Spirometry quality predictors in a large multistate prospective study

Kaitlyn G. Lawrence*, W. Braxton Jackson II*, Steven Ramsey, Richard K. Kwok, Lawrence S. Engel, Matthew D. Curry, Dale P. Sandler

*Co-first authors

ATS/ERS: American Thoracic Society/European Respiratory Society
CI: Confidence interval
DWH: Deepwater Horizon
FEV1: Forced expiratory volume in one second
FVC: Forced vital capacity
GuLF STUDY: Gulf Long-Term Follow-up Study
mL: milliliter

Running head: Predictors of spirometry quality

Notes: Kaitlyn G. Lawrence and W. Braxton Jackson contributed equal work as co-first authors of this manuscript.

Funding: This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES 102945). The authors declare no conflicts of interest.

Author affiliations: Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC (Kaitlyn G. Lawrence, Lawrence S. Engel, Richard K. Kwok, Dale P. Sandler); Social & Scientific Systems, Inc., Durham, NC (W. Braxton Jackson II, Matthew D. Curry); Office of the Director, National Institute of Environmental Health Sciences, Research Triangle Park, NC (Richard K. Kwok); Dept. of Epidemiology, UNC Chapel Hill, Chapel Hill, NC (Lawrence S. Engel)

ABSTRACT

Background: The Gulf Long-Term Follow-up (GuLF) Study is a prospective cohort study of health effects associated with oil spill response and clean-up following the 2010 *Deepwater Horizon* Disaster *(DWH).* As part of the study, spirometry testing of lung function was carried out in home visits across multiple states. Few studies have described factors associated with spirometry test failure in field-based settings. **Objective:** Our objective was to identify what factors, if any, predicted test failure among GuLF Study participants who completed spirometry testing in a non-traditional setting. **Methods:** Trained examiners administered spirometry (May 2011 – May 2013) to 10,019 participants living in US Gulf States (LA, MX, TX, AL, FL) using an Easy-on ultrasonic spirometer. We applied American Thoracic Society/European Respiratory Society quality criteria to determine quality test failure and identified factors predictive of failure using both a Stepwise and a LASSO model. We calculated odds ratios and 95% confidence intervals (CIs) for associations of selected factors with test failure. **Results:** Among GuLF Study participants who conducted spirometry Black participants (OR: 1.39, 95% CI: 1.23,1.56); males (OR:1.61, 95% CI: 1.41,1.83); and those making less than $20,000 per year (OR: 1.45, 95% CI: 1.26,1.69) were more likely to fail quality testing, while those who were obese were less likely to fail (OR: 0.61, 95% CI: 0.42,0.89). **Conclusion**: Field-based studies involving spirometry should identify and account for participant factors that may influence test failure. Coaching that is tailored to those less likely to have experience with spirometry may help reduce test failure rates.

INTRODUCTION

In 2011, National Institute of Environmental Health Sciences investigators launched the Gulf Long-Term Follow-up (GuLF) Study, a large prospective cohort study aimed at evaluating both short- and long-term health effects following the 2010 *Deepwater Horizon* Disaster *(DWH)* [1, 2]. Primary environmental hazards released into the environment included crude oil, chemical dispersants [3], and combustion by-products from burning crude oil and flaring natural gas[4]. Based on previous toxicologic knowledge of chemical hazards and related target organs, as well as previous epidemiologic findings [5], investigators considered respiratory health a primary health endpoint of interest for the study. To evaluate lung function in this cohort, trained examiners administered spirometry testing during an in-home exam carried out among eligible participants living in the five Gulf states of Alabama (AL), Mississippi (MI), Florida (FL), Louisiana (LA), and Texas (TX). Spirometry, a type of pulmonary function test that measures lung capacity and airflow [6], is an objective measure of respiratory health widely used in clinical settings. In research settings, spirometry is an attractive choice given that tests are inexpensive and non-invasive [6]. Quality control of spirometry ensures that the lung function parameters generated are reliable and therefore valid for accurate comparison between individuals in a given study. Quality control efforts were considered especially important in the GuLF Study given its home rather than clinic setting.

Spirometry quality depends on three factors: 1) accurate and precise instrumentation; 2) effective test administration (coaching) by the examiner/technician; and 3) a concerted effort on the part of the participant [6]. Of these factors (device, examiner, and participant), the examiner has the greatest influence on test quality through effective coaching on maximal inhalation, and forceful and complete exhalation [7]. Additionally, there are three key aspects of coaching which include full inhalation, forceful exhalation, and complete exhalation, with incomplete inhalation as the most common reason for low quality tests [8]. However, many GuLF Study participants were drawn from a community that faces social and economic barriers that may additionally threaten their respiratory health and nearly half (49.4%) of study participants report low utilization to medical care. These traits are common in other large post-disaster cohorts that have also measured spirometry (such as the World Trade Center) and may pose risk of selection bias for those excluded due to test failure. For example, lack of familiarity with protocols and subsequent potential discomfort with testing administrators may reduce the proportion of persons who achieve passing quality tests.

To reduce misclassification of spirometry results, testing sessions are typically evaluated for quality against a set of standard criteria for technical acceptability and reproducibility established by the American Thoracic Society and European Respiratory Society (ATS/ERS) [9, 10]. These standards were set up such that 90% of experienced examiners could pass them [11]. Multiple studies have shown that high quality spirometry can be achieved in research and clinical practice through comprehensive quality control programs that include technician training, quality scoring of testing, and technician feedback and retraining [8, 12].

Still, several studies have found that certain participant factors increase the likelihood of spirometry test failure while others have experienced near perfect quality scores. A study of older adults found that spirometry failure was associated with lower cognitive function among adults[13] while another study among men in Norway found spirometry failures on both acceptability and reproducibility criteria were more prevalent in never smokers, single men, and those with respiratory symptoms[14]. These studies were carried out in a clinical setting whereas the GuLF Study testing was carried out in the homes of participants. Thus, the objective of our study was to identify what factors, if any, predict spirometry test failure among GuLF Study participants who conducted spirometry testing in a non-traditional setting.

METHODS

Study Design

Participants were initially enrolled via telephone interview between March 2011-March 2013 (N=32,608). **Figure 1** shows a map of participant residences at enrollment. Of those enrolled, 11,193 English- or Spanish- speaking participants living in the five US Gulf states (Texas, Louisiana, Mississippi, Alabama, Florida) completed a home visit between May 2011-May 2013. During the home visit, trained examiners (N=49) administered a more detailed health questionnaire, collected biologic samples and anthropometric measures, and administered pulmonary function tests [2]. All home visit participants provided verbal consent at the enrollment telephone interview and written informed consent for all activities conducted during the home visit. This study received approval from the National Institute of Environmental Health Sciences Institutional Review Board.

Spirometry

Trained examiners coached the participants on using an ultrasonic spirometer (Easy-On spirometer; ndd Medical Technologies; Andover MA). This spirometer has built-in quality control software that provides real-time feedback to the examiner about acceptability and reproducibility of the test session. Examiners received periodic feedback about their quality scores and, if necessary, booster training to improve their performance. The spirometry measures of interest were the forced vital capacity (FVC), and the forced expiratory volume in one second (FEV_1) and the FEV_1/FVC ratio, derived from these maneuvers [12, 15, 16]. The forced vital capacity is the total volume of air that can be exhaled during a maximal forced expiration effort whereas the FEV_1 is the amount of air you can force from your lungs in one second.

Among home visit participants (n=11,193), some did not complete a spirometry test due to refusal (n=110), early visit termination (n=75), or a technical problem (n=74) and others were not eligible for spirometry due to an American Thoracic Society or study specific medical exclusion criteria (n=716). For a small number, the reason for missing spirometry test data was not recorded (n=178). A total of 10,040 participants were included. Participants performed pre-bronchodilator spirometry tests in their own home, seated, with a nose clip. Following 2005 ATS/ERS guidelines [6], participants conducted the forced vital capacity maneuver until either achieving three acceptable maneuvers or completing 8 maneuvers overall. Spirometry tests were considered 'acceptable' if they were free from artifacts (i.e. cough during the first second of exhalation, glottis closure that affected the measurement, early termination or cut-off, less than maximal effort during the test, leak, or obstructed mouthpiece); had good starts (extrapolated volume <5% of FVC or 0.15L, whichever was greater); and had satisfactory exhalation (duration of ≥6 seconds or a 1 second plateau in the volume-time curve, or if the subject could not, or should not continue to exhale).

10,019 spirometry test results were overread by a spirometry expert and assigned a final quality score. Specifically, quality scores (A, B, C, D, or F;) were assigned to each FEV_1 and FVC curve based on the 2005 ATS/ERS (**Table 1**). Of the 10,019 participants who took a spirometry test at the home visit, 75% received a passing quality score, defined as A, B, or C, while 25% received a failing quality score defined as D or F, following the standardized ATS/ERS criteria for both FEV_1 and FVC. The analytic sample included 8,466 participants with full predictor information. The distributions of raw spirometry measurements and percent predicted values as calculated by Quanjer et al.[17] for those who failed quality standards and those who passed are shown in **Table 2a and 2b,** respectively.

Participant characteristics

Participant characteristics were self-reported at the enrollment interview and included: age, gender (female, male), self-reported race (African American/Black, white, other), Hispanic ethnicity (Hispanic, non-Hispanic), annual household income (<$20,000, $20,000-$50,000, >$50,000), educational attainment (less than High School/Equivalent, High School Diploma/GED, Some College/2 year degree, 4 year college graduate or more), pre-spill lung disease (asthma, bronchitis, or emphysema; yes or no), smoking status (never, former, light current (≤20 cigarettes per day), heavy (>20 cigarettes per day) current), always/usually worried about future health (yes, no), always/usually worried about paying rent or buying food (yes, no), previous oil industry experience (yes, no), previous oil spill cleanup experience (yes, no), wheeze (yes, no), tightness in chest (yes, no), and shortness of breath (yes, no). At the home visit, to reduce technical errors of measurement, trained examiners measured height and weight in triplicate, and the average of these measures was used to calculate BMI (kg/m^2; <18.5 (underweight), 18.5-24.9 (normal weight), 25-29.9 (overweight), ≥30 (obese)).

Oil spill response and cleanup work

A structured interview at enrollment was used to collect information on jobs/tasks performed as a *DWH* response worker including information on exposure to burning crude oil and natural gas. A detailed description of oil spill response and cleanup work exposures can be found elsewhere [18]. Those who worked at least one day on any job or task related to the oil spill were classified as workers and all others were classified as non-workers. For those classified as workers, industrial hygienists grouped jobs and tasks, based on an approximate intensity of exposure reported as level of total hydrocarbons, into one of six broad hierarchical job classes: response (highest exposure), operations, clean-up on water, decontamination, clean-up on land, and support work (lowest exposure). Using exposure measurements taken during the oil spill response, daily arithmetic means in parts per million (ppm) of estimated THC exposure were developed for all possible combinations of jobs, tasks, locations, and time periods worked. These estimates were linked to participant-reported oil spill work histories using a job-exposure matrix.

We used the maximum daily exposure across all days worked to define an ordinal THC exposure level scale using a pseudo-log scale based on the empirical range of job/task specific-exposures as follows: 1 (≤0.29 ppm); 2 (0.30-0.99 ppm); 3 (1.00-2.99 ppm), 4 (≥3.00 ppm). Potential exposure to burning oil/gas was assessed from self-reported responses to questions on the task of *in situ* burning and working near the wellhead where controlled oil burning/gas flaring occurred (based on proximity to the well-head and/or name of the vessel they worked on). Workers were classified into high, medium, low, and no potential exposure to burning oil/gas.

Examiner characteristics

Examiner characteristics that we analyzed included gender (female/male), examiner-participant gender match (female/female, male/male, female/male, male/female) and the examiner's state of hire (LA, MI, FL, TX, AL). This sometimes differed from that of the participant's residential location as examiners were asked to travel to home visits proximal to their own location of service including traveling across state lines if necessary. We also attempted to analyze examiner credentials but there was little variability as most technicians had a similar certification. Unfortunately, we did not have information on the examiner's years of experience in the field or the examiner's race, although we hypothesized that these characteristics might influence test quality.

Statistical Analysis

We first determined the proportion of participants who failed spirometry quality criteria (score of failing= "D" or "F"). We selected a group of participant and examiner variables *a priori* based on factors known to potentially influence spirometry. We used two methods for exploring quality predictors: stepwise selection, and Least Absolute Shrinkage and Selection Operator (LASSO) regression as described by John et al. to select predictors of spirometry quality in our sample using the GLMSELECT procedure in SAS (version 9.4) [19]. This method included the use of Schwarz Bayesian Criterion for variable selection [20, 21]. Finally, we used multivariable logistic regression to calculate adjusted odds ratios and 95% confidence intervals (CIs) for each selected predictor to assess the directionality and magnitude of the relationship between predictor and spirometry test failure. Each model was mutually adjusted for all other variables selected.

RESULTS

Participant and examiner characteristics are shown for those failing or meeting quality metrics in **Table 3**. We observed some qualitative proportional differences between those who passed and those who failed spirometry testing. Those who failed were younger (25% vs 18% <30 years old), more likely to be male (83% vs 76%), to report African American/Black race (42% vs 32%), and have an annual household income <$20,000 per year (46% vs 38%).

The stepwise model selected as predictive of quality test failure participant gender, self-reported race, income, and BMI but not spill response and cleanup work exposures of worker status, estimated exposure to total hydrocarbons, or burning oil/flaring natural gas. **Table 4 shows** the results of the logistic regression-estimated adjusted odds ratios and 95% CIs associated with each STEPWISE-selected variable. Compared to women, men were at higher odds of failing the spirometry test (OR:1.61, 95% CI: 1.41,1.83). Compared to white participants, African American/Black participants were also more likely to fail spirometry (OR:1.39, 95% CI: 1.23,1.56). Compared to participants with a normal weight (BMI=18.5-24.9), those who were overweight (BMI=25-29.9) were less likely to fail spirometry testing (OR: 0.58, 95% CI: 0.40,0.85) as were those who were obese (BMI ≥30) (OR: 0.61, 95% CI: 0.42, 0.89). Low SES indicated by lower income placed participants at higher risk of failed spirometry testing. No other participant characteristics were shown to predict spirometry. In the LASSO analysis, self-reported race was the only predictor of spirometry test failure.

DISCUSSION

This study sought to describe pass/failure rates of spirometry testing administered in the field at home visits across a large geographic area and to identify any predictors associated with spirometry testing quality. Of the 10,019 GuLF Study participants who conducted spirometry testing, 75% met passing acceptability and reproducibility criteria. Participant characteristics that predicted spirometry failure in our analysis included gender, self-reported race, income, and BMI. We used stepwise selection, and LASSO to compare results, and found that self-reported race was identified as a predictor, regardless of which selection method was used even after including socioeconomic factors associated with self-reported race. Given the long history of inappropriate correction for self-reported race in pulmonary function measurements[22] we want to make clear that we adjust for self-reported race here as an indicator of social factors (not genetic factors) such as socioeconomic status and/or systemic racism. Thus, with self-reported race as a primary indicator of spirometry test failure, we hypothesis that this is attributable to residual confounding captured by self-reported race that is not captured by our other socioeconomic status indicators of income or education.

Spirometry is known to be highly influenced by participant effort, participant-examiner cooperation, and examiner proficiency as a spirometry “coach” [6, 23]. Standardized criteria are common solutions to identify and help mitigate the impact of this unwanted variability[24]. However, there is little knowledge about the degree to which this evaluation approach can help to minimize undesirable influence in a cohort such as the GuLF Study, with its large geographical range, need for multiple technicians, and location in a region with low access/experience with healthcare. Participants who lacked regular access to health care are likely to have had less familiarity with medical tests performed.

Unlike prior studies, GuLF Study investigators faced unique challenges in obtaining spirometry including the large geographic study area (the entire US Gulf region), which hindered the ability to provide regular in-person oversight of agents in subjects' homes, as might be done in a clinic, the lack of a controlled testing environment, and the accelerated time frame given the nature of the disaster (making it harder for investigators to pick and retain only the very best examiners). Investigators also had to prioritize examiners who had a range of skills (e.g. phlebotomy), given the diverse tasks required by these personnel during the home visit exam. GuLF Study investigators monitored examiner performance, and did supplemental trainings as needed, but this was rarely done in person (due to the large geographic area the study staff were deployed to), and so routine in-person oversight of a home visit examiner administering tests in the home setting was rare. However, participants did have in-person training sessions and refresher trainings as needed throughout the study period.

Our study findings agree with some prior research suggesting that being male and having a lower socioeconomic status are associated with lower quality spirometry tests [8, 25]. Other studies, such as those following the *World Trade Center* disaster, did not identify participant or examiner factors that influenced spirometry failure [9], potentially because spirometry in those studies was conducted in a central clinic. However, findings from clinic-based studies of cohorts with lung disease or risk factors for lung disease have concluded that a rigorous quality control program can overcome negative influences of participant spirometry factors including participant age [25]. One exception was a field based study of children and adolescents that found no participant factors associated with spirometry quality [26].

Previous research suggests that failure to perform reproducible spirometry may itself be an indicator of respiratory ill health [27]. If there is a causal relationship between an exposure of interest and respiratory health, spirometry quality is an important point of consideration for collider stratification bias. Though we could not directly test this hypothesis using measured spirometry (because those who failed could not have their lung function reliably measured) we did not find evidence showing that respiratory symptoms were associated with spirometry test reproducibility [28]. Nonetheless, we agree with others that including participants regardless of test failure or adopting a more liberal criteria of reproducibility[28, 29] could help to minimize introduction of bias. At the very least, researchers should consider the implications of and trade-offs in selecting inclusion criteria. In other work, we have shown that including participants with lower quality scores did not materially affect exposure-outcome relationships [30].

While our aim was to describe spirometry test failure among GuLF Study participants specifically, these findings add important insight for other large-scale epidemiologic studies and consideration of spirometry performance conducted outside of a clinical setting. Previous work describing the GuLF Study cohort overall, also suggests sociodemographic differences between those eligible for a home visit compared to those, which may additionally limit the generalizability of findings: Of those eligible for a home visit (N=25,304) those who completed a home visit (N=11,193) showed a higher participation rate of self-reported Black participants (27.4% vs. 34.7%) and those making <$20,000 per year (30.6% vs. 37.2%)[2]. A major strength of our quality assessment is the amount of information we collected on participant characteristics that allowed us to assess a wide range of factors, including socioeconomic status measures, demographics, and exposure and health related information, in relation to spirometry failure. Another strength of our analysis is the large sample size, which provided sufficient power to assess spirometry quality across a large number of mutually adjusted participant characteristics. By examining these factors, we are able to provide practical guidance for future cohort studies conducting spirometry in the home setting. Specifically, we suggest that future large, geographically distributed studies pay extra attention to participants in potential subgroups at greater risk of spirometry failure, as shown here. This has consequences for who can contribute spirometry data to an epidemiologic study.

Limitations of our study include the fact that home visit examiners, though well-trained in spirometry, also had to have expertise across multiple anthropometric and sample collection methods, thereby potentially limiting outstanding examiners who just specialized in spirometry. Given the limited number of examiners used in our study, we may have been underpowered to look across multiple examiner characteristics. Additionally, with more examiners, we could have looked at predictors of trajectories of scores attributed to examiners with specific characteristics (e.g. those examiners who had spirometry scores which were consistently good; improved over time; worsened over time; or were consistently poor). Such an analysis could help identify which examiners are worth retraining and which are not. Aside from training, assessment of self-reported race or ethnicity participant-examiner discordance may provide deeper insight into coaching effects given that investigators were made aware of a few instances of perceived overt racism during the field work (though examiners were instructed to leave if they felt they were in danger). However, we did not collect information on the race/ethnicity of the home visit examiners.

Our study population includes those living in the Gulf coast states who were involved in the response effort following the *DWH* disaster. The socio-demographic characteristics in this population, including those with potentially very low experience with spirometry/ health examinations may differ significantly from that of other studies of spirometry quality. Thus, results may not be directly comparable to other study populations with spirometry testing. This paper helps identify factors that testers or trainers should be aware of that can reduce the quality of spirometry tests and should receive attention during training/testing as the most common cause of erroneous results.

Conclusions

This study identified participant characteristics associated with spirometry test failure that maybe generalizable to other large-scale and/or home examination settings (i.e., outside of a controlled clinical setting). Participants self-reporting as black race, male gender, and annual household income <$50,000 were significantly more likely to fail spirometry quality testing, whereas participants with ≥25 BMI were less likely to fail spirometry quality testing. We did not find that oil spill related exposures, unique to our cohort, were predictive of spirometry test failure. Although we did not evaluate why these traits predicted spirometry failure, that we identified such differences is important for future work. Future study designs and analytic protocols should consider and mitigate the potential for selection and information bias [31]. Investigators of studies with pulmonary function as a main outcome of interest should pay close attention to participant characteristics that predict test failure.

REFERENCES

1. Engel, L.S., et al., *The Gulf Long-Term Follow-Up Study (GuLF STUDY): Biospecimen collection at enrollment.* J Toxicol Environ Health A, 2017. **80**(4): p. 218-229.
2. Kwok, R.K., et al., *The GuLF STUDY: A Prospective Study of Persons Involved in the Deepwater Horizon Oil Spill Response and Clean-Up.* Environ Health Perspect, 2017. **125**(4): p. 570-578.
3. National Commission on the BP Deepwater Horizon Oil Spill and Offshore Drilling, *Deep Water: The Gulf Oil Disaster and the Future of Offshore Drilling, Report to the President*. 2011: Washington, DC.
4. Allen, A.A., et al., *The Use of Controlled Burning during the Gulf of Mexico Deepwater Horizon MC-252 Oil Spill Response.* International Oil Spill Conference Proceedings, 2011. **2011**(1): p. abs194.
5. Aguilera, F., et al., *Review on the effects of exposure to spilled oils on human health.* Journal of Applied Toxicology, 2010. **30**(4): p. 291-301.
6. Miller, M.R., et al., *Standardisation of spirometry.* European Respiratory Journal, 2005. **26**(2): p. 319-338.
7. Haynes, J.M., *Quality Assurance of the Pulmonary Function Technologist.* Respiratory Care, 2012. **57**(1): p. 114-126.
8. Enright, P.L., *How to make sure your spirometry tests are of good quality.* Respiratory care, 2003. **48**(8): p. 773-776.
9. Enright, P.L., et al., *Quality of spirometry performed by 13,599 participants in the world trade center worker and volunteer medical screening program.* Respiratory Care, 2010. **55**(3): p. 303-309.
10. Miller, M.R., et al., *Improved survival prediction from lung function data in a large population sample.* Respiratory Medicine, 2009. **103**(3): p. 442-448.
11. Enright, P. and T. Schermer, *Don't pay for poor quality spirometry tests.* Primary Care Respiratory Journal, 2013. **22**(1): p. 15-16.
12. Ackermann-Liebrich, U., et al., *Lung function and long term exposure to air pollutants in Switzerland. Study on Air Pollution and Lung Diseases in Adults (SAPALDIA) Team.* American Journal of Respiratory and Critical Care Medicine, 1997. **155**(1): p. 122-129.
13. Turkeshi, E., et al., *Predictors of poor-quality spirometry in two cohorts of older adults in Russia and Belgium: a cross-sectional study.* NPJ primary care respiratory medicine, 2015. **25**(1): p. 1-6.
14. Humerfelt, S., et al., *Predictors of spirometric test failure: a comparison of the 1983 and 1993 acceptability criteria from the European Community for Coal and Steel.* Occupational and Environmental Medicine, 1995. **52**(8): p. 547-553.
15. Buist, A.S., et al., *International variation in the prevalence of COPD (The BOLD Study): a population-based prevalence study.* The Lancet. **370**(9589): p. 741-750.
16. Kan, H.D., et al., *Traffic exposure and lung function in adults: the Atherosclerosis Risk in Communities study.* Thorax, 2007. **62**(10): p. 873-879.
17. Quanjer, P.H., et al., *Multi-ethnic reference values for spirometry for the 3-95-yr age range: the global lung function 2012 equations.* Eur Respir J, 2012. **40**(6): p. 1324-43.
18. Stewart, P.A., et al., *Development of a total hydrocarbon ordinal job-exposure matrix for workers responding to the Deepwater Horizon disaster: The GuLF STUDY.* J Expo Sci Environ Epidemiol, 2018. **28**(3): p. 223-230.

19. Cohen, R.A. *Introducing the GLMSELECT procedure for model selection*. in *Proceedings of the Thirty-First Annual SAS Users Group International Conference*. 2006. Citeseer.
20. Kass, R.E. and L. Wasserman, *A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion.* Journal of the american statistical association, 1995. **90**(431): p. 928-934.
21. Bhattacharya, S. and P.D. McNicholas, *A LASSO-penalized BIC for mixture model selection.* Advances in Data Analysis and Classification, 2014. **8**(1): p. 45-61.
22. Braun, L., *Spirometry, measurement, and race in the nineteenth century.* Journal of the History of Medicine and Allied Sciences, 2005. **60**(2): p. 135-169.
23. Upton, M.N., et al., *Improving the quality of spirometry in an epidemiological study.* Public Health, 2000. **114**(5): p. 353-360.
24. Eaton, T., et al., *Spirometry in primary care practice: the importance of quality assurance and the impact of spirometry workshops.* Chest, 1999. **116**(2): p. 416-423.
25. Bellia, V., et al., *Quality control of spirometry in the elderly. The SA.R.A. study. SAlute Respiration nell'Anziano = Respiratory Health in the Elderly.* Am J Respir Crit Care Med, 2000. **161**(4 Pt 1): p. 1094-100.
26. Enright, P.L., et al., *Quality of Spirometry Test Performance in Children and Adolescents*: Experience in a Large Field Study.* Chest, 2000. **118**(3): p. 665-671.
27. Eisen, E.A., *Standardizing spirometry: problems and prospects.* Occup Med, 1987. **2**(2): p. 213-25.
28. Eisen, E.A., et al., *Selection effects of repeatability criteria applied to lung spirometry.* American Journal of Epidemiology, 1984. **120**(5): p. 734-42.
29. Hankinson, J.L. and K.M. Bang, *Acceptability and reproducibility criteria of the American Thoracic Society as observed in a sample of the general population.* Am Rev Respir Dis, 1991. **143**(3): p. 516-21.
30. Gam, K.B., et al., *Lung Function in Oil Spill Response Workers 1-3 Years After the Deepwater Horizon Disaster.* Epidemiology, 2018. **29**(3): p. 315-322.
31. Cole, S.R. and M.A. Hernan, *Constructing inverse probability weights for marginal structural models.* Am J Epidemiol, 2008. **168**(6): p. 656-64.

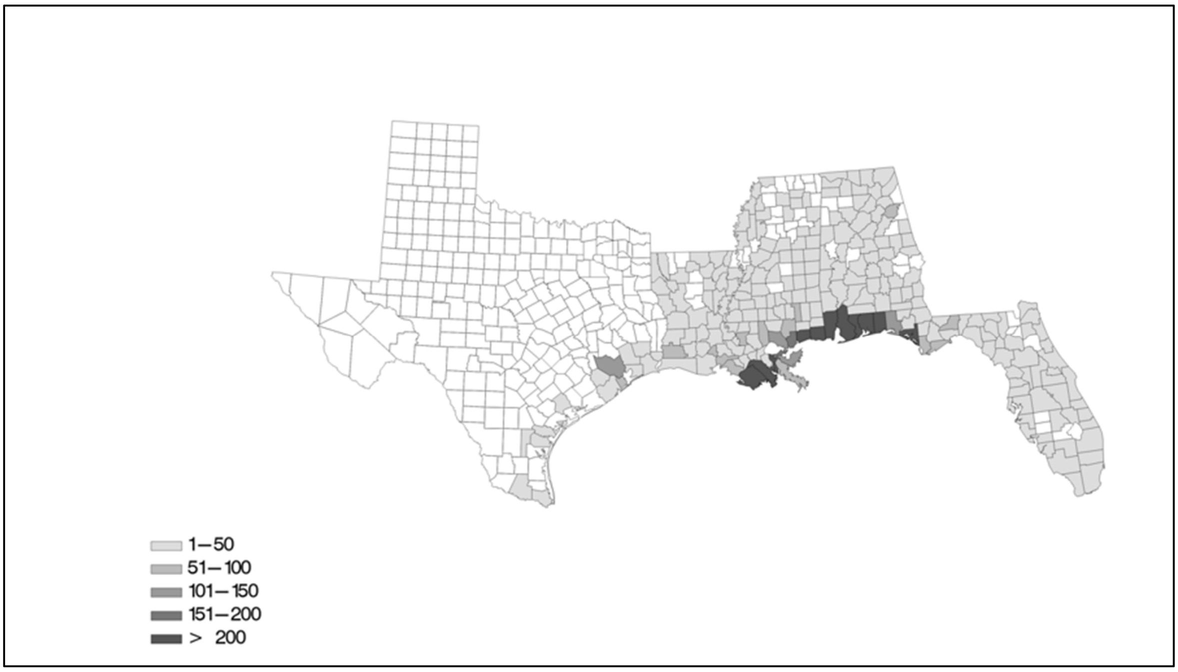

Figure 1. Number of GuLF Study participants eligible for spirometry testing at a home visit by county (N=11,193).

Table 1. Distribution of spirometry quality score (N=10,019) for both FEV_1 and FVC

Spirometry quality score	Frequency (N)	Percent (%)
A (Pass): ≤50 mL variability in FVC or FEV_1 and 3 acceptable tests	2881	29
B (Pass): >50 & ≤100 mL variability in FVC or FEV_1 and 3 acceptable tests	3106	31
C(Pass)[a]: >100 & ≤100 mL variability in FVC or FEV_1 and 3 acceptable tests	1501	15
D (Fail):>150 & ≤200 mL variability in FVC or FEV_1 or <3 acceptable tests	940	9
F (Fail):>200mL or <2 acceptable tests	1591	16

[a]Some participants were included in the passing group if over-reader determined that tests were representative of meeting quality C definition

Table 2a. Raw spirometry values among passing spirometry quality tests (N=7,488[a])

	Mean	Median	25th percentile	75th percentile	SD
FEV_1(mL)	3,145	3,110	2,574	3,698	797
FVC(mL)	4,023	4,000	3,325	4,656	962
FEV_1/FVC (%)	78.23	79.45	74.94	83	7.07
FEV_1 percent predicted[b,c]	88.57	89.43	79.15	98.96	16.23
FVC percent predicted[b,c]	91.52	91.84	82.20	101.08	15.50
FEV_1/FVC percent predicted[b,c]	96.51	97.96	92.46	102.32	8.67

SD=standard deviation

[a]N=1 participant excluded from distributions due to implausible spirometry values

[b]Calculated using Quanjer et al. 2021 GLI equations

[c]N=40 missing data for percent predicted equations

Table 2b. Raw spirometry values among failing quality tests (N=2,531)

	Mean	Median	25th percentile	75th percentile	SD
FEV_1(mL)	3357	3301	2710	3910	1375
FVC(mL)	4335	4312	3590	5009	1106
FEV_1/FVC (%)	78.02	79.72	72.34	84.67	21.24
FEV_1 percent predicted[a,b]	88.03	89.23	77.60	99.99	18.64
FVC percent predicted[a,b]	93.85	92.90	82.27	104.04	20.02
FEV_1/FVC percent predicted[a,b]	94.08	97.22	88.06	103.09	14.05

SD=standard deviation

[a]Calculated using Quanjer et al. 2021 GLI equations

[b]N=10 missing data for percent predicted equations

Table 3. Characteristics of GuLF STUDY participants who took a spirometry exam at a home visit by spirometry quality (N=8,466)[a,b]

	Passing[b] spirometry quality score N(%)	Failed spirometry quality score N(%)
Age		
<30 years	1161(18)	531(25)
30-60 years	4532(72)	1423(66)
>60 years	629(10)	190(9)
Gender		
Male	4834(76)	1782(83)
Female	1488(24)	362(17)
Self-reported Race		
White	3620(57)	1038(48)
African American/Black	2041(32)	911(42)
Other	661(10)	195(9)
Hispanic ethnicity		
Hispanic	393(6)	105(5)
Non-Hispanic	5929(94)	2039(95)
Annual household income		
Less than $20,000/year	2368(38)	987(46)
$20,001 to $50,000/year	2107(33)	702(33)
Greater than $50,000/year	1847(29)	455(21)
Highest educational attainment		
Less than high school/equivalent	1189(19)	505(24)
High school diploma/GED	2107(33)	769(36)
Some college/2-year degree	1991(31)	592(28)
4-year college graduate or more	1035(16)	278(13)
Employment status		
Employed	3537(55)	1087(51)
Unemployed	1627(26)	671(31)
Other	1158(19)	386(18)
Pre-spill respiratory disease diagnosis		
Yes	910(14)	275(13)
Smoking status		
Heavy Current Smoker (≥1 pack per day)	784(13)	256(12)
Light Current Smoker (< 1 pack per day)	1544(24)	512(24)
Former Smoker	1330(21)	409(19)
Never Smoker	2664(42)	967(45)
Body mass index, kg/m^2		
Underweight (< 18.5)	87(1)	45(2)
Healthy (18.5 to 24.9)	1358(21)	584(27)
Overweight (25 to 29.9)	2155(34)	647(30)

Obese (≥30)	2722(43)	868(41)
Always/usually worried about future health		
Yes	2807(44)	1085(51)
Always/usually worried about paying rent or buying food		
Yes	3028(48)	1100 (51)
Previous oil industry experience		
Yes	1813(29)	699(33)
Previous oil spill cleanup experience		
Yes	1078(17)	348(16)
Wheeze at time of enrollment		
Yes	775(12)	320(15)
Tightness in chest		
Yes	271(10)	500(8)
Shortness of breath		
Yes	682(11)	294(14)
Home visit examiner and participant gender match		
Yes	1719(27)	474(22)
No	4603(73)	1670(78)
Location of home visit		
Alabama	1814(29)	570(27)
Florida	1780(28)	577(27)
Louisiana	1456(23)	492(23)
Mississippi	1100(17)	431(20)
Texas	172(3)	74(3)
Gender of home visit examiner		
Male	429(7)	166(8)
Female	5893(93)	1978(92)
Deepwater Horizon oil spill cleanup hierarchical job class		
Non-worker	1200(19)	431(20)
Response worker	967(15)	325(15)
Operations worker	1052(17)	375(18)
Decontamination worker	755(12)	232(11)
Cleanup on water worker	1011(16)	358(17)
Cleanup on land worker	825(13)	278(13)
Support worker	512(8)	145(7)

[a] Passing spirometry quality defined as meeting 2005 ATS criteria of 3 acceptable curves with 150 mL reproducibility for both FEV_1 and FVC
[b]All characteristics are self-reported except for BMI

Table 4. Stepwise predictors of spirometry test failure and odds of failure (N=8,466)

Predictor	Odds Ratio(95% Confidence Interval)[a]	Odds Ratio(95% Confidence Interval)[b]
Self-reported Race		
White	Ref	Ref
African American/Black	1.56(1.40,1.73)	1.39(1.23,1.56)
Other	1.03(0.87,1.22)	0.98(0.81,1.17)
Gender		
Male	1.51(1.33,1.72)	1.61(1.41,1.83)
Female	Ref	Ref
BMI, kg/m^2		
<18.5 (underweight)	0.83(0.57,1.21)	0.81(0.56,1.19)
18.5-24.9 (normal)	Ref	Ref
25-29.9 (overweight)	0.58(0.40,0.84)	0.58(0.40,0.85)
≥30 (obese)	0.62(0.43,0.89)	0.61(0.42,0.89)
Annual household income		
<$20,000	1.69(1.49,1.92)	1.45(1.26,1.67)
$20,000-50,000	1.35(1.18,1.55)	1.25(1.09,1.43)
>$50,000	Ref	Ref

[a]Unadjusted

[b] Mutually adjusted for all predictor variables