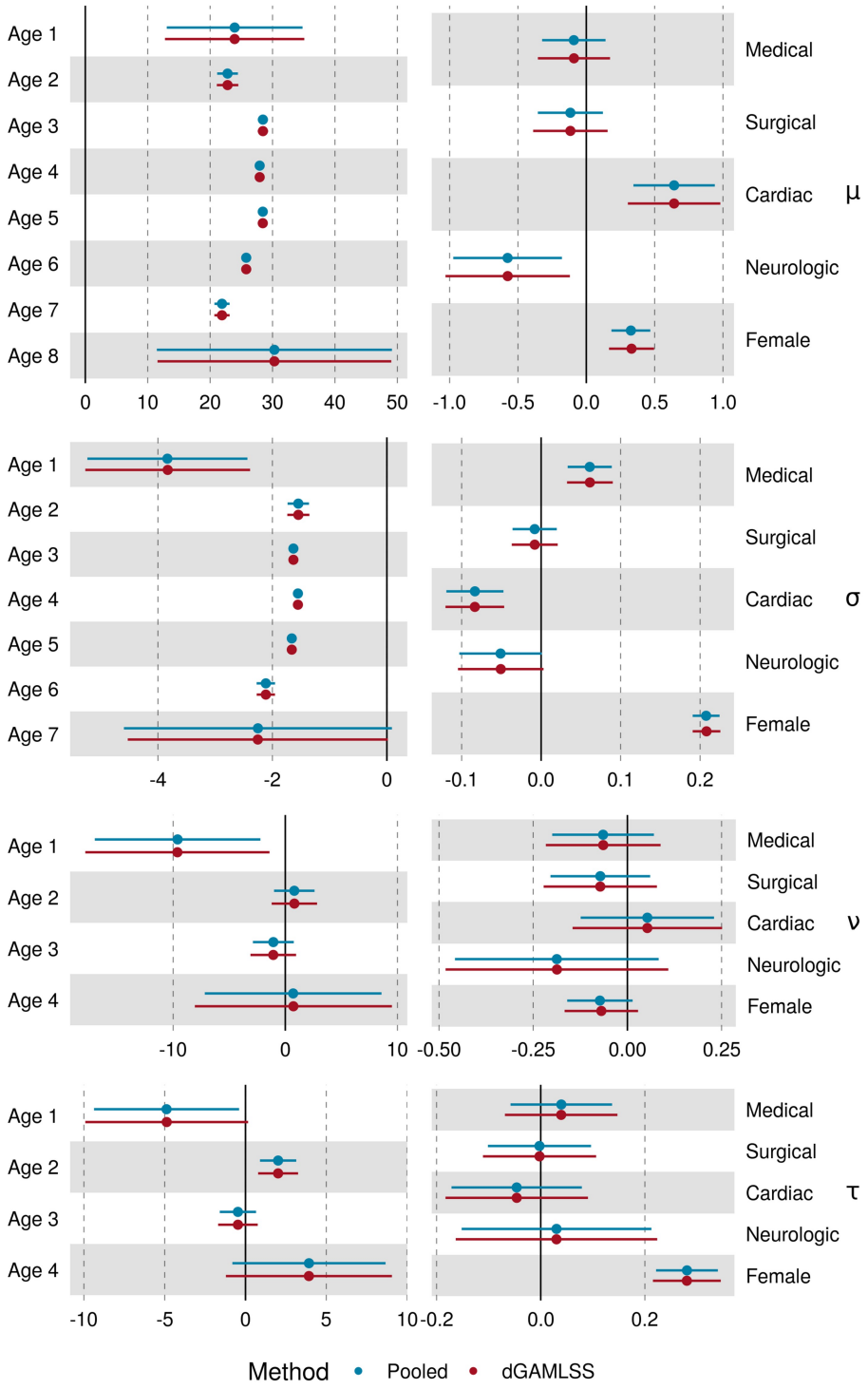

Figure 2. Forest plot of dGAMLSS and pooled GAMLSS coefficients for each parameter. Coefficients for each age spline basis, as well as for sex and admission type, are shown. Age spline bases are numbered from young age to old age. Dots represent point estimates and solid lines represent 95% confidence interval. Vertical solid lines indicate no effect.

effect for sex. The B-spline design matrix is specified to have 20 knots, allowing for a high maximum amount of flexibility, such that automated penalization will regularize the effective degrees of freedom. Based on recommendations from Rigby and Stasinopoulos, we use BIC for penalty selection (Rigby and Stasinopoulos 2005).

The automated penalty dGAMLSS model took one communication round to establish the spline basis and 57 communication rounds to converge. The final automated penalty dGAMLSS model chose 7.32 EDF for the μ term, 6.05 EDF for the σ term, and 3.01 EDF for the ν term. Meanwhile, the pooled GAMLSS model chose for 16.41 EDF for the μ term, 5.62 EDF for the σ term, and 6.03 EDF for the ν term. The final BIC of the dGAMLSS and pooled models were 646 260.3 and 646 345.7, respectively, indicating the lower EDF selected by the dGAMLSS model may provide a better fit than the higher EDF selected by the pooled GAMLSS model. Outside of numerical differences in optimization, additional differences

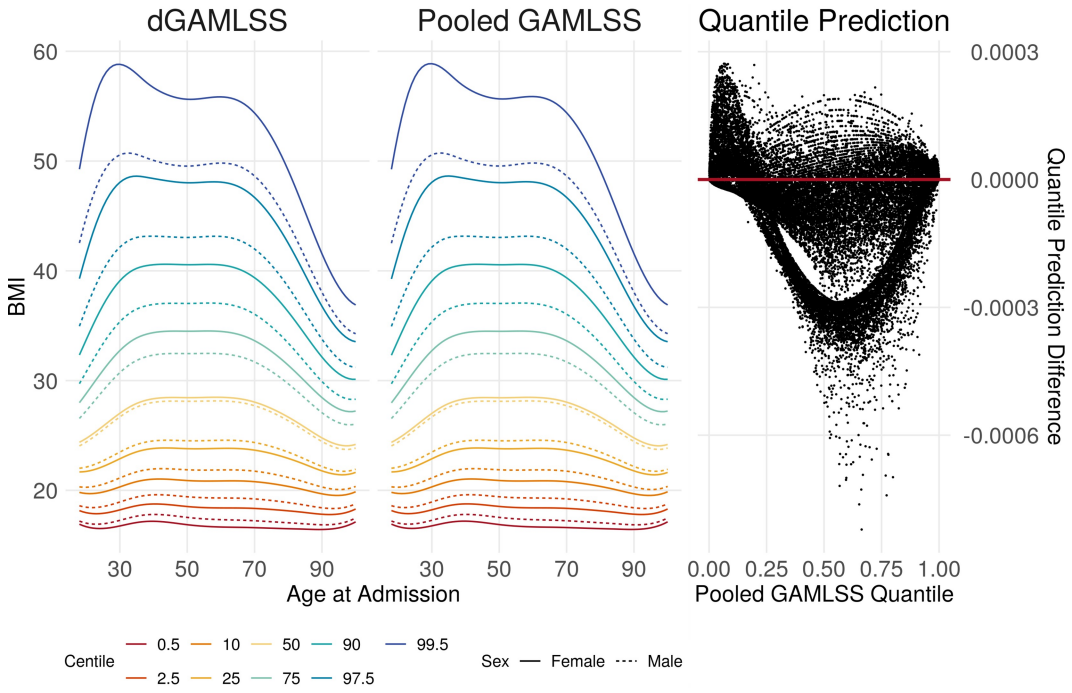

Figure 3. Left: dGAMLSS and pooled GAMLSS reference charts for BMI in ICU patients. Colors represent different centiles, with the 2.5 and 97.5 centile lines demonstrating a potential reference range defined by the inner 95% of patients. Solid lines represent reference centiles for females, while dotted lines represent reference centiles for males. Right: Scatterplot comparison of predicted quantiles for every patient. Pooled GAMLSS quantile predictions are shown on the *x*-axis, and differences between dGAMLSS and pooled quantile predictions are shown on the *y*-axis.

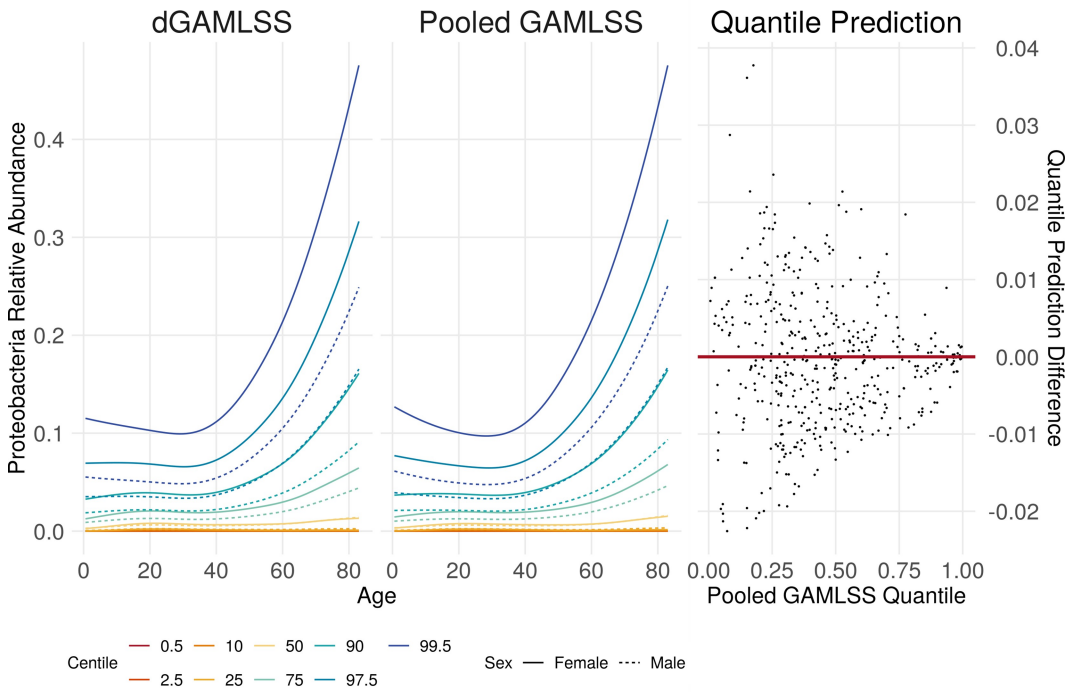

Figure 4. Left: dGAMLSS and pooled GAMLSS reference charts for Proteobacteria relative abundance. Colors represent different centiles, with the 2.5 and 97.5 centile lines demonstrating a potential reference range defined by the inner 95% of participants. Solid lines represent reference centiles for females, while dotted lines represent reference centiles for males. Right: Scatterplot comparison of predicted quantiles for every patient. Pooled GAMLSS quantile predictions are shown on the *x*-axis, and differences between dGAMLSS and pooled quantile predictions are shown on the *y*-axis.

in EDF selection may be due to different knot placement for the spline basis or the use of different penalty selection criteria.

dGAMLSS reference GMV charts appeared qualitatively similar to pooled reference GMV charts (Fig. 5, left). In these charts, we note a large increase in GMV from around age 3 to 10, steady decreases in GMV from age 10 to 50, and relatively steeper GMV decreases at older ages. Variances seem to be higher at younger ages, and slight skewness towards higher volume is observed across all ages. Male brains tended to have higher GMV. Small differences between dGAMLSS and pooled GAMLSS predictions can be observed for the extreme centiles.

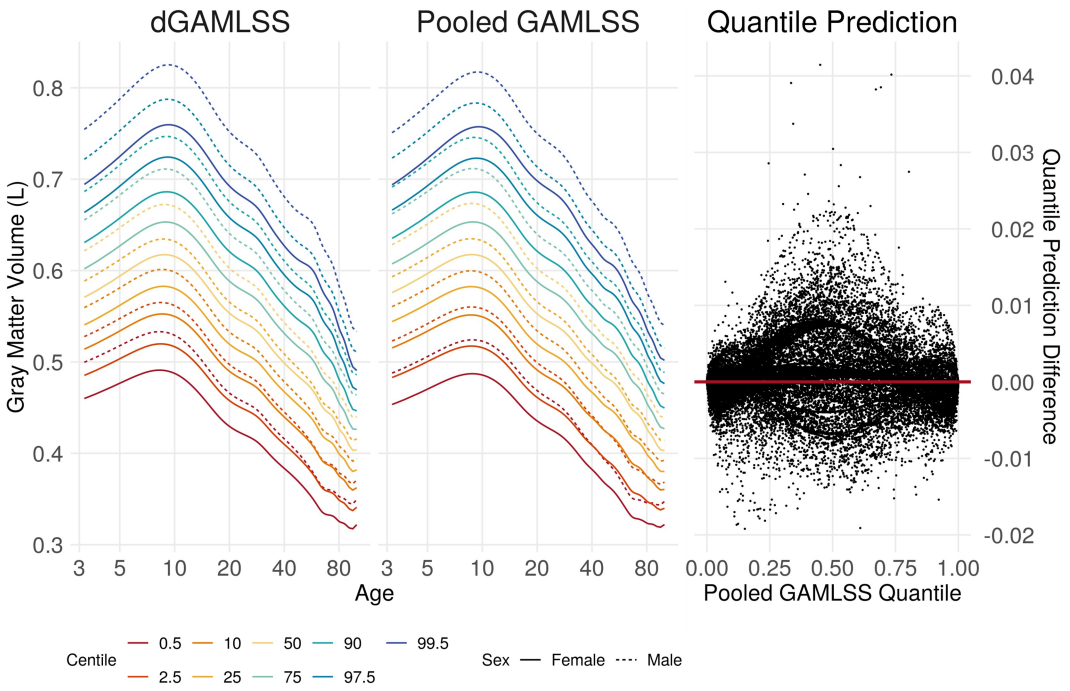

Figure 5. Left: dGAMLSS and pooled GAMLSS reference charts for gray matter volume, in liters, in healthy participants, plotted on a log age scale to better demonstrate large changes in young brains. Colors represent different centiles, with the 2.5 and 97.5 centile lines demonstrating a potential reference range defined by the inner 95% of participants. Solid lines represent reference centiles for females, while dotted lines represent reference centiles for males. Bottom: Scatterplot comparison of predicted quantiles for every participant. Pooled GAMLSS quantile predictions are shown on the *x*-axis, and differences between dGAMLSS and pooled quantile predictions are shown on the *y*-axis.

Finally, quantile predictions between automated penalty dGAMLSS and pooled GAMLSS fits were again nearly identical (Fig. 5, right)—most differences between predicted quantiles were <0.02, though a few subjects demonstrated relatively larger differences. The correlation between dGAMLSS and pooled quantile predictions was 0.9999. Overall, this analysis demonstrates the capacity of automated penalty dGAMLSS to successfully replicate a pooled analysis without *a priori* knowledge of parameter-wise effective degrees of freedom for smooth terms.

4 Discussion

In this manuscript, we develop dGAMLSS, an exact, privacy-preserving algorithm for fitting semi-parametric GAMLSS in the federated setting. We show dGAMLSS fits nearly-identical models to pooled GAMLSS in the fixed effect smooth term setting, including coefficient values and Wald-type inference. In the fixed penalty setting, we show dGAMLSS can also achieve similar performance while allowing for potentially higher amounts of non-linearity in certain local regions compared to others. Finally, in the automated penalty setting, we provide machinery for automated penalty selection based on distributed GAIC and GCV criteria and again demonstrate that dGAMLSS is able to reproduce the gold-standard GAMLSS population reference charts.

Notably, as with pooled GAMLSS and other multi-site models, standardization of data measurement, processing, and storage across sites is necessary in order to generate robust and generalizable models and inference under the dGAMLSS framework. In the population reference charts built in this manuscript, covariates for site effects are included in each model in the form of admission type for BMI modeling, dataset identity for microbiome modeling, and study identity for brain volume modeling. The site effects in these GAMLSS models are estimated for one or more moments and allow for semi-parametric estimation of site effects. However, if site-based effects are more complex in any given moment due to underlying discrepancies in the data collection process, including markedly different measurement devices or measurement definitions, such site effects may not be adequately accounted for. Thus, when applying dGAMLSS to real-world data, care must be taken to assess the quality and consistency of the data collection process across candidate sites. Additionally, in the setting of highly correlated or high-dimensional covariates, both pooled GAMLSS and dGAMLSS may encounter convergence challenges. As in standard regression, multicollinearity can produce unstable coefficient estimates and inconsistent inference, issues that are further compounded by computational dependencies between lower- and higher-order moment estimation. While such instability may affect the interpretability of covariate effects, it is unlikely to substantially impair overall model fit. High-dimensional covariates pose similar, though typically less severe, difficulties, with the stability of each covariate largely dependent on its effective sample size.

Further work is necessary to extend dGAMLSS to longitudinal settings by incorporating subject-wise random effects as well as to survival analysis settings by incorporating censored data capabilities (Li *et al.* 2022, Luo *et al.* 2022b, Masciocchi *et al.* 2022, Rahimian *et al.* 2022). Cross-sectionally, an investigation into how to incorporate modern statistical ideas for testing the overall significance of smooth terms in the automated penalty GAMLSS setting would also be beneficial to allow for statistically sound inference (Nychka 1988, Marra and Wood 2012, Wood 2013a,b). Notably, this significance testing limitation also exists in the pooled GAMLSS setting.

In a similar vein, the pooled GAMLSS algorithm (Rigby and Stasinopoulos 2005) utilizes backfitting cycles within the

inner Newton-Raphson cycles. This allows GAMLSS to uniquely fit multiple smooth terms and spline-based interactions for each parameter without running into a problem of near-singular matrices and the resulting coefficient instability. However, in the distributed setting, true backfitting is prohibitive—in the setting of just one smooth term, true backfitting requires at least two additional communication rounds per inner cycle. The minimum number of communication rounds per inner cycle increases by two for each additional smooth term, and these communication rounds are multiplied if multiple iterations of backfitting are required in a given inner cycle. While fitting one smooth term per parameter may be adequate in the development of reference charts where age is often the primary smooth term, the development of dGAMLSS extensions that allow for communication-efficient backfitting is important. While dGAMLSS can theoretically fit multiple splines per parameter without backfitting, the development of communication-efficient backfitting would allow for improved computational stability when estimating multiple smooth terms per parameter or spline-based interaction terms (Daniel *et al.* 2024).

Exploration into how dGAMLSS can be made more communication-efficient would allow for further uptake and utility of the algorithm (Duan *et al.* 2019, Tong *et al.* 2020, Luo *et al.* 2022a). Empirically, models seem to converge in under 100 communication rounds—while fitting such a model is feasible in practice, minimizing communication costs would reduce barriers to use. Other approaches to reducing barriers to dGAMLSS use may be the further development of software specialized to the task of automating communication rounds in federated analysis settings (Gazula *et al.* 2020, Rootes-Murdy *et al.* 2022).

Privacy-wise, the proposed dGAMLSS algorithm follows similar distribution ideas as other federated learning methods, which involve sending two coefficient-based matrices related to the first and second derivatives of overall site-wise likelihoods (Li *et al.* 2020, Pfitzner *et al.* 2021, Yin *et al.* 2021). However, since the dGAMLSS algorithm is not shown to be differentially private, iterative exchanges of summary statistics could lead to incremental information disclosure and increase re-identification risk (Kim *et al.* 2020). There may be increased privacy risk for sites with small sample sizes relative to the number of covariates, individuals with extreme values for the outcome or one or more covariates, or individuals who have health information in outside databases and could be identified via linking to these outside databases. Such risks are also elevated in the automated penalty setting, since each round of penalty selection requires communication of the above matrices for each set of coefficient locations across a grid of proposed penalties.

Ultimately, we demonstrate the utility of dGAMLSS for building population reference charts in clinical, genomics, and neuroimaging settings where outcomes follow drastically different GAMLSS-family distributions. In the context of modern EHRs and big data efforts where large amounts of data are readily available within institution-specific databases, dGAMLSS promises the potential for real-time, heterogeneity-aware, privacy-preserving reference charts.

Author contributions

Fengling Hu (Conceptualization [equal], Data curation [equal], Formal analysis [equal], Investigation [equal], Methodology [equal], Software [equal], Validation [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jiayi Tong (Formal analysis [equal], Investigation [equal], Methodology [equal], Resources [equal], Validation [equal], Writing—review & editing [equal]), Margaret Gardner (Resources [equal], Validation [equal], Writing—review & editing [equal]), Andrew A. Chen (Resources [equal], Supervision [equal], Validation [equal], Writing—review & editing [equal]), Richard A.I. Bethlehem (Resources [equal], Validation [equal], Writing—review & editing [equal]), Jakob Seidlitz (Resources [equal], Validation [equal], Writing—review & editing [equal]), Hongzhe Li (Formal analysis [equal], Supervision [equal], Validation [equal], Writing—review & editing [equal]), Aaron Alexander-Bloch (Resources [equal], Supervision [equal], Validation [equal], Writing—review & editing [equal]), Yong Chen (Conceptualization [equal], Investigation [equal], Methodology [equal], Resources [equal], Supervision [equal], Validation [equal], Writing—review & editing [equal]), and Russell T. Shinohara (Conceptualization [equal], Formal analysis [equal], Funding acquisition [equal], Investigation [equal], Methodology [equal], Project administration [equal], Resources [equal], Supervision [equal], Validation [equal], Writing—review & editing [equal])

Supplementary data

Supplementary data are available at *Bioinformatics* online.

Conflict of interest: None declared.

Funding

F.H. and R.T.S. were supported by R01MH123550, R01MH112847, R01MH123563, R01NS112274, and R01NS060910 from the National Institutes of Health. F.H. was supported by the National Institutes of Health Medical Scientist Training Program T32 GM07170. M.G., J.S., and A.A.B. were supported by R01MH134896, R01MH132934, and R01MH133843 from the National Institutes of Health. M.G., R.T.S., J.S., and A.A.B. were supported by "RAising the Investment in Sex and gender Evidence (RAISE) Pilot Grant Award" funded by FOCUS on Health and Leadership for Women and Penn PROMOTES Research on Sex and Gender in Health. R.T.S. receives consulting income from Octave Bioscience and compensation for reviewership duties from the American Medical Association. A.A.C. receives compensation for reviewership duties from the American Medical Association. J.S., R.A.I.B., and A.A.-B. hold shares in Centile Bioscience, and J.S. and R.A.I.B. are directors of Centile Bioscience. Funding sources were not involved in study design, data analysis, manuscript preparation, or submission decisions.

Data availability

An R package containing all code for running dGAMLSS, performing distributed inference, and building distributed reference charts is provided at: https://github.com/hufengling/dGAMLSS. All code for analysis and manuscript preparation is provided at: https://github.com/hufengling/dGAMLSS_analyses. Included in the dGAMLSS package and analysis code are vignettes demonstrating how to set up dataframes and R

objects for distributed computation, as well as functions simulating how coefficient updates are performed at the central site.

The MIMIC-IV database is publicly available via PhysioNet (A Goldberger *et al.* 2000, Johnson *et al.*). The microbiome dataset is publicly available and was accessed via: https://github.com/borenstein-lab/microbiome-metabolome-curated-data (Muller *et al.* 2022). Availability of the LBCC database is managed at the discretion of each primary study and is previously described (Bethlehem *et al.* 2022, Gardner *et al.* 2024). Links to open and semi-open access datasets are listed at: https://github.com/brainchart/Lifespan.

References

Bethlehem RAI, Seidlitz J, White SR *et al.* Brain charts for the human lifespan. *Nature* 2022;**604**:525–33.

Borghi E, de Onis M, Garza C *et al.* Construction of the World Health Organization Child Growth Standards: selection of methods for attained growth curves. *Stat Med* 2006;**25**:247–65.

Brown J, Cook A, Lane K *et al.* PS2-9: the NIH health care systems research collaboratory. *Clinic Med Res* 2013;**11**:152.

Chu E, Keshavarz A, Boyd S *et al.* A distributed algorithm for fitting generalized additive models. *Optim Eng* 2013;**14**:213–24.

Cole TJ, Stanojevic S, Stocks J *et al.* Age- and size-related reference ranges: a case study of spirometry through childhood and adulthood. *Stat Med* 2009;**28**:880–98.

Cole TJ, Freeman JV, Preece MA *et al.* Body mass index reference curves for the UK, 1990. *Arch Dis Child* 1995;**73**:25–9.

Collins FS, Hudson KL, Briggs JP *et al.* PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014;**21**:576–7.

Daniel S, Bernd B, David R *et al.* Privacy-preserving and lossless distributed estimation of high-dimensional generalized additive mixed models. *Stat Comput* 2024;**34**:31.

Duan R, Boland MR, Moore JH *et al.* ODAL: a one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac Symp Biocomput* 2019;**24**:30–41.

Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci* 1996;**11**:89–121.

Gardner M, Shinohara RT, Bethlehem RAI *et al.* ComBatLS: A location- AND scale-preserving method for multi-site image harmonization. *Hum Brain Mapp* 2024;**46**:e70197.

Gazula H, Kelly R, Romero J *et al.* COINSTAC: collaborative informatics and neuroimaging suite toolkit for anonymous computation. *JOSS* 2020;**5**:2166.

Goldberger AL, Amaral LA, Glass L *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 2000;**101**:e215–e220.

Hastie T, Tibshirani R. Generalized additive models. *Statist Sci* 1986;**1**:297–310.

Ho NT, Li F, Wang S *et al.* metamicrobiomeR: an R package for analysis of microbiome relative abundance data using zero-inflated beta GAMLSS and meta-analysis across studies using random effects models. *BMC Bioinformatics* 2019;**20**:188.

Hripcsak G, Duke JD, Shah NH *et al.* Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;**216**:574–8.

Johnson AEW, Bulgarelli L, Shen L *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 2023;**10**:219.

Jones G, Barker A. Reference intervals. *Clin Biochem Rev* 2008;**29** Suppl 1:S93–S97.

Kia SM *et al.* 2021. Federated multi-site normative modeling using hierarchical Bayesian Regression. 2021.05.28.446120.

Kim M, Lee J, Ohno-Machado L *et al.* Secure and differentially private logistic regression for horizontally distributed data. *IEEE Trans Inform Forensic Secur* 2020;**15**:695–710.

Li L, Fan Y, Tse M *et al.* A review of applications in federated learning. *Comput Ind Eng* 2020;**149**:106854.

Li W, Tong J, Anjum MM *et al.* Federated learning algorithms for generalized mixed-effects model (GLMM) on horizontally partitioned data from distributed sources. *BMC Med Inform Decis Mak* 2022;**22**:269.

Luo C, Duan R, Naj AC *et al.* ODACH: a one-shot distributed algorithm for cox model with heterogeneous multi-center data. *Sci Rep* 2022a;**12**:6627.

Luo C, Islam MN, Sheils NE *et al.* dPQL: a lossless distributed algorithm for generalized linear mixed model with application to privacy-preserving hospital profiling. *J Am Med Inform Assoc* 2022b;**29**:1366–71.

Marra G, Wood SN. Coverage properties of confidence intervals for generalized additive model components. *Scand J Stat* 2012;**39**:53–74.

Masciocchi C *et al.* 2022. Federated Cox Proportional Hazards Model with multicentric privacy-preserving LASSO feature selection for survival analysis from the perspective of personalized medicine. In: *2022 IEEE 35th International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, Shenzhen, China, pp. 25–31.

Muller E, Algavi YM, Borenstein E *et al.* The gut microbiome-metabolome dataset collection: a curated resource for integrative meta-analysis. *NPJ Biofilms Microbiomes* 2022;**8**:79–7.

Nychka D. Bayesian confidence intervals for smoothing splines. *J Am Stat Assoc* 1988;**83**:1134–43.

O'Connor PJ. Normative data: their definition, interpretation, and importance for primary care physicians. *Fam Med* 1990;**22**:307–11.

Pfitzner B, Steckhan N, Arnrich B *et al.* Federated learning in a medical context: a systematic literature review. *ACM Trans Internet Technol* 2021;**21**:1–50.

Platt R, Brown JS, Robb M *et al.* The FDA sentinel initiative - an evolving national resource. *N Engl J Med* 2018;**379**:2091–3.

Rahimian S, Kerkouche R, Kurth I *et al.* Practical challenges in differentially-private federated survival analysis of medical data. In: *Proceedings of the Conference on Health, Inference, and Learning*. Virtual, pp. 411–425, 2022.

Rigby RA, Stasinopoulos DM. A semi-parametric additive model for variance heterogeneity. *Stat Comput* 1996;**6**:57–65.

Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J Royal Statis Society C Appl Stat* 2005;**54**:507–54.

Rigby RA, Stasinopoulos DM. Smooth centile curves for skew and kurtotic data modelled using the Box-Cox power exponential distribution. *Stat Med* 2004;**23**:3053–76.

Rootes-Murdy K, Gazula H, Verner E *et al.* Federated analysis of neuroimaging data: a review of the field. *Neuroinformatics* 2022;**20**:377–90.

Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. *J Stat Soft* 2007;**23**:1–46.

Tong J, Duan R, Li R *et al.* Robust-ODAL: learning from heterogeneous health systems without sharing patient-level data. *Pac Symp Biocomput* 2020;**25**:695–706.

Weir CB, Jan A. 2024. BMI classification percentile and cut off points. In: *StatPearls*. Treasure Island (FL): StatPearls Publishing.

WHO Multicentre Growth Reference Study Group. WHO child growth standards based on length/height, weight and age. *Acta Paediatr Suppl* 2006;**450**:76–85.

Wood SN. A simple test for random effects in regression models. *Biometrika* 2013a;**100**:1005–10.

Wood SN. On p-values for smooth components of an extended generalized additive model. *Biometrika* 2013b;**100**:221–8.

World Health Organization (WHO). WHO Child Growth Standards WHO Child Growth Standards: length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development. 2006.

Yin X, Zhu Y, Hu J. A comprehensive survey of privacy-preserving federated learning: a taxonomy, review, and future directions. *ACM Comput Surv* 2021;**54**:131:1–36.

Bioinformatics, 2026, 42, 1–12
https://doi.org/10.1093/bioinformatics/btaf625
Original Paper