Research Article ISSN 2639-944X

Journal of Medical - Clinical Research & Reviews

Creating Risk-Adjusted Outcome Measures Using Baseline Clinical Data across Multiple Physical Therapy Industry Stakeholders

Adam D Lutz, PhD, PT, DPT, MBA[1*], John M Brooks, PhD[2,3], Cole G Chapman, PhD[2,4] and Natalie L Myers, PhD, ATC[5]

[1]*Behavior Frontiers, LLC, 100 North Pacific Coast Highway, Suite 1400, El Segundo, CA 90245. USA.*

[2]*SC SmartState Center of Effectiveness Research in Orthopaedics, 200 Patewood Dr Suite B350, Greenville, SC 29615, USA.*

[3]*Department of Health Services Policy and Management, University of South Carolina; Arnold School of Public, 921 Assembly Street, Columbia, SC 29208, USA*

[4]*College of Pharmacy, Division of Health Services Research, University of Iowa, 345 180 South Grand Ave, Iowa City, IA 52242, USA.*

[5]*ATI Physical Therapy; 2001Butterfield Road, Suite 1600 Downers Grove, IL, 60515, USA and Institute for Musculoskeletal Advancement, 2001 Butterfield Road, Downers Grove, IL, 60515, USA.*

***Correspondence:**

Adam D. Lutz, PhD, PT, DPT, MBA, Behavior Frontiers, LLC, 100 North Pacific Coast Highway, Suite 1400, El Segundo, CA 90245. USA.

Received: 02 Oct 2026; **Accepted:** 05 Nov 2026; **Published:** 18 Nov 2026

Citation: Adam D Lutz, John M Brooks, Cole G Chapman, et al. Creating Risk-Adjusted Outcome Measures Using Baseline Clinical Data across Multiple Physical Therapy Industry Stakeholders. J Med - Clin Res & Rev. 2026; 10(11): 1-12.

ABSTRACT

Background: *Value-based reimbursement models require fair assessment of provider performance through risk adjustment that accounts for patient factors outside the provider's control. The purpose of this study was to evaluate the feasibility of aggregating data and risk-adjusting PRO change across organizations and to assess physical therapist performance for back, neck, and shoulder conditions.*

Methods: *Stakeholders identified baseline patient factors for risk adjustment and provided 2017–2018 data for the qDASH, MDQ, and NDI. Separate models were developed for each PRO using complete episodes, evaluated by explained variance in PRO change, and used to generate therapist-level risk-adjusted performance estimates.*

Results: *Despite consensus on baseline factors and PROs, substantial missing data required simplified risk-adjustment models. Of >4 million potential episodes, 275,459 were included in the final models, which explained 18%–37% of PRO change. Payer type was the strongest consistent predictor, while baseline PRO had the greatest potential influence on model estimates.*

Conclusions: *This study shows that risk-adjusted performance measurement using commonly used PROs is practical with information collected in existing EMRs. Risk-adjustment variables outside of provider control explain a substantial portion of variation in episode-level outcome change, and risk-adjustment is necessary for meaningful assessment of performance.*

Keywords

Performance Measurement; Value-Based Healthcare; Quality; Modified Low Back Pain Disability Questionnaire; Neck Disability Index; Quick Disabilities of the Arm, Shoulder, and Hand

Introduction

The passing of the Affordable Care Act (ACA) marked an important shift in reimbursement strategy from the traditional fee-for-service model to a value-based approach in musculoskeletal (MSK) healthcare [1]. This shift to a value-based approach to

MSK care values a patient outcome framework for treatment decisions that promotes consistent evaluation from the clinician in the probability of improved or risk of poor outcomes [2]. Value-based physical therapy models positively affect overall healthcare costs for the consumer with the implementation of early physical therapy (PT), direct access, opioid reduction, and referral strategies [3-6]. Consequently, the need to evaluate patient success and performance is a clinical necessity and can be done using patient reported outcome measures (PROMs).

Patient outcomes are dynamic and complex and can be affected by patient characteristics. As such, patient characteristics should be considered prior to drawing conclusions about the clinical efficacy of a treatment or the quality of care a patient receives [7]. One method to help account for these patient specific factors is the incorporation of risk-adjusted changes in patient reported outcomes (PROs) [8-14]. There has been growing acceptance of the implementation of risk-adjusted PRO measurements in clinical practice and has been used in PT for many years [7,8,15]. Many important baseline patient factors to consider have been outlined in the available literature [8-13,16-18]. Accounting for this important baseline information is required to establish a fair comparison of clinical performance across physical therapy providers. However, the implementation of risk-adjusted PROs can pose real-world clinical practice challenges. For example, baseline patient characteristics may differ across the diversity of electronic medical record (EMR) systems used within the medical community. This limits the feasibility of standardized quality assessment across practices and undermines broader value-based reimbursement efforts. While it is important to acknowledge this limitation, it is not unobtainable to overcome, and the ability to aggregate and blend clinical and demographic data across EMRs to support meaningful risk adjusted comparisons in PT should be investigated.

Determining the PROMs that should be risk adjusted can be an arduous task. First, in a multicenter collaboration it's imperative to find similarities in reporting structures, and second to recognize which body regions represent a high-volume of caseloads across the system of stakeholders. Not only do the neck, back, and spine represent high-volumes of MSK impairments [19-21], but baseline factors and risk-adjusted algorithms have been explored on commonly utilized legacy PROMs such as the quick Disability of Arm Shoulder Hand (qDASH), Modified Low Back Disability Questionnaire (MDQ), and Neck Disability Index (NDI) [13,22]. While the healthcare system is certainly adopting alterative systems like Patient-Reported Outcomes Measurement Information System (PROMIS) and Focus on Therapeutic Outcomes (FOTO), these legacy outcome measures remain prevalent within clinical workflow and registries.

Physical therapy practices that operate outside of platforms like PROMIS and FOTO would benefit from scalable, risk-adjusted outcome models that leverage existing EMR data that commonly use legacy PROs to support clinical practice. Therefore, the purposes of this study were to aggregate information from stakeholder EMRs, establish a common analytical database for analysis, develop risk-adjusted performance measure algorithms for the shoulder, back, and neck using this analytical database, and lastly to demonstrate variations in physical therapists' performance after risk adjustment. It was hypothesized that are harmonizing and pooling stakeholder EMR data into a common analytical database and applying risk-adjustment algorithms, significant variation in physical therapists' performance will remain across therapists for shoulder, neck, and back care.

Materials and Methods

This retrospective study followed the Reporting of Studies Conducted Using Observational Routinely Collected Data initiative and was reviewed and approved by the University of South Carolina Institutional Review Board (Pro00070869) [23]. Industry stakeholders within the Alliance for Physical Therapy Quality and Innovation (APTQI) performed a literature review (Table 1) and produced a conceptual model containing patient-specific baseline factors thought to influence PRO change during an episode of care commonly available across the contributing body regions (Table 2, Specification 1). Industry stakeholders were made up of a large, national physical therapy practice; a medium-sized, regional physical therapy practice; a large regional health system; and a large national electronic medical record (EMR) provider. These groups were chosen for their coverage of the United States and diversity of covered lives. In addition, the APTQI-led team agreed that the quick Disability of Arm Shoulder Hand (qDASH), Modified Low Back Disability Questionnaire (MDQ), and Neck Disability Index (NDI) were commonly utilized PROs within their EMRs for the management of patients with shoulder, back, and neck complaints, respectively. To allow comparison across instruments, all PRO measures (PROMs) were normalized to a 0-100 scale, with higher scores reflecting better function and lower disability (0=worst outcome, 100 = best outcome). The transformations were as follows:

1) MDQ and NDI: These PROMs are scored 0-50, with higher scores indicating greater disability. To align with the 0-100 (best=100) framework, we applied the below transformation

a. Normalized Score = (100 – (2 X Raw Score). For example, an NDI score of 26/50 becomes 100 – (2*26) = 48

1) qDASH: This PROM ranges from 0-100, with higher scores indicating greater disability. To reverse the direction so that higher scores represent better function we applied the below transformation

a. Normalized Score = 100 - Raw Score. For example, a qDASH score of 65/100 becomes 100 – 65 = 35

While the knee represents a large proportion of MSK related injuries, the outcomes used to assess knee functional improvement were too diverse amongst the stakeholders associated with the project. Diversity of measures across the stakeholders included the Lower Extremity Functional Scale, Knee Injury and Osteoarthritis Outcome Score, and the International Knee Documentation System. The four stakeholders within the APTQI agreed to supply data from their respective EMRs to project investigators containing these PROs and baseline patient factors for episodes of

Table 1: Literature Review of Previous Risk-Adjustment Factors.

	Literature Review					
	Deutscher et al. [14]	Gozalo et al. [8]	Lutz et al. [13]	Resnik et al. [10]	Resnik et al. [11]	Yen et al. [12]
Patient Population	Low Back	Lumbar Shoulder Knee Neck Foot/Ankle Hip Wrist/Hand Elbow Ribs Craniofacial	Low Back Neck	Perspective article	Low Back	Low Back
Factors Considered						
Payer Type	X	X	X	X	X	X
Anthropometry (Height, Weight, &/or BMI)			X			
Diagnosis				X	X	
Days Since Onset/ Chronicity	X	X	X	X	X	X
Physical Therapy Following Related Surgery	X	X	X	X	X	X
Number of Surgeries	X	X		X	X	
Initial Outcome Score	X	X	X	X	X	
Initial General Health Survey			X			
Initial Functional Comorbidity Survey		X				
Initial Fear Avoidance		X				
Comorbid History	X		X			X
Medication Use for Condition	X					
Previous Treatment for Condition	X					
Sex	X	X	X	X	X	
Age	X	X	X	X	X	X
Exercise History	X			X	X	
Socioeconomic Status			X			
Employment History				X	X	
Years of Experience as a Physical Therapist					X	
PTA Involved in Care					X	
Provider Count & Mix in Clinic					X	
Clinic Productivity				X	X	
Referral Source – Type				X	X	

PTA= Physical therapy assistant.

care related to the back, neck, and shoulder for patients treated in 2017 and 2018 as outlined in Table 2. Using these data, the project investigators used the following steps to address the aims of this paper.

Step 1: Aggregate Information from Stakeholders EMRs and Establish a Common Analytical Database Across Stakeholders

A request for proposal was issued to obtain all data points related to episodes of care from each industry stakeholder's database. The extracted data were then standardized to a set of common factors (e.g., consistent variable names, coding schemes, and value ranges). After this harmonization process, the four databases were merged into a single integrated dataset.

Table 2 lists the baseline factors identified for consideration in the request for proposal following a literature review by the stakeholders (Table 1). Of note, ZIP code-level median income data were obtained from the US Census Bureau by CERortho and assigned to each patient based on reported residence ZIP code [24].

Step 2: Develop Risk-Adjustment Performance Measure Algorithms

A patient's episode was eligible for inclusion in a risk-adjustment model if the episode contained an initial and final PRO and had complete data for the baseline demographic and clinical variables specified as model predictors. The initial PRO was defined as the first patient-reported outcome collected at the initial physical therapy visit, while the final PRO was defined as the last outcome measure obtained. It was apparent after Step 1 that most episodes in the common database had missing information for baseline factors in the (APTQI-led) conceptual model (Table 2, Specification 1). Alternative risk-adjustment models (Table 2, Specifications 2 and 3) including fewer baseline factors were developed with the goal of increasing the number of included patient episodes.

Table 2: Model Factors Included Across Risk-Adjustment Model Specifications.

Factors Considered	Definition	Included in Risk-Adjustment Model Specification		
		1	2	3
Patient-Reported Outcome Change	(patient-reported outcome score at episode discharge) – (patient-reported outcome score at episode initial visit). All outcome scores scaled from 0 (worst) to 100 (best).	X	X	X
Payer Indicator Variables				
Auto	1 if patient insured through Auto insurance, 0 otherwise.	X	X	X
Commercial	1 if patient insured through Commercial insurance, 0 otherwise.	X	X	X
Medicaid	1 if patient insured through Medicaid insurance, 0 otherwise.	X	X	X
Medicare	1 if patient insured through Medicaid insurance, 0 otherwise.	X	X	X
Other	1 if patient insured through Other insurance, 0 otherwise.	X	X	X
Worker's Comp	1 if patient insured through Worker's Compensation, 0 otherwise.	X	X	X
Initial Outcome Score	Patient initial outcome score (0-100)	X	X	X
Baseline Health Status				
Height: Centimeters	Patient height in centimeters	X		
Weight: Pounds	Patient weight in pounds	X		
BMI	Patient BMI score		X	
Acute	1 if patient injury is acute, 0 otherwise.	X	X	X
Chronic	1 if patient injury is chronic, 0 otherwise.	X	X	X
First PT Visit Before Surgery	1 if patient index visit occurred before surgery, 0 otherwise.	X	X	X
First PT Visit After Surgery	1 if patient index visit occurred after surgery, 0 otherwise.	X	X	X
Index Pain (0 - 10)	Patient index pain score	X		
Veterans RAND 12-Item Health Survey Mental Component Score	Patient index mental health score	X		
Veterans RAND 12-Item Health Survey Physical Component Score	Patient index physical health score	X		
Socio-Demographics				
Female	1 if patient is female, 0 otherwise.	X	X	X
Male	1 if patient is male, 0 otherwise.	X	X	X
Age	Age of patient at index visit	X	X	X
Race		X		
Asian/Pacific	1 if patient is Asian/Pacific origin, 0 otherwise.	X		
African American	1 if patient is African American, 0 otherwise.	X		
Caucasian	1 if patient is Caucasian, 0 otherwise.	X		
Eskimo/American Indian	1 if patient is Eskimo/American Indian, 0 otherwise.	X		
Hispanic	1 if patient is Hispanic, 0 otherwise.	X		
Patient Zip Code Median Income	Median income in patient residence Zip Code in 10K (e.g. 55,000 specified as 5.5)	X	X	X
Comorbidities at Index				
Health History: Arthritis	1 if patient has a history of arthritis, 0 otherwise.	X	X	
Health History: Blood Clot or DVT	1 if patient has a history of blood clots, 0 otherwise.	X	X	
Health History: Breathing Difficulties/Asthma	1 if patient has a history of breathing difficulties, 0 otherwise.	X	X	
Health History: Cancer	1 if patient has a history of cancer, 0 otherwise.	X	X	X
Health History: Diabetes Mellitus	1 if patient has a history of diabetes, 0 otherwise.	X	X	X
Health History: Kidney Condition	1 if patient has a history of kidney conditions, 0 otherwise.	X	X	
Health History: Osteoporosis	1 if patient has a history of osteoporosis, 0 otherwise.	X	X	
Health History: Pacemaker	1 if patient has a pacemaker, 0 otherwise.	X	X	
Health History: Psychological Condition	1 if patient has a history of psychological conditions, 0 otherwise.	X	X	
Health History: Seizures	1 if patient has a history of seizures, 0 otherwise.	X	X	
Health History: Smoking	1 if patient has a history of smoking, 0 otherwise.	X	X	X
Health History: CVA	1 if patient had a previous stroke, 0 otherwise.	X	X	
Health History: Unexplained Weight Loss	1 if patient had a previous unexplained weight loss, 0 otherwise.	X	X	X

DVT= Deep Vein Thrombosis; CVA= Cerebrovascular Accident; Comorbidities coded as 0(no) or 1 (yes) for each category.

Table 3: Descriptive Information in the Functional Model Specification 2 for the Shoulder, Neck, and Back. Data are presented as median (interquartile range) or % (count).

Factors	Shoulder	Neck	Back
Episode count	75,294	61,047	139,118
Visit count	13 (7,20)	12 (8,17)	12 (8,17)
PRO change	16 (2,31)	10 (0,20)	8 (-1,17)
Outcome score at initial	55 (36,73)	64 (52, 78)	64 (51,77)
Payer Variables (ref = auto)			
Commercial	57% (42,661)	48% (29,241)	50% (68,972)
Medicaid	5% (3,762)	8% (5,026)	10% (13,352)
Medicare	22% (16,603)	18% (11,160)	26% (36,396)
Other	2% (1,580)	2% (2,187)	2% (2,187)
Workers Comp	11% (8,380)	6% (3,869)	7% (10,106)
Auto	3% (2,308)	18% (10,859)	6% (8,078)
Health Status at Index			
BMI	27 (24, 32)	27 (24,31)	28 (25,33)
Chronic (ref=acute)	86% (64,668)	91% (55,449)	90% (125,129)
1st PT visit after surgery (ref=before)	28% (21,163)	6% (3,553)	7% (10,214)
Socio-demographics			
Male (ref = female)	48% (36,090)	36% (21,895)	42% (58,856)
Age	56 (44,66)	53 (40,64)	56 (40,68)
Patient zip code median income ($10,000)	6.44 (5.17, 8.06)	6.31 (5.06,7.91)	6.22 (4.97,7.79)
Comorbidities at Index			
Health History: Arthritis	37% (28,215)	38% (23,187)	42% (58,235)
Health History: Blood Clor or DVT	4% (2,972)	4% (2,262)	4% (6,206)
Health History: Breathing Difficulties/Asthma	16% (12,412)	19% (11,409)	19% (26,747)
Health History: Cancer	9% (6,568)	8% (4,811)	9% (12,551)
Health History: Diabetes Mellitus	15% (11,231)	12% (7,517)	15% (20,814)
Health History: Kidney Condition	7% (4,900)	6% (3,787)	8% (10,515)
Health History: Osteoporosis	9% (6,566)	9% (5,537)	11% (15,375)
Health History: Pacemaker	1% (751)	1% (660)	1% (2,010)
Health History: Psychological Condition	12% (9,404)	15% (9,009)	14% (19,784)
Health History: Seizures	2% (1,233)	2% (1,260)	2% (2,704)
Health History: Smoking	13% (9,869)	15% (9,020)	16% (21,860)
Health History: CVA	2% (1,720)	2% (1,489)	3% (4,003)
Health History: Unexplained Weight Loss	2% (1,115)	2% (1,427)	2% (3,091)

DVT= Deep Vein Thrombosis; CVA=Cerebrovascular Accident; Comorbidities coded as 0(no) or 1 (yes) for each category

Statistical Approach

The three model specifications were estimated for each body region using nine multivariable linear regression to contrast their abilities to control for baseline patient factors. PRO change over an episode of care was the dependent variable in the risk-adjustment regression models. Models were estimated using robust standard errors. Use of robust standard errors coupled with our large sample sizes yield regression parameter estimates which are normally distributed regardless of the distribution of the error terms within each model [25]. The R^2 statistics was reported as part of the regression model to estimate the percentage of variation in PRO change attributable to the set of baseline factors. The F-statistics for the regressions were used to indicate whether the set of baseline factors specified in each model describe a significant portion of the variation in outcome change. A priori significance (α) was set at 0.05 and model estimation produced unstandardized coefficients. The authors acknowledge that other statistical approaches could been implemented; however, we are confident that the approach selected yields scientifically sound, reproducible, and interpretable results. Although data were contributed by four healthcare systems, all variables were harmonized and merged into a single integrated data set, with the patient as the unit of analysis. While outpatient sample size was large, the limited number of higher-level clusters (stakeholders n=4) was insufficient to support stable estimation of random effects in multilevel hierarchical models. For this reason, we used multivariable linear regressions. This approach allowed us to evaluate patient-level predictors in a unified framework while maintaining parsimony and interpretability. Tables providing summary statistics, patient counts, and estimated parameters by body region by model specification were developed and provided to all APTQI stakeholders to inform the team of potential tradeoffs associated with each model specification. The APTQI stakeholders compared the number of patient episodes (Figure 1) with complete data available for each model specification across all body regions.

Step 3: Demonstrate Variation in Physical Therapist Performance

After Risk-Adjustment

For risk-adjustment model specification selected by the APTQI stakeholders in Step 2, we calculated and assessed the distribution of risk-adjusted provider performance measures generated by these models. For each body region, the actual change in PROs across the care episode was calculated for each patient by subtracting the initial PRO from the final PRO. For example, a patient with an initial PRO of 40 and a final PRO of 70 would have an actual change of +30 points. Using our risk-adjustment model, which accounts for baseline patient characteristics such as age, sex, comorbidities, and baseline function, we predicted the expected risk-adjusted PRO change for each patient. For instance, based on the baseline profile, the model might predict a PRO change of +25 points. The risk-adjusted PRO change was then calculated as the difference between the actual PRO change and predicted change which allows for the determination of a residual PRO change (equation below).

Residual PRO change = Actual PRO change – risk-adjusted PRO predicted change
In the example above: 30 (actual) − 25 (predicted) = +5, indicating the patient improved 5 points more than expected given their baseline characteristics. A negative residual would indicate the patient improved less than expected.

To evaluate performance at the therapist level, these residuals were aggregated by averaging across all patients treated by the same physical therapist. For example, if a therapist treated three patients with residuals of +5, −2, and +1, the mean residual for that therapist would be (+5 − 2 + 1)/3 = +1.33. Positive mean residuals indicate that patients on average improved more than expected, while negative values indicate less-than-expected improvement.

Results

Patient Episodes Available for Analyses

Figure 1 provides the number of unique patient episodes received from the four data contributors and the number of episodes available for use in risk-adjustment modeling by body region for the three model specifications that varied with the number of specified baseline factors. CERortho received data that represented 4,076,745 unique potential patient episodes. Of these total episodes, 2,207,457 were clearly identified as related to the shoulder, neck, or back; 495,837 of whom included at least 1 PRO, and 330,828 episodes ultimately included both initial and final PROs that matched the body region of interest and were consistent with the request for proposal. Among patient episodes with index and discharge patient reported outcomes, 5,411 combined patient episodes (i.e., back, neck, and shoulder models combined) were available for Specification 1; 275,459 combined patient episodes were available for Specification 2; and 317,094 combined patient episodes were available for Specification 3.

Rationale of Risk-Adjustment Estimation Model Specifications

After reviewing the tradeoffs between sample sizes and model specifications, the APTQI stakeholders chose Specification 2 as the basis for risk-adjustment. Relative to Specification 1, Specification 2 excluded race, pain, and the Veterans RAND 12-Item Health Survey (VR-12) (mental and physical component scores) [26]. These variables were the highest percentage missing across the risk-adjustment factors and were not reliably stored by all industry stakeholders.

Relative to Specification 3, Specification 2 includes BMI and several additional baseline comorbidities. Baseline BMI had a statistically significant effect on outcome change scores for 2 of the 3 body regions. In addition, many of the added comorbidities also had significant effects on the outcome of change scores. As a result, industry stakeholders and the CERortho team agreed that Specification 2 provided a strong foundation to assess the practicality of risk-adjustment modeling given the information collected by physical therapy practices. Thus, the remainder of the text focuses on the specification 2 model.

Descriptive data on the baseline factors for the 3 body regions

Tables 3 contains summary statistics for the patient episodes available for risk-adjustment by model specification 2 for shoulder, neck, and back, respectively. The patients represented across the entire sample were predominantly female (58%), with conditions that were considered chronic (>= 90 days, 88%). Most patients in the sample used commercial insurance (51%); followed by Medicare (23%); Medicaid, Worker's Compensation, Auto/Personal Injury; and "Other" (e.g., Uninsured/Self-Pay). The PRO scores for each body region were normalized to a 0 (worst) to 100 (best) scale. Median normalized baseline PRO scores were 55 (interquartile range [IQR] 36,73) for qDASH, 64 (IQR 52,78) for NDI, and 64 (IQR 51,77) for MDQ. Commonly observed comorbidities were Arthritis (40%), Breathing Difficulties (19%), Psychological Conditions (15%), Smoking (15%), and Diabetes (14%).

Patient reported outcome (PRO) change over an episode of care was the dependent variable in the risk-adjustment regression models. Median change in PRO was 16 (IQR 2,31) for qDASH, 10 (IQR 0,20) for NDI, and 8 (IQR -1,17) for MDQ. Visits were not considered as an outcome or factor in the risk-adjustment; however, for context the median visits per episode were 13 (IQR 7,20) for qDASH, 12 (IQR 8,17) for NDI, and 12 (IQR 8,17) for MDQ.

Risk-Adjustment Model Estimates

Table 4 contains the risk-adjustment model estimates for shoulder, neck, and back, respectively. The observations in these analyses are single episodes of care as described in Specification 2 in Figure 1. The R^2 statistics presented in these tables estimate the percentage of variation in PRO change attributable to the set of baseline factors specified in each model. In Specification 2, the specified baseline factors explained 37% of the PRO variation in qDASH change, 18% in NDI change, and 20% in MDQ change. In the 3 regression models, the F-statistics were large and statistically significant, indicating the importance of these baseline factors on PRO change over the episode of care. Since some of these factors

are outside the control of the provider, the results suggest that risk-adjustment is necessary to yield performance measures that can provide meaningful comparisons across providers. Partial F-tests estimating the relative importance of grouped variables are presented in Table 5. Across all body regions, initial PRO change was the strongest predictor of outcome change, surpassing the contributions of payer, health status, socio-demographic, and comorbidity variables. Payer and health status factors contributed additional explanatory power, particularly in shoulder and back cohorts, while socio-demographics and comorbidities had smaller but statistically significant effects. Notably, the predictive importance of baseline health status was minimal in neck patients compared to shoulder and back.

Performance Measure Distributions

Figures 2a-c establish the distribution of performance measures across physical therapists for conditions related to the shoulder, neck, and back, respectively. Each figure contains performance measures for only the physical therapists in the database with 40 or more complete patient episodes (including initial and final outcome with all requisite predictor variables) in the respective body region. The Y-axis represents the residual PRO change, and each observation is the average value for a single physical therapist. Physical therapists are distributed along the X-axis from lowest to highest performance measures. Physical therapists with a performance measure greater than zero on the Y-axis performed better than average given the characteristics of their patients. The opposite is true of the physical therapists with a performance measure below zero on the Y-axis.

Discussion

Calculation of physical therapist risk-adjusted performance

Table 4: Regression Coefficients (B) with 95% confidence intervals (CI) for shoulder, neck, and back patients in the functional model specification 2.

Factors	Shoulder	Neck	Back
Episode count	75,294	61,047	139,118
Outcome score at initial	-0.55 (-0.56,-0.54)***	-0.34 (-0.35,-0.33)***	-0.36 (-0.36,-0.35)***
Payer Variables (ref = auto)			
Commercial	4.48 (3.69,5.28)***	-1.22 (-1.55,-0.89)***	1.37 (1.02,1.72)***
Medicaid	-2.27 (-3.26,-1.28)***	-5.95 (-6.46,-5.44)***	-4.33 (-4.75,-3.91)***
Medicare	4.92 (4.07,5.76)***	0.15 (-0.28,0.58)	1.29 (0.91,1.68)***
Other	2.73 (1.59,3.86)***	-2.78 -3.70,-6.14)***	-0.70 (-1.38,-0.01)*
Workers Comp	-5.38 (-6.27,-4.50)***	-6.69 (-7.25,-6.14)***	-4.06 (-4.51,-3.61)***
Health Status at Index			
BMI	-0.05 (-0.08,-0.03)***	0.01 (-0.01,0.02)	-0.11 (-0.12,-0.10)***
Chronic (ref=acute)	-1.88 (-2.26.-1.50)***	-0.47 (-0.86,-0.08)*	-0.45 (-0.70,0.20)***
1st PT visit after surgery (ref=before)	6.25 (5.92,6.57)***	-1.14 -1.61,-0.66)***	-0.13 (-0.42,0.15)
Socio-demographics			
Male (ref = female)	1.77 (1.52,2.03)***	0.13 (-0.10.0.37)	0.75 (0.59,0.90)***
Age	-0.09 (-0.10,-0.08)***	-0.05 (-0.06,-0.04)***	-0.05 (-0.06,-0.05)***
Patient zip code median income	0.26 (0.21,0.31)***	0.08 (0.03,0.13)**	0.22 (0.19,0.25)***
Comorbidities at Index			
Health History: Arthritis	-1.17 (-1.45,-0.88)***	-1.42 (-1.67,-1.16)***	-2.36 (-2.54,-2.19)***
Health History: Blood Clor or DVT	-1.58 (-2.23,-0.93)***	-1.02 (-1.60,-0.44)***	-1.23 (-1.59,-0.87)***
Health History: Breathing Difficulties/Asthma	-0.91 (-1.26,-0.56)***	-1.17 (-1.46,-0.87)***	-1.43 (-1.63,-1.24)***
Health History: Cancer	0.52 (0.10,0.94)**	0.36 (-0.03,0.74)	0.53 (0.28,0.78)***
Health History: Diabetes Mellitus	-1.31 (-1.69,-0.96)***	0.20 (-0.15,0.56)	-0.71 (-0.93,-0.49)***
Health History: Kidney Condition	-0.75 (-1.27,-0.22)***	-0.66 (-1.13,-0.19)***	-0.75 (-1.03,-0.46)***
Health History: Osteoporosis	-0.61 (-1.06,-0.32)**	-0.48 (-0.87,-0.09)*	-0.69 (-0.94,-0.45)***
Health History: Pacemaker	0.37 (-0.81,1.56)	0.14 (-0.89,1.16)	0.05 (-0.55,0.65)
Health History: Psychological Condition	-3.04 (-3.45,-2.62)***	-3.02 (-3.37,-2.67)***	-2.94 (-3.16,-2.71)***
Health History: Seizures	-2.53 (-3.61,-1.45)***	-2.11 (-2.99,-1.24)***	-1.53 (-2.19,-0.96)***
Health History: Smoking	-1.83 (-2.23,-1.44)***	-1.06 (-1.39,-0.73)***	-1.98 (-2.19,-1.76)***
Health History: CVA	-2.73 (-3.62,-1.84)***	-0.52 (-1.27,-0.22)	-0.68 (-1.14,-0.22)***
Health History: Unexplained Weight Loss	-1.95 (-3.07,-0.83)***	-2.65 (-3.42,-1.89)***	-2.28 (-2.79,-1.76)***
Constant	51.91 (50.76,53.06)***	37.92 (36.99,38.84)***	39.29 (38.61,39.97)***
Summary Statistics			
R^2	0.37	0.18	0.20
Adjusted R^2	0.37	0.18	0.20
Residual Standard Error	16.48 (df=75,265)	13.41 (df=61,020)	13.68 (df=139,088)
F statistic	2,005 (df=25;75,265)***	494.20 (df=25;61,020)***	1,115 (df=25;139.088)***

Note: * p <0.05, ** p <0.01, *** p <0.001; DVT= Deep Vein Thrombosis; CVA= Cerebrovascular Accident; Comorbidities coded as 0(no) or 1 (yes) for each category.

Table 5: Partial F-Statistics Indicating the Relative Strength of Baseline Factor Groups in Predicting Outcome Change by Body Region.

	Body Region		
Baseline Factor Group	Shoulder (N = 75,294)	Neck (N = 61,047)	Back (N = 139,118)
Initial PRO change	31,984.1***	10,309.8***	25,666.4***
Payer Variables[a]	526.8***	260.8***	526.4***
Health Status at Initial[b]	587.3***	9.9***	128.3***
Socio-Demographics[c]	234.1***	51.6***	200.7***
Comorbidities at Initial[d]	69.9***	72.7***	257.6***

a. Commercial, Medicaid, Medicare, Other, Worker's Comp (reference = auto).
b. BMI, Chronic (ref = acute), First PT Visit After Surgery (ref = before).
c. Male (ref = female), Age, Patient Zip Code Median Income.
d. Health History: Arthritis, Health History: Blood Clot or DVT, Health History: Breathing Difficulties / Asthma, Health History: Cancer, Health History: Diabetes Mellitus, Health History: Kidney Condition, Health History: Osteoporosis, Health History: Pacemaker, Health History: Psychological Condition, Health History: Seizures, Health History: Smoking, Health History: CVA, Health History: Unexplained Weight Loss
* p < .05, ** p < .01, *** p < .0001

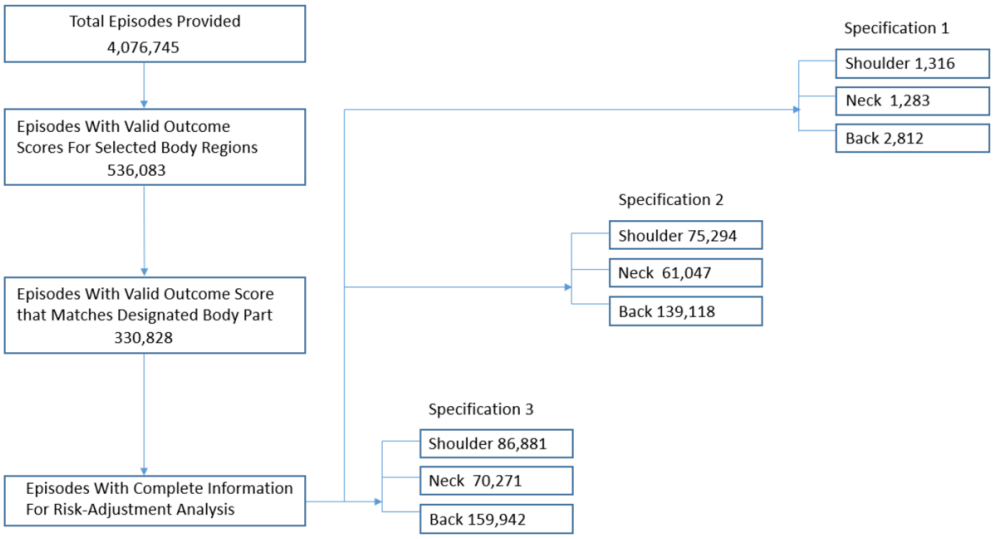
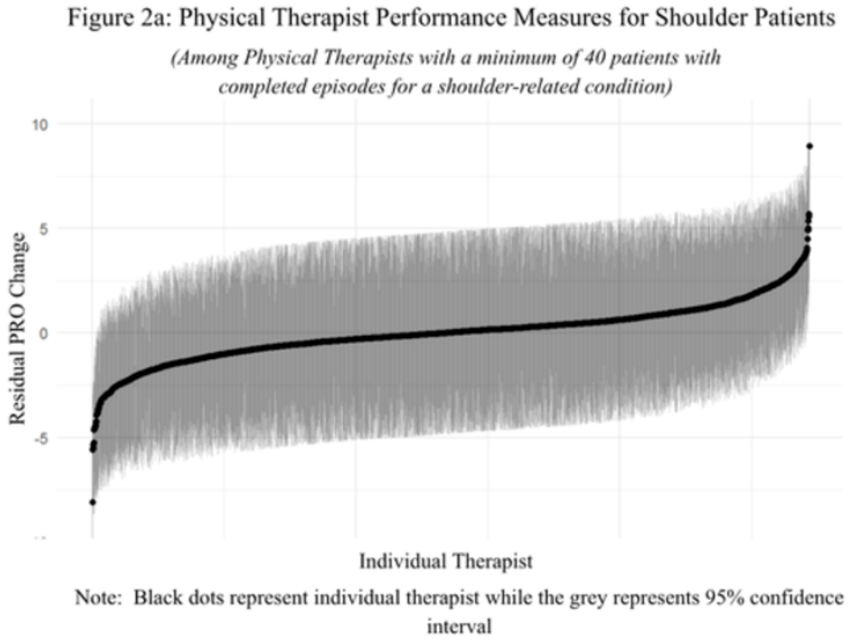
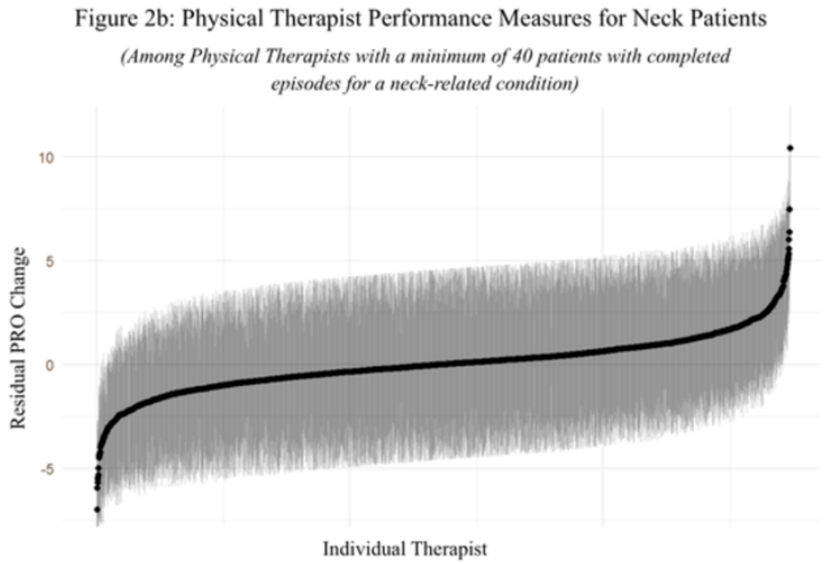
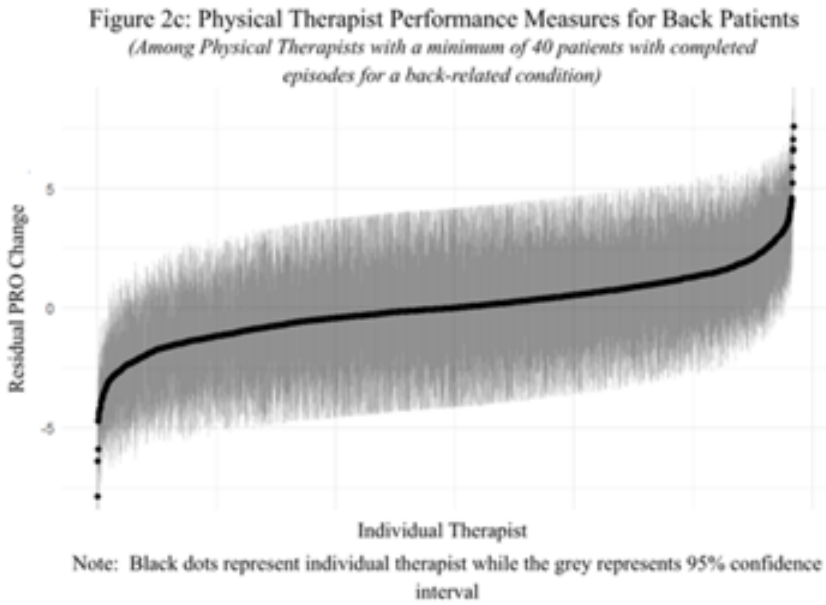


Figure 1: Sample Size Summary for Risk-Adjustment Analysis by Body Region and Model Specification.







Figures 2a-c: Physical Therapist Performance Measures for Shoulder, Neck, and Back, respectively (physical therapists treating a minimum of 40 shoulder patients).

measures using data from disparate EMRs was practical and important to fairly assess clinical performance. Further, these results are generalizable to all outpatient physical therapy providers in the United States due to the diversity of included physical therapists and patients in the sample. The results are also easy to implement with straightforward calculations. The estimated linear regression coefficients for the patient baseline factors on PRO change have direct interpretations. For factors specified as binary variables, the estimated coefficient equals the expected difference in the PRO change score between patients with that status or condition at baseline and those who did not. For example, in the model specification estimating qDASH change (Table 4), patients with commercial insurance were expected to have PRO change that was 4.5 points greater than patients with auto insurance (the reference group) over the care episode. In contrast, patients with Worker's Compensation insurance were expected to have PRO change that was 5.4 points less than patients in the reference group (auto insurance). For continuous factors, the coefficient estimates reflect the difference in expected outcome change for a one unit change in the factor. Again, using the model specification estimating qDASH change (Table 4), a representative 10-unit difference in BMI (e.g., 27 vs. 37) was associated with a 0.5 unit decrease in expected PRO change over the episode of care. Meaning BMI was negatively associated with PRO change and that patients with a higher BMI demonstrated slightly smaller improvements in PRO change compared to those with lower BMI.

The study also highlighted several challenges. First, collection of baseline health history and demographic information differed between organizations. Second, even when comparable factors

were collected across organizations, it was sometimes structured or formatted in a manner that did not allow for aggregation at the episode level. EMRs were developed to document information necessary for reimbursement of patients in the traditional fee-for-service reimbursement model, to facilitate communication, and collect patient information relevant to patient care, among other things, but not to be aggregated in outcomes registries and leveraged for research and performance measurement purposes. Finally, the frequency of PRO collection can be driven by state, payer, or organizational policy; therefore, the availability of PROs beyond the initial physical therapy evaluation varied widely. These differences influenced our ability to aggregate information to risk-adjust at the physical therapist, clinic, and/or organizational levels.

Seminal work using risk-adjusted PROs as performance measures in the physical therapy industry by Resnik et al. [9-11,18], Yen et al. [12] and Gozalo et al. [8] compared data from different organizations; however, the organizations all used a common EMR for data collection, the Focus On Therapeutic Outcomes, Inc. (FOTO) platform. These works were integral to the industry stakeholder team in understanding important risk-adjustment factors to consider. Results from this study generally aligned with results from these prior works, with similar statistical significance and explanatory power of the regression models. While direct comparison of explained variance with many prior studies is challenging due to differing methodologies, Deutscher et al. reported 37% explained variance, and Yen et al. reported 31% explained variance in their models predicting functional change in patients with lower back pain using FOTO functional scales; while Lutz et al. reported 19% explained variances in models predicting change in NDI and MDQ [12-14]. Our investigation yielded comparable explained variance, 18%, 20%, and 37%, for the models predicting change in NDI, MDQ, and qDASH, respectively. It should be noted that these results included a smaller set of baseline patient factors due to the challenges described above when combining multiple data sources. Further, our approach excluded therapist-level information and instead assigned all unexplained variance to the therapist providing care. While we highlighted above that unknown or unmeasured variables surely make up some portion of this unexplained variance, we hypothesize that the variability in explained variance between the 3 outcomes is associated with 2 key tenets—1) statistical properties and specific questions queried in the outcomes, and 2) relative variability in performance relative to outcomes. The NDI and MDQ have less initial disability and significantly less change from initial to final outcome; potentially reducing the ability of the independent variables to explain PRO change. Additionally, there may be greater variability in physical therapy provider performance for conditions related to the spine; therefore, yielding more "unexplained variance". Finally, this study yielded a set of predictor variables that aligned closely with prior investigations. The most important predictors across the 3 outcomes were the initial PRO and payer class; these variables were commonly the most important predictors across common investigations.

Limitations

This study used observational data, and our performance measurement approach adjusts only for baseline patient factors that were available in the integrated dataset. Consequently, any variation in patient outcomes not explained by these measured factors is attributed to provider performance. While this framework is consistent with common approaches in risk-adjusted performance measurement, it is important to acknowledge that unmeasured variables such as patient motivation, adherence, social support, or clinical nuances not captured in the EMR may also influence outcomes. There are likely other unmeasured patient factors that influence patient outcomes. As a result, the measures calculated here will be a biased measure of a physical therapist's performance if a therapist cares for patients with a disproportionate distribution of these unmeasured factors. Another limitation was the low number of patients with complete data. Patient outcomes for patients included in the risk-adjustment models may not be representative of those that were excluded, and these differences are not reflected in the risk-adjustment models estimated here. This investigation was completed retrospectively, and organizations were not able to develop a data capture and storage strategy necessary to comply with the requested variable fields. While this is important for generalizability, it should be noted that the results are in line with prior work by Deutscher et al., Lutz et al., and Yen et al.; and the challenges were associated with data capture versus practicality, the key goal of this investigation [12-14]. Finally, we were unable to control for timeliness of final PRO. Due to the other constraints limiting inclusion, we believe that the pragmatic nature of the study allows such inclusion for the purposes of demonstrating practicality. However, it is important to consider the timeliness of the final outcome when building a risk-adjusting process at the organization level.

Conclusions

This study demonstrated that risk-adjusted performance measurement is practical using the most used and freely available PROs and patient-level information that is collectable from existing EMR systems. Patient baseline factors outside the control of providers explain a substantial portion of the variation in episode-level outcome changes across patients. As a result, risk-adjustment for these factors is necessary for meaningful performance comparisons across providers. Baseline disability level, payer status, socioeconomic characteristics, and comorbidities at care initiation all had significant effects on subsequent outcome changes and should be "risk-adjusted" to ensure that performance measures provide fair comparisons across physical therapists. Due to the wide variation in data collection and storage, standardization of the data collected in this industry is needed for consistent, industry-wide performance measurement. Physical therapy providers must be committed to the collection of baseline patient information and patient-reported outcomes across each episode of care to utilize this approach. Future work in this field should examine prospective data capture to establish the generalizability of the approach used in this limited, retrospective set.

Funding

This project was funded by the Alliance for Physical Therapy Quality and Innovation ($144,218) in August of 2017.

Declaration of Interest

When this manuscript was written AL was an employee of ATI Physical Therapy, one of the industry stakeholders and data contributors, and a research associate with the Center for Effectiveness Research in Orthopedics, the investigating body. The project began when Adam Lutz was a PhD candidate with the University of South Carolina.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work, the author(s) used ChatGPT, a large language model developed by OpenAI, to assist with manuscript preparation, including refining language, improving clarity, organizing study rationale, and drafting/editing selected sections. The tool was also used to assist with interpretation of statistical output. After using this tool, the author(s) reviewed and edited all content as needed and take full responsibility for the content of the published article.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, AL, upon reasonable request.

References

1. Burwell SM. Setting value-based payment goals HHS efforts to improve US health care. N Engl J Med. 2015; 372: 897-899.
2. Lentz TA, Goode AP, Thigpen CA, et al. Value-based care for musculoskeletal pain are physical therapists ready to deliver. Phys ther. 2020; 100: 621-632.
3. Denninger TR, Cook CE, Chapman CG, et al. The Influence of Patient Choice of First Provider on Costs and Outcomes Analysis From a Physical Therapy Patient Registry. J Orthop Sports Phys Ther. 2018; 48: 63-71.
4. Ojha HA, Snyder RS, Davenport TE. Direct access compared with referred physical therapy episodes of care a systematic review. Phys Ther. 2014; 94: 14-30.
5. Childs JD, Fritz JM, Wu SS, et al. Implications of early and guideline adherent physical therapy for low back pain on utilization and costs. BMC Health Serv Res. 2015; 15: 150.
6. Frogner BK, Harwood K, Andrilla CHA, et al. Physical Therapy as the first point of care to treat low back pain An instrumental variables approach to estimate impact on opioid prescription health care utilization and costs. Health Serv Res. 2018; 53: 4629-4646.
7. Iezzoni LI. Using risk-adjusted outcomes to assess clinical practice an overview of issues pertaining to risk adjustment. Ann Thorac Surg. 1994; 58: 1822-1826.
8. Gozalo PL, Resnik LJ, Silver B. Benchmarking outpatient rehabilitation clinics using functional status outcomes. Health Serv Res. 2016; 51: 768-789.
9. Resnik L, Hart DL. Using clinical outcomes to identify expert physical therapists. Phys Ther. 2003; 83: 990-1002.
10. Resnik L, Liu D, Hart DL, et al. Benchmarking physical therapy clinic performance statistical methods to enhance internal validity when using observational data. Phys Ther. 2008; 88: 1078-1087.
11. Resnik L, Liu D, Mor V, et al. Predictors of physical therapy clinic performance in the treatment of patients with low back pain syndromes. Phys Ther. 2008; 88: 989-1004.
12. Yen SC, Corkery MB, Chui KK, et al. Risk adjustment for lumbar dysfunction comparison of linear mixed models with and without inclusion of between-clinic variation as a random effect. Phys Ther. 2015; 95: 1692-1702.
13. Lutz AD, Brooks JM, Chapman CG, et al. Risk Adjustment of the Modified Low Back Pain Disability Questionnaire and Neck Disability Index to Benchmark Physical Therapist Performance Analysis From an Outcomes Registry. Phys Ther. 2020; 100: 609-620.
14. Deutscher D, Werneke MW, Hayes D, et al. Impact of risk adjustment on provider ranking for patients with low back pain receiving physical therapy. J Ortho Sports Phys Ther. 2018; 48: 637-648.
15. Deutscher D, Horn SD, Dickstein R, et al. Associations between treatment processes patient characteristics and outcomes in outpatient physical therapy practice. Arch Phys Med Rehabil. 2009; 90: 1349-1363.
16. Resnik L, Feng Z, Hart DL. State regulation and the delivery of physical therapy services. Health Serv Res. 2006; 41: 1296-1316.
17. Resnik L, Dobrzykowski E. Guide to outcomes measurement for patients with low back pain syndromes. J Orthop Sports Phys Ther. 2003; 33: 307-316.
18. Resnik L, Jensen GM. Using clinical outcomes to explore the theory of expert practice in physical therapy. Phys Ther. 2003; 83: 1090-106.
19. Holder NL, Clark HA, DiBlasio JM, et al. Cause prevalence and response to occupational musculoskeletal injuries reported by physical therapists and physical therapist assistants. Phys ther. 1999; 79: 642-652.
20. Lock C, Allgar V, Jones K, et al. Prevalence of back neck and shoulder problems in the inner city implications for the provision of physiotherapy services in primary healthcare. Physiother Res Int. 1999; 4: 161-169.
21. Hodgetts CJ, Leboeuf-Yde C, Beynon A, et al. Shoulder pain prevalence by age and within occupational groups a systematic review. Arch Physiother. 2021; 11: 24.
22. Uhl TL, Smith-Forbes EV, Nitz AJ. Factors influencing final outcomes in patients with shoulder pain A retrospective review. J Hand Ther. 2017; 30: 200-207.
23. Benchimol E, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected

health Data RECORD statement. PLoS ONE. 2015; 12: e1001885.

24. Bureau USC. American Community Survey, 2016 American Community Survey 5-Year Estimates Table S1901. 2018.
25. Freedman DA. On the so-called Huber sandwich estimator and robust standard errors. The American Statistician. 2006; 60: 299-302.
26. Kazis L, Selim A, Rogers W, et al. Dissemination of Methods and Results From the Veterans Health Study Final Comments and Implications for Future Monitoring Strategies Within and Outside the Veterans Healthcare System. J Ambul Care Manage. 2006; 29: 310-319.

© 2026 Adam D. Lutz, et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License