UC Davis

UC Davis Previously Published Works

Title

A refined spirometry dataset for comparing segmented (piecewise) linear models to that of GAMLSS.

Permalink

https://escholarship.org/uc/item/0wd9m52g

Author

Zavorsky, Gerald

Publication Date

2024-12-01

DOI

10.1016/j.dib.2024.111062

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Data in Brief 57 (2024) 111062

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A refined spirometry dataset for comparing segmented (piecewise) linear models to that of GAMLSS

Gerald Stanley Zavorsky

Department of Physiology and Membrane Biology, Tupper Hall, Rm 4327, 1275 Med Sciences Drive, University of California, Davis, CA 95616, United States

ARTICLE INFO

Article history:
Received 25 September 2024
Accepted 15 October 2024
Available online 23 October 2024

Dataset link: Refined NHANES 2007-2012 spirometry dataset for the comparison of segmented (piecewise) linear models to that of GAMLSS (Reference data)

Keywords:
Biostatistics
Spirometry
Lung diffusing capacity
Restriction
Airway obstruction

ABSTRACT

Generalized Additive Models for Location, Scale, and Shape (GAMLSS) are widely used for developing spirometric reference equations but are often complex, requiring additional spline tables. This study explores the potential of Segmented (piecewise) Linear Regression as an alternative, comparing its predictive accuracy to GAMLSS and examining the agreement between the two methods. Spirometry data from nearly 16,600 patients, deemed Grade "A" and "B" acceptable from the NHANES 2007-2012 dataset, was analyzed. The dataset includes both nominal and scalar variables. Reference equations for forced expiratory volume in 1 s (FEV_1), forced vital capacity (FVC), and the ratio (FEV_1/FVC) were generated using GAMLSS (FEV_1, FVC, FEV_1/FVC), Segmented Linear Regression (FEV_1, FVC) and multiple linear regression (FEV_1/FVC). *K*-fold cross-validation was employed to compare prediction accuracy, using root-mean-square error (RMSE) and correlation coefficients. Agreement in classifying spirometric patterns (i.e. airway obstruction, restrictive spirometry pattern, mixed obstructive and restrictive disorder) was evaluated with the kappa statistic. This study uniquely compares the models by incorporating the lower limit of normal (LLN) using fitted z-scores of −1.645 or −1.96. The dataset is publicly

E-mail address: gszavorsky@ucdavis.edu

https://doi.org/10.1016/j.dib.2024.111062
2352-3409/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

available in SPSS (.sav) and .csv formats through the Mendeley Data repository.

© 2024 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Specifications Table

Subject	*Health and* Medical Sciences
Specific subject area	Pulmonary and Respiratory Medicine.
Type of data	Filtered spirometry data (continuous data) included 16,596 subjects who could perform technically acceptable spirometry maneuvers. Age, weight, height, sex, and self-selected race/ethnicity are reported. As well, results of Generalized Additive Models of Location, Scale and Shape (GAMLSS) & segmented linear regression (SLR) modeling are provided in nominal formal. The data is in SPSS (.sav) and .csv format.
Data collection	Spirometry was performed in the standing position in those 6 to 80 years of age unless the participant was physically impaired. Data is from the 2007 to 2012 National Health and National Examination Survey (NHANES) dataset on spirometry measures. Excluded from testing were examinees that had current chest pain or a physical problem with forceful expiration, were taking supplemental oxygen, had recent surgery of the eye, chest, or abdomen, had a recent heart attack, stroke, tuberculosis exposure, or had recently coughed up blood. Adults with a personal history of detached retina or a collapsed lung and children with painful ear infections were also excluded.
Data source location	The National Health and National Examination Survey (NHANES) follows a complex, multistage probability sampling design aimed to be representative of the U.S. population. The survey is conducted in various locations across the United States, chosen randomly to ensure geographic and demographic diversity. Still, specific cities are not typically disclosed in public datasets to protect participant confidentiality. However, NHANES sampling is designed to cover a broad range of urban and rural areas across different regions in the United States, ensuring nationwide representation rather than focusing on specific cities. The data is collected by mobile examination centers (MECs) traveling to different locations yearly. https://wwwn.cdc.gov/nchs/nhanes/
Data accessibility	Repository name: Mendeley Data Data identification number: 10.17632/dwjykg3xww.1 Direct URL to data: https://data.mendeley.com/datasets/dwjykg3xww/1 The data are in two formats: A) SPSS (.sav) format. Once the file is open you will find the labels for each parameter under the "VARIABLE VIEW" tab. B) .csv format.
Related research article	G.S. Zavorsky, Debunking the GAMLSS myth: Simplicity reigns in pulmonary function diagnostics, Respir Med (2024) https://doi.org/10.1016/j.rmed.2024.107836

1. Value of the Data

- **Representative of the U.S. population**: NHANES follows a complex, multistage sampling process, ensuring that the data collected reflects a broad cross-section of the U.S. population. This makes the data highly representative and valuable for understanding lung function trends and disparities across different demographics, including age, race, and ethnicity.
- **High data quality and standards**: The spirometry data in NHANES meets strict technical standards set by the American Thoracic Society (ATS) and the European Respiratory Society (ERS). This ensures that the measurements for forced expiratory volume (FEV_1), forced vital

capacity (FVC), and the FEV_1/FVC ratio are reliable and of high quality, suitable for clinical and research applications.

- **Extensive coverage of demographic groups**: The dataset includes spirometry data from over 16,000 participants from various racial and ethnic backgrounds (White, Black, Mexican American, Other Hispanic, and multi-racial). This allows for robust analysis of lung function differences across populations, contributing to developing race- or ethnicity-specific reference values, which are important for equitable healthcare.
- **Facilitates longitudinal and cross-sectional analysis**: The comprehensive nature of the dataset, collected over several years, makes it ideal for both cross-sectional studies and longitudinal analyses of lung function trends. It can be used to track how lung function changes with age, across different population groups, or in response to environmental and health factors.
- **Supports comparisons of modeling techniques**: This refined dataset has been used to compare different statistical modeling techniques, such as Generalized Additive Models for Location Scale and Shape (GAMLSS) and segmented linear regression (SLR) for the development of reference equations for use across the lifespan. This comparison allows researchers to explore complex and simplified lung function diagnostics approaches, ultimately improving the accuracy and accessibility of predictive models. Specifically, this dataset is unique in that includes the results of GAMLSS and SLR modeling using the lower limit of normal (LLN) for fitted z-scores of –1.645 or –1.96 for between model comparison.

2. Background

The NHANES 2007–2012 spirometry dataset is a key component of the National Health and Nutrition Examination Survey, a program designed to assess the health and nutritional status of the U.S. population through comprehensive, nationwide data collection. This dataset provides high-quality lung function measurements, including forced expiratory volume (FEV_1) and forced vital capacity (FVC), meeting the rigorous technical standards set by the American Thoracic Society (ATS) and European Respiratory Society (ERS). It includes data from over 16,000 participants across diverse racial and ethnic groups, enabling the development of accurate reference values for pulmonary diagnostics. The data is highly representative of the U.S. population and has been used to evaluate the effectiveness of different statistical modeling techniques, such as Generalized Additive Models for Location Scale and Shape (GAMLSS) and Segmented (Piecewise) Linear Regression (SLR), for predicting lung function. The dataset's extensive demographic coverage and adherence to strict data quality protocols make it an invaluable resource for researchers and clinicians focused on respiratory health.

The current guidelines advocate for GAMLSS-derived reference equations, which require spline tables [1–4]. However, previous research has shown that simpler SLR models – which do not require supplementary spline tables – provide similar predictive accuracies as GAMLSS for pulmonary diffusing capacity [5,6]. In those studies, predictive accuracies were defined by the root mean square error (RMSE), and obtained through repeated subsampling via the holdout method [5] or repeated *K*-fold cross-validation [6]. As such, it was thought that simple multiple linear regression or SLR would show comparable accuracies to GAMLSS when spirometric reference equations were developed.

This data article adds value by offering a comprehensive dataset that can be used for secondary analyses, comparative studies, and modeling efforts [7]. It enhances the primary findings of the published research study [8] by providing the raw data needed to validate and expand upon those results, thereby supporting broader research initiatives in pulmonary function diagnostics. Specifically, results from GAMLSS, SLR and multiple linear regression are provided in the dataset, and the R-code for GAMLSS and SLR replication is provided in the supplementary material section of this data-in-brief article.

3. Data Description

The refined dataset of 16,596 subjects is available in SPSS (.sav) format, which requires SPSS software to open [7]. However, there is an identical data file in .csv format [7]. Once the SPSS file is open, individuals can navigate to the "Variable View" tab. In this view, the "Label" column describes each parameter and its units (if applicable). For nominal variables, the "Values" column in the "Variable View" tab categorizes the variable. For example, "Sex" is a nominal variable; in the "Label" column, it is labeled as "Sex" (male or female). In the "Values" column, the coding is provided as 0 = female and 1 = male. The "Variable View" tab in the SPSS file of comprehensively explaining each variable and its corresponding coding.

The NHANES 2007-2012 spirometry dataset provides detailed lung function measurements for a representative sample of the U.S. population. Key variables measured include:

1. **Forced expiratory volume in 1 s (FEV_1)**: The amount of air a person can forcefully exhale in one second after taking a deep breath.
2. **Forced vital capacity (FVC)**: The total amount of air exhaled during a forced breath.
3. **FEV_1/FVC ratio**: A calculated ratio used to diagnose airflow obstruction and other lung disorders.
4. **Demographic information**: Age, sex, race/ethnicity (White, Black, Mexican American, Other Hispanic, multi-racial), and other sociodemographic variables are included, providing context for analyzing lung function across different groups.
5. **Body measurements**: Variables like height, weight, and Body Mass Index (BMI) are also recorded, as these factors influence lung function.
6. **Z-scores**: The dataset includes z-scores for FEV_1, FVC, and FEV_1/FVC ratio, which are standardized scores comparing an individual's values to population norms based on Global Lung Function Initiative (GLI) reference equations (i.e., race-neutral equations) [2].
7. **Modeling results**: Airflow obstruction, possible restriction, or mixed disorder results – as defined elsewhere [9] – were classified by the developed GAMLSS and SLR models for FEV_1 & FVC, and, then GAMLSS and multiple linear regression for the FEV_1/FVC ratio. This is the unique feature of this refined dataset.

All spirometry measurements meet or exceed the technical acceptability standards set by ATS and ERS in 2005 [10], ensuring the reliability and accuracy of the data. The dataset covers individuals aged 6–80, providing a wide age range for analysis of lung function across the lifespan. Additionally, the dataset facilitates comparisons across racial and ethnic groups, making it a valuable resource for understanding population-specific lung function trends and disparities.

4. Experimental Design, Materials and Methods

This is a cross-sectional study, as the spirometry data was collected at a single time. This snapshot captures a variety of ages without considering individual changes over time. Yet, the design allows for analysing trends and associations related to age in lung function across a large sample.

This dataset [7] includes only NHANES participants who met "A" and "B" grade acceptability standards as defined by the National Institute for Occupational Safety and Health (NIOSH). The criteria for spirometry values were as follows: for an "A" grade, three acceptable curves were required, with the largest and second-largest values within 100 ml and no more than a 50 ml difference from the last maneuver. Three acceptable curves were required for a "B" grade, with the largest and second-largest values within 150 ml, meeting the minimum criteria set by the American Thoracic Society's 2005 guidelines [10]. Based on these criteria, 16,596 subjects were retained from an initial pool of around 30,000.

The dataset included both males and females, ranging in age from 6 to 80 years, across five racial/ethnic categories (White, Black, Mexican American, Other Hispanic, and multi-racial). Since

the focus was not on comparing spirometry differences across races or ethnicities, but rather on comparing the RMSE between different modeling techniques, all racial/ethnic groups were analyzed together as one pooled group.

Reference equations for FEV_1, FVC, and the FEV_1/FVC ratio were developed using three modeling techniques: GAMLSS and segmented linear regression (SLR) for FEV_1 and FVC, and multiple linear regression and GAMLSS for the FEV_1/FVC ratio. SLR was not used for the FEV_1/FVC ratio, as no significant breakpoint was identified across the lifespan.

All analyses were performed in the R programming environment, with age, height, and weight identified as key predictors.

4.1. Statistical Analyses

Two primary statistical models were compared: GAMLSS and segmented linear regression. For prediction accuracy, *K*-fold cross-validation (with 10 folds) was used to estimate the root-mean-square error (RMSE) and correlation coefficients for FEV_1, FVC, and FEV_1/FVC. The agreement between these two models in classifying spirometric patterns (e.g., airflow obstruction, restrictive spirometry) was assessed using the unweighted Kappa statistic**.** Additionally, Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) were employed to compare the goodness-of-fit between models, and paired *t*-tests were used to evaluate differences in z-scores produced by the models. McNemar's test was applied to assess differences in the classification of spirometry patterns, and multiple comparison corrections were made using the Benjamini-Hochberg procedure to control for false discovery rates [11].

These statistical analyses ensured a thorough comparison between the simpler segmented regression and the more complex GAMLSS models in predicting lung function across the dataset.

The GAMLSS models [12] and SLR [13] were implemented using R CRAN packages. All the R-package names are provided in the supplementary file to this data article that includes the "bare bones" R-code. Statistical significance was defined as $p < 0.05$.

Limitations

The refined dataset, available through the Mendeley Data repository [7] has certain limitations that should be acknowledged. Firstly, only a portion of the NHANES data was included, as the dataset was restricted to individuals who successfully completed technically valid spirometry tests. As a result, the exclusion of participants who could not perform these maneuvers might have affected the outcomes, potentially leading to different conclusions if the full NHANES dataset had been analyzed. Additionally, the decision to combine racial and ethnic groups may have concealed important differences in lung function between populations, limiting the findings' relevance for specific demographic groups. Finally, since the data is cross-sectional, it does not allow for the analysis of lung function changes over time, which could provide more detailed insights into the effects of factors like growth, aging, and environmental exposures.

Ethics Statement

Ethics approval was unnecessary as this was de-identified publicly available data from NHANES.

Data availability

Refined NHANES 2007-2012 spirometry dataset for the comparison of segmented (piecewise) linear models to that of GAMLSS (Reference data) (Mendeley Data)

CRediT Author Statement

Gerald Stanley Zavorsky: Conceptualization, Software, Writing – original draft, Formal analysis, Methodology, Investigation, Writing – review & editing.

Acknowledgments

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

The research article related to this publication is found in the journal Respiratory Medicine [8].

Declaration of Competing Interest

Gerald S. Zavorsky is a member of the Global Lung Function Initiative (GLI) Network. The GLI Network has published reference equations for spirometry, DLCO, and static lung volumes using GAMLSS models. Gerald S. Zavorsky is the current co-chair of the European Respiratory Society Task Force on interpreting pulmonary diffusing capacity for nitric oxide.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2024.111062.

References

[1] P.H. Quanjer, S. Stanojevic, T.J. Cole, X. Baur, G.L. Hall, B.H. Culver, P.L. Enright, J.L. Hankinson, M.S. Ip, J. Zheng, J. Stocks, Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations, Eur. Respir. J. 40 (6) (2012) 1324–1343, doi:10.1183/09031936.00080312.

[2] C. Bowerman, N.R. Bhakta, D. Brazzale, B.R. Cooper, J. Cooper, L. Gochicoa-Rangel, J. Haynes, D.A. Kaminsky, L.T.T. Lan, R. Masekela, M.C. McCormack, I. Steenbruggen, S. Stanojevic, A race-neutral approach to the interpretation of lung function measurements, Am. J. Respir. Crit. Care Med. 207 (6) (2023) 768–774, doi:10.1164/rccm.202205-0963OC.

[3] G.L. Hall, N. Filipow, G. Ruppel, T. Okitika, B. Thompson, J. Kirkby, I. Steenbruggen, B.G. Cooper, S. Stanojevic, Official ERS technical standard: global lung function initiative reference values for static lung volumes in individuals of European ancestry, Eur. Respir. J. 57 (3) (2021), doi:10.1183/13993003.00289-2020.

[4] S. Stanojevic, B.L. Graham, B.G. Cooper, B.R. Thompson, K.W. Carter, R.W. Francis, G.L. Hall, Official ERS technical standards: global lung function initiative reference values for the carbon monoxide transfer factor for Caucasians, Eur. Respir. J. 50 (3) (2017), doi:10.1183/13993003.00010-2017.

[5] G.S. Zavorsky, J. Cao, Reference equations for pulmonary diffusing capacity using segmented regression show similar predictive accuracy as GAMLSS models, BMJ Open Respir. Res. 9 (1) (2022), doi:10.1136/bmjresp-2021-001087.

[6] L.G. Gochicoa-Rangel, A. De-Los-Santos Martinez, A. Reyes-Garcia, D.M. Briseno, M.H. Vargas, I. Lechuga-Trejo, C. Guzman-Valderrabano, L. Torre-Bouscoulet, G.S. Zavorsky, Reference equations for DLNO & DLCO in Mexican hispanics: influence of altitude and race, BMJ Open Respir. Res. (2024) [in press], doi:10.1136/bmjresp-2024-002341.

[7] G. Zavorsky, "Refined NHANES 2007-2012 spirometry dataset for the comparison of segmented (piecewise) linear models to that of GAMLSS", Mendeley Data, V1, doi: 10.17632/dwjykg3xww.1, V 4, 2024.

[8] G.S. Zavorsky, Debunking the GAMLSS myth: simplicity reigns in pulmonary function diagnostics, Respir. Med. (2024) 107836, doi:10.1016/j.rmed.2024.107836.

[9] S. Stanojevic, D.A. Kaminsky, M.R. Miller, B. Thompson, A. Aliverti, I. Barjaktarevic, B.G. Cooper, B. Culver, E. Derom, G.L. Hall, T.S. Hallstrand, J.D. Leuppi, N. MacIntyre, M. McCormack, M. Rosenfeld, E.R. Swenson, ERS/ATS technical standard on interpretive strategies for routine lung function tests, Eur. Respir. J. 60 (1) (2022), doi:10.1183/13993003.01499-2021.

[10] R. Pellegrino, G. Viegi, V. Brusasco, R.O. Crapo, F. Burgos, R. Casaburi, A. Coates, C.P. van der Grinten, P. Gustafsson, J. Hankinson, R. Jensen, D.C. Johnson, N. MacIntyre, R. McKay, M.R. Miller, D. Navajas, O.F. Pedersen, J. Wanger, Interpretative strategies for lung function tests, Eur. Respir. J. 26 (5) (2005) 948–968, doi:10.1183/09031936.05.00035205.

[11] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, Ann. Stat. 29 (4) (2001) 1165–1188.

[12] R.A. Rigby, D.M. Stasinopoulos, Generalized additive models for location, scale, and shape, Appl. Stat. 54 (3) (2005) 507–554.

[13] V.M.R. Muggeo, Segmented: an R package to fit regression models with broken-line relationships, R News 8/1 (2008) 20-25.