Risk-based Assessment of Heterogeneity of Treatment Effect Using PATH Strategies in LESS Trial

Sponsored by:

Department of Biostatistics, University of Washington
Department of Radiology and Neurological Surgery, University of Washington

Investigators:

Yitao Wu, MS
University of Washington,
Seattle WA
206-226-7972
yitaow2@uw.edu

Joanna Xiao, MS
University of Washington,
Seattle WA
206-468-4156
ziyuxiao@uw.edu

Pinyan Liu, MS
University of Washington,
Seattle WA
206-209-7501
pliu5@uw.edu

Protocol Version: V1.0.20201215

TABLE OF CONTENTS

1. Background

1.1 Current status and limitation of back pain study

The prevention of back pain itself is difficult, and treatments for back pain are rarely curative. Preventing back-related functional limitations and disability is viewed as the most important outcome domain in back pain research. Back-related functional limitations and disability are typically measured using patient-reported outcomes (PROs), in which individuals report the extent to which back pain limits their mobility and ability to perform activities of daily living. A current limitation is that no administrative based prognostic score for PRO-measured back-related functional limitations have been generated for back pain research. Therefore, our first goal of this project is to develop and validate a predictive algorithm for coding administrative health data that can provide a prognostic score for functional limitations related to back pain. The goal of developing this outcome prediction model is to then use this as a pre-specified derived covariate that would then be calculated in the LESS cohort. We aim to stratify patients within clinical trials and examine risk-based variation in treatment effects by evaluating whether the effect of treatment is modified by this score. We are interested in illustrating one of the strategies that is suggested in the PATH statement, and this has not been illustrated in musculoskeletal research.

1.2 BOLD cohort

The algorithm will be developed using the Back-Pain Outcomes Using Longitudinal Data (BOLD) Cohort. The BOLD cohort includes about 5200 patients 65 years or older initiating a new episode of care for back pain from 3 integrated health care systems. [Harvard Vanguard (Boston), Henry Ford Health System (Detroit), Kaiser-Permanente Northern California]. The Roland-Morris Disability Questionnaire (0-24 numerical rating scale [RMDQ]) and back pain (0-10 numerical rating scale [NRS]) were used in BOLD to assess back-related functional limitations and disability at the study baseline and at follow-up assessments 3 months, 6 months, 12 months, and 24 months after the baseline assessment.

1.3 LESS trial

This is a randomized, double-blinded trial of epidural glucocorticoid injections for spinal stenosis. Epidural glucocorticoid injections are widely used to treat symptoms of lumbar spinal stenosis, a common cause of pain and disability in older adults. However, rigorous data are lacking regarding the effectiveness and safety of these injections. In this trial, the researchers randomly assigned 400 patients who had lumbar central spinal stenosis and

moderate-to-severe leg pain and disability to receive epidural injections of glucocorticoids plus lidocaine or lidocaine alone. The result is that at 6 weeks, there were no significant between-group differences in the RMDQ score or the intensity of leg pain. Therefore, in the treatment of lumbar spinal stenosis, epidural injection of glucocorticoids plus lidocaine offered minimal or no short-term benefit as compared with epidural injection of lidocaine alone.

1.4 PATH Statement

Predictive Heterogeneity of Treatment Effect (HTE) analyses have been described as approaches that provide predictions of potential outcomes in a particular patient with one intervention versus an alternative, taking into account multiple relevant patient characteristics.[4] The goal of predictive HTE analysis is to develop models that can be used to predict which of two or more treatments will be better for a particular individual, taking into account multiple relevant variables and the PATH guidance is limited to predictive approaches to HTE.[5,6] In short, such analyses aim to provide patient-centered evidence in support of decision making.

The PATH (Predictive Approaches to Treatment effect Heterogeneity) Statement was developed using a multidisciplinary technical expert panel, targeted literature reviews, simulations to characterize potential problems with predictive approaches, and a deliberative process engaging an expert panel[7]. This statement recommended a promising approach, “risk modeling”, in which treatment effects are estimated in strata of predicted risk (Figure 1). To be specific, first, a multivariable regression model that predicts risk for an outcome (usually the primary study outcome) is identified from external sources (an “external model”) or developed directly on the trial population without a term for treatment assignment (an “internal model”). Next, this model is applied to stratify patients within trials and examine risk-based variation in treatment effects. Such method avoids the low power and multiplicity issues in the “effect modeling” method, which is similar to conventional subgroup analysis. Thus, we will select the risking modeling methods for prediction and evaluation.

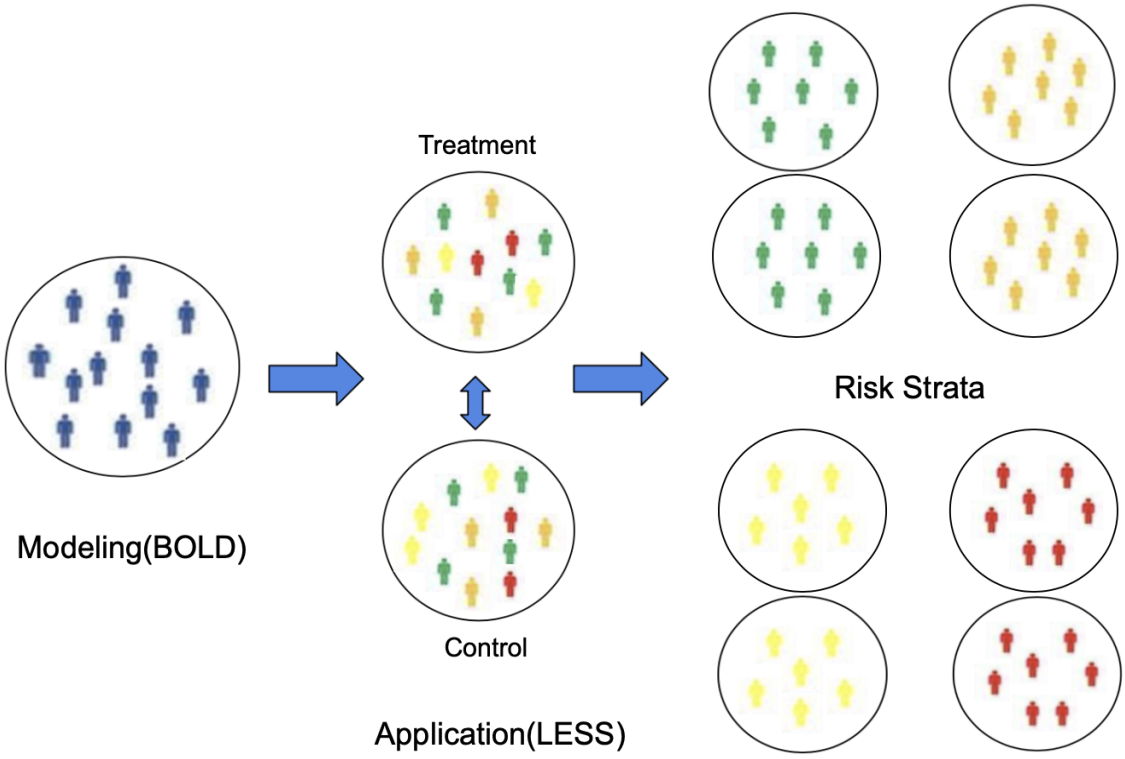

Figure 1. Modeling process including: 1) Develope a multivariable regression-based model that predicts risk for outcomes(usually select primary outcome) using BOLD data 2) Apply the model to stratify patients within the LESS trial and examine risk-based variation in treatment effects within each risk strata

2. Statistical analysis plan

2.1 General Scientific Questions

Will it be possible to develop and validate a prognostic variable for functional limitations related to back pain based on baseline characteristics? How does this prognostic score perform to compare the effect of treatment relative to the control condition in patients who had lumbar central spinal stenosis and moderate-to-severe leg pain and disability with respect to outcomes?

2.2 Specific Aims

2.2.1 Aim 1

Develop and validate a predictive algorithm that can provide a prognostic score for functional limitations related to back pain using data from BOLD (EHR and PROs).

2.2.2 Aim 2

Apply the prognostic score developed in Aim 1 to stratify patients within the LESS cohort to examine risk-based variation in treatment effects.

2.3 Datasets

For BOLD cohort data, administrative health data are available from electronic health records (EHR) for use in development of the algorithm with demographic variables such as age, sex, race, ethnicity, and insurance status. International Classification of Diseases, Ninth Revision (ICD-9) codes, Current Procedural Terminology (CPT) codes that reflect medical procedures or services performed, as well as pharmacy data are also included. EHR data is available beginning 12 months prior to baseline until 24 months after the baseline assessment. PRO datasets include all the outcomes and demographic data of patients.
For the LESS data, we also have their PRO and EHR data. But only a little more than half have the EMR data. Besides, we do have pain and function outcomes, and multiple follow-up times.

2.4 Outcomes of Interest

2.4.1 Aim 1

2.4.1.1 Primary outcome

- 3-month RMDQ risk scores (continuous)
- 3-month back pain NRS risk scores (continuous)

2.4.1.2 Secondary outcome

- 3-month probability of 30% improvement in RMDQ/back pain NRS scores (binary)
- 3-month probability of 50% improvement in RMDQ/back pain NRS scores (binary)

If time permitted:

- 24-month RMDQ/back pain NRS risk scores (continuous).
- 24-month probability of 30% improvement in RMDQ/back pain NRS scores (binary)
- 24-month probability of 50% improvement in RMDQ/back pain NRS scores (binary)

2.4.2 Aim 2

2.4.2.1 Primary outcome

- 6-week RMDQ risk scores (continuous)
- 6-week back pain NRS (continuous)

2.4.2.2 Secondary outcome

- 6-week probability of 30% or 50% improvement in RMDQ risk scores (binary)
- 6-week probability of 30% or 50% improvement in back pain NRS (binary)

In order to ensure the power of analyses, we will treat both RMDQ and NRS as continuous and binary. Besides, when the baseline score is taken into account, a 30% improvement was considered a useful threshold for identifying clinically meaningful improvement on each of these measures[3]. And we are also interested in the 50% improvement. Therefore, the transformed binary outcomes will be 30% or 50% improvement.

2.5 Predictors of Interest

Baseline characteristics will be chosen from the 12-month period prior to the index visit, including PRO and EHR data.

Predictors from PRO data are patient self-report age, gender, ethnicity, education, employment status, marital status, smoking status, hospital sites, baseline RMDQ, baseline leg pain NRS, baseline back pain NRS, BPI score, EQ5D-index, EQ5D-VAS, falls and confidence from patients, etc. Predictors from EHR data include ICD9 code, CPT code, Quan comorbidity score, insurance type, etc.

BPI (Brief Pain Inventory) score: Here, we calculated the meanbpi=mean(bpi_enjoy,bpi_general,bpi_mood,bpi_walk,bpi_work,bpi_relate,bpi_sleep)
Falls: In the past 3 weeks how many times have patients fallen? (baseline)
confidence from patients: How confident is the patient that the back or leg pain will be completely gone or will be much better 3 months from now?
Quan comorbidity score: We calculated this score using electronic medical record data for the 365 days before the index visit.
Baseline diagnosis: We used ICD-9-CM [8] codes to categorize the patient's back pain diagnosis into one of the following categories: (1) back pain only, (2) back and leg pain, (3) spinal stenosis, or (4) other.
Current procedure terminology: We divided them into four categories (1) manual, (2) imaging, (3) injection, or (4) spine surgery code.

Noticing that half of the patients do not have electronic health records in the LESS cohort, we will also construct a reduced model with the variables in the PRO dataset.

2.6 Type of Analysis

(1) Regression Modeling

(2) Method evaluation
(3) Descriptive statistics
(4) Prediction and validation

2.7 Analysis Approach and Special Issues

2.7.1 Aim 1: Methods for building the predictive model

We will utilize the target cohort (BOLD) to develop multivariable regression models that predict the risk of the outcome. Our primary outcome will be the 3-month RMDQ score (continuous variable). We will develop predictive models for both 3-month RMDQ and NRS using the least absolute shrinkage and selection (LASSO) regression method, a machine learning statistical model. The characteristics of the LASSO method is to penalize the absolute value of the model coefficients, and then we eliminate coefficients that become zero from the model.

Two types of models will be constructed in our analysis; a full model includes variables from both PRO and EHR data and a reduced model only includes variables from PRO data. For the full model, we also plan to evaluate a series of models ranging in complexity from a simple parsimonious model to more complex models by adding more refined coding of CPT codes, finer categorization of comorbidity, etc.

We will first split the whole BOLD cohort into a training set and a testing set with a ratio of 4:1. Since we categorize the patient's back pain diagnosis into one of the following categories using ICD-9 code: (1) back pain only, (2) back and leg pain, (3) spinal stenosis, or (4) other, we will make sure there will be enough patients (20% of each category) in the testing set. Then, we plan to do a 5-fold cross validation on the training set to choose the best model in this step, which has the highest r square. We will evaluate the performance of the selected models on the testing set and summarize the result for both full model and reduced model. Using variables selected by LASSO, we will also construct linear regression models and compare the performance with our models.

In order to evaluate model prediction for the continuous outcomes, we will calculate how well the models explain variations in outcomes using the coefficients of determination (R^2). To evaluate model discrimination for the dichotomous outcomes, we will construct receiver operator characteristics (ROC) curves on the validation data using the predicted probabilities derived from the training data and AUC (area under the ROC curve) will be one of the criterion for choosing the best model. Scatter plot comparing actual RMDQ with predicted RMDQ will also be investigated to evaluate modeling performance.

Finally, on the testing test, we report unstandardized model coefficients (for linear regressions), odds ratios (for logistic regressions).

2.7.2 Aim2: Treatment Effect Modeling to Identify HTE

For the full model that has variables in the EMR dataset, we will use it to predict the 6-week RMDQ risk scores in the subset of LESS cohort who both have EMR and PRO data and evaluate heterogeneity of treatment effect. In general, we will divide the patients in LESS into 4 groups (quartiles) based on their RMDQ risk scores. Later, t-test(no) or linear regression (including an interaction term between the predictive risk score and the treatment) will be used to test treatment effects within each stratum. The predictive risk score could be either continuous or probabilities representing each risk strata. The model will be: 6-week RMDQ ~ Predictive scores + treatment group + Predictive scores*treatment group+same covariates at baseline. Here, we can evaluate whether the association between treatment and outcome (6-week RMDQ) is the same for patients with different predictive scores. And then we will estimate relative and absolute treatment effects for all outcomes of interest to identify treatment effects. We will use forest plots to illustrate the treatment effects in the various risk strata. It is important to evaluate treatment effects in both scales, as-assuming a non-zero treatment effect-treatment effect cannot remain constant on both the relative and the absolute scale at the same time. Linear regression and logistic for the evaluation of absolute and relative treatment effects can be considered, as long as this is done consistently in all risk strata.

Risk group 0,1,2,3

Risk1+risk2+risk3+
For the reduced model, we will evaluate the treatment effect in the whole LESS cohort and also in the LESS cohort who have both EMR and PRO data.

2.8 Result presentation and evaluation

- The number of patients and person years by treatment arm along with the number of outcomes
- Report metrics for model performance for outcome prediction on the RCT, including measures of discrimination and calibration (when appropriate)
- Report distribution of predicted risk (or the risk score) in each group of the trial and in the overall study population
- Report outcome rates and both relative and absolute risk reduction across risk strata
- Event rates, hazard ratios and absolute risk differences in risk strata for a selected outcome, both in tables and in graphs

- Mean differences in t-tests will only be reported as absolute results if the outcome is a continuous variable. In other words, relative differences may not be reported in t-tests.
- To test the consistency of the relative treatment effect across prognostic risk, a continuous measure of risk (for example, the logit of risk) may be used in an interaction term with treatment group indicator.

2.9 List of Tables and Figures

2.9.1 Aim 1

- Table 1.Baseline demographic data on the population: age, gender, race, education, working status, smoking status, lawyer, Quan comorbidity, baseline diagnosis category, cpt code and marital status, stratified by sites (for both BOLD and LESS).
- Table 2. Baseline PROs by site. PROs include the following measures: RMDQ score, back pain NRS score, leg pain NRS score, BPI score, BPI-Sleep score, PHQ-4 score, EQ5D Index (baseline), EQ5D-index, EQ5D VAS score, falls and confidence (for both BOLD and LESS).
- Table 3. Coefficients or adjusted odds ratios (95% CIs) from generalized multivariable linear and logistic regression models with continuous and dichotomous RMDQ and NRS Outcomes.
- Figure 1. Unadjusted means and standard errors for patient-reported outcomes (PRO) at baseline, 3 month, 6 month and 12 month in BOLD dataset.
- Figure 2. Distribution of changes of PROs from 0 to 90 days in BOLD dataset
- Table 4. Tables showing the values of the AUC and R square for comparison of the various models.

2.9.2 Aim 2

- Table 5. Coefficients and P-Values from t-tests or linear regression models with continuous and dichotomous RMDQ and NRS Outcomes.
- Figure 3. A forest plot to illustrate the treatment effects in the various risk strata.

2.10 Data domains needed

- Prescriptions
- Comorbidity-Charlson
- VitalSigns-BMI
- PROs

3. Revision History

Original – 28 Oct 2020
Revisions for SAP - 20 Nov 2020
Revisions for time-period of outcomes - 1 Dec 2020
Final revisions - 15 Dec 2020

4. References

[1] JarvikJG, Gold LS, Comstock BA, et al. Association of early imaging for back pain with clinical outcomes in older adults.
[2] JarvikJG, Gold LS, Tan K, et al. Long-term outcomes of a large, prospective observational cohort of older adults with back pain.
[3]Ostelo, Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change, 2008.
[4] Kent, David M., et al. "The predictive approaches to treatment effect heterogeneity (PATH) statement." Annals of internal medicine 172.1 (2020): 35-45.
[5] Kent DM , Steyerberg E , and van Klaveren D . Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018;363:k4245. [PMID: 30530757] doi:10.1136/bmj.k4245
[6] Varadhan R , Segal JB , Boyd CM , et al. A framework for the analysis of heterogeneity of treatment effect in patient-centered outcomes research. J Clin Epidemiol. 2013;66:818-25. [PMID: 23651763] doi:10.1016/j.jclinepi.2013.02.009
[7] Rekkas, Alexandros, et al. "A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases." arXiv preprint arXiv:2010.06430 (2020).
[8] StatisticsNCfH.InternationalClassificationofDiseases,NinthRevision (ICD-9). 2009. Available at: http://www.cdc.gov/nchs/icd/icd9.htm. Accessed February 12, 2018.

5. Appendices

Tables & Figures

Table 1. Descriptive statistics of demographic characteristics of patients.

	Henry Ford (N=743)	Kaiser (N=2732)	Harvard Vanguard (N=584)	Total (N=4059)
age				
Mean (SD)	73.7 (6.88)	73.8 (6.66)	73.9 (6.93)	73.8 (6.74)
Median [Min, Max]	72.0 [65.0, 101]	73.0 [65.0, 98.0]	72.5 [65.0, 94.0]	73.0 [65.0, 101]
sex				
0	527 (70.9%)	1737 (63.6%)	371 (63.5%)	2635 (64.9%)
1	216 (29.1%)	995 (36.4%)	213 (36.5%)	1424 (35.1%)
hispanic				
0	731 (98.4%)	2511 (91.9%)	576 (98.6%)	3818 (94.1%)
1	12 (1.6%)	221 (8.1%)	8 (1.4%)	241 (5.9%)
education				
0	112 (15.1%)	119 (4.4%)	6 (1.0%)	237 (5.8%)
1	266 (35.8%)	552 (20.2%)	140 (24.0%)	958 (23.6%)
2	33 (4.4%)	116 (4.2%)	7 (1.2%)	156 (3.8%)
3	185 (24.9%)	855 (31.3%)	42 (7.2%)	1082 (26.7%)
4	67 (9.0%)	528 (19.3%)	292 (50.0%)	887 (21.9%)
5	80 (10.8%)	562 (20.6%)	97 (16.6%)	739 (18.2%)
Working full-time				
0	713 (96.0%)	2585 (94.6%)	543 (93.0%)	3841 (94.6%)
1	30 (4.0%)	147 (5.4%)	41 (7.0%)	218 (5.4%)
Working part-time				

0	722 (97.2%)	2568 (94.0%)	548 (93.8%)	3838 (94.6%)
1	21 (2.8%)	164 (6.0%)	36 (6.2%)	221 (5.4%)
Retired(not due to ill health)				
0	127 (17.1%)	513 (18.8%)	88 (15.1%)	728 (17.9%)
1	616 (82.9%)	2219 (81.2%)	496 (84.9%)	3331 (82.1%)
On leave of absense				
0	740 (99.6%)	2721 (99.6%)	583 (99.8%)	4044 (99.6%)
1	3 (0.4%)	11 (0.4%)	1 (0.2%)	15 (0.4%)
Unemployed and looking for work				
0	743 (100%)	2708 (99.1%)	582 (99.7%)	4033 (99.4%)
1	0 (0%)	24 (0.9%)	2 (0.3%)	26 (0.6%)
Retired or disabled because of ill health				
0	693 (93.3%)	2679 (98.1%)	578 (99.0%)	3950 (97.3%)
1	50 (6.7%)	53 (1.9%)	6 (1.0%)	109 (2.7%)
Homemaker				
0	729 (98.1%)	2688 (98.4%)	584 (100%)	4001 (98.6%)
1	14 (1.9%)	44 (1.6%)	0 (0%)	58 (1.4%)
Other				
0	734 (98.8%)	2662 (97.4%)	582 (99.7%)	3978 (98.0%)
1	9 (1.2%)	70 (2.6%)	2 (0.3%)	81 (2.0%)
marital				

0	342 (46.0%)	1551 (56.8%)	456 (78.1%)	2349 (57.9%)
1	4 (0.5%)	48 (1.8%)	1 (0.2%)	53 (1.3%)
2	7 (0.9%)	21 (0.8%)	1 (0.2%)	29 (0.7%)
3	100 (13.5%)	335 (12.3%)	23 (3.9%)	458 (11.3%)
4	49 (6.6%)	150 (5.5%)	15 (2.6%)	214 (5.3%)
5	241 (32.4%)	627 (23.0%)	88 (15.1%)	956 (23.6%)
lawyer				
0	738 (99.3%)	2718 (99.5%)	584 (100%)	4040 (99.5%)
1	5 (0.7%)	14 (0.5%)	0 (0%)	19 (0.5%)
smokingstatus				
0	521 (70.1%)	1348 (49.3%)	351 (60.1%)	2220 (54.7%)
1	137 (18.4%)	1244 (45.5%)	215 (36.8%)	1596 (39.3%)
2	85 (11.4%)	140 (5.1%)	18 (3.1%)	243 (6.0%)
Baseline diagnosis category				
Back pain only	678 (91.3%)	1671 (61.2%)	382 (65.4%)	2731 (67.3%)
Back and leg pain	43 (5.8%)	755 (27.6%)	91 (15.6%)	889 (21.9%)
Spinal stenosis	13 (1.7%)	127 (4.6%)	62 (10.6%)	202 (5.0%)
Others	9 (1.2%)	179 (6.6%)	49 (8.4%)	237 (5.8%)
Manual spine-related CPT				
No	287 (38.6%)	2709 (99.2%)	565 (96.7%)	3561 (87.7%)
Yes	456 (61.4%)	23 (0.8%)	19 (3.3%)	498 (12.3%)
Percutaneous spine-related CPT				

No	732 (98.5%)	2694 (98.6%)	574 (98.3%)	4000 (98.5%)
Yes	11 (1.5%)	38 (1.4%)	10 (1.7%)	59 (1.5%)
Spine image-related CPT				
No	473 (63.7%)	2011 (73.6%)	455 (77.9%)	2939 (72.4%)
Yes	270 (36.3%)	721 (26.4%)	129 (22.1%)	1120 (27.6%)
Spine surgery-related CPT				
No	743 (100%)	2724 (99.7%)	582 (99.7%)	4049 (99.8%)
Yes	0 (0%)	8 (0.3%)	2 (0.3%)	10 (0.2%)

	Henry Ford (N=743)	**Kaiser (N=2732)**	**Harvard Vanguard (N=584)**	**Total (N=4059)**
RDQ score(baseline)				
Mean (SD)	12.8 (5.83)	9.57 (6.02)	6.65 (6.49)	9.74 (6.30)
Median [Min, Max]	13.0 [0, 24.0]	10.0 [0, 24.0]	5.00 [0, 24.0]	10.0 [0, 24.0]
NRS leg pain(baseline)				
Mean (SD)	4.22 (3.57)	3.46 (3.21)	2.64 (3.08)	3.48 (3.29)
Median [Min, Max]	5.00 [0, 10.0]	3.00 [0, 10.0]	1.00 [0, 10.0]	3.00 [0, 10.0]
NRS back pain(baseline)				
Mean (SD)	6.20 (2.73)	4.85 (2.73)	4.48 (2.70)	5.04 (2.78)
Median [Min, Max]	7.00 [0, 10.0]	5.00 [0, 10.0]	5.00 [0, 10.0]	5.00 [0, 10.0]
Brief Pain Inventory score(baseline)				
Mean (SD)	3.81 (2.48)	3.39 (2.43)	2.62 (2.31)	3.36 (2.44)

Median [Min, Max]	3.86 [0, 10.0]	3.14 [0, 10.0]	2.00 [0, 10.0]	3.14 [0, 10.0]
EQ5D score 0-100(baseline)				
Mean (SD)	74.3 (17.9)	74.1 (18.3)	77.1 (17.1)	74.5 (18.1)
Median [Min, Max]	75.0 [0, 100]	80.0 [0, 100]	80.0 [7.50, 100]	80.0 [0, 100]
EQ5D score 0-1(baseline)				
Mean (SD)	0.692 (0.194)	0.760 (0.165)	0.805 (0.149)	0.754 (0.172)
Median [Min, Max]	0.761 [-0.0384, 1.00]	0.778 [-0.0402, 1.00]	0.810 [0.0494, 1.00]	0.778 [-0.0402, 1.00]
Times of fallen in past 3 weeks(baseline)				
Mean (SD)	0.110 (0.518)	0.128 (0.520)	0.0171 (0.193)	0.109 (0.487)
Median [Min, Max]	0 [0, 10.0]	0 [0, 8.00]	0 [0, 3.00]	0 [0, 10.0]
Patients' expectation in 3 months				
Mean (SD)	4.82 (3.64)	5.77 (3.59)	5.03 (3.99)	5.49 (3.68)
Median [Min, Max]	5.00 [0, 10.0]	6.00 [0, 10.0]	5.00 [0, 10.0]	5.00 [0, 10.0]

Less trial

	Henry Ford (N=107)	**Kaiser (N=59)**	**Harvard Vanguard (N=25)**	**Total (N=191)**
age				
Mean (SD)	66.4 (9.70)	68.7 (9.98)	68.4 (8.90)	67.4 (9.70)
Median [Min, Max]	66.0 [50.0, 87.0]	70.0 [51.0, 96.0]	68.0 [52.0, 86.0]	68.0 [50.0, 96.0]
sex				

0	66 (61.7%)	35 (59.3%)	22 (88.0%)	123 (64.4%)
1	41 (38.3%)	24 (40.7%)	3 (12.0%)	68 (35.6%)
hispanic				
0	103 (96.3%)	55 (93.2%)	24 (96.0%)	182 (95.3%)
1	4 (3.7%)	4 (6.8%)	1 (4.0%)	9 (4.7%)
education				
0	17 (15.9%)	1 (1.7%)	0 (0%)	18 (9.4%)
1	35 (32.7%)	11 (18.6%)	5 (20.0%)	51 (26.7%)
2	3 (2.8%)	0 (0%)	0 (0%)	3 (1.6%)
3	32 (29.9%)	20 (33.9%)	5 (20.0%)	57 (29.8%)
4	9 (8.4%)	10 (16.9%)	5 (20.0%)	24 (12.6%)
5	11 (10.3%)	17 (28.8%)	10 (40.0%)	38 (19.9%)
Working full-time				
0	90 (84.1%)	41 (69.5%)	19 (76.0%)	150 (78.5%)
1	17 (15.9%)	18 (30.5%)	6 (24.0%)	41 (21.5%)
Working part-time				
0	102 (95.3%)	51 (86.4%)	21 (84.0%)	174 (91.1%)
1	5 (4.7%)	8 (13.6%)	4 (16.0%)	17 (8.9%)
Retired(not due to ill health)				
0	52 (48.6%)	30 (50.8%)	16 (64.0%)	98 (51.3%)
1	55 (51.4%)	29 (49.2%)	9 (36.0%)	93 (48.7%)
On leave of absence				
0	100 (93.5%)	59 (100%)	25 (100%)	184 (96.3%)
1	7 (6.5%)	0 (0%)	0 (0%)	7 (3.7%)
Unemployed and looking for work				
0	104 (97.2%)	59 (100%)	22 (88.0%)	185 (96.9%)
1	3 (2.8%)	0 (0%)	3 (12.0%)	6 (3.1%)

Retired or disabled because of ill health				
0	90 (84.1%)	56 (94.9%)	23 (92.0%)	169 (88.5%)
1	17 (15.9%)	3 (5.1%)	2 (8.0%)	22 (11.5%)
Homemaker				
0	106 (99.1%)	58 (98.3%)	23 (92.0%)	187 (97.9%)
1	1 (0.9%)	1 (1.7%)	2 (8.0%)	4 (2.1%)
Other				
0	104 (97.2%)	59 (100%)	25 (100%)	188 (98.4%)
1	3 (2.8%)	0 (0%)	0 (0%)	3 (1.6%)
marital				
0	50 (46.7%)	34 (57.6%)	10 (40.0%)	94 (49.2%)
1	1 (0.9%)	2 (3.4%)	0 (0%)	3 (1.6%)
2	1 (0.9%)	2 (3.4%)	0 (0%)	3 (1.6%)
3	18 (16.8%)	8 (13.6%)	6 (24.0%)	32 (16.8%)
4	9 (8.4%)	5 (8.5%)	3 (12.0%)	17 (8.9%)
5	28 (26.2%)	8 (13.6%)	6 (24.0%)	42 (22.0%)
lawyer				
0	104 (97.2%)	58 (98.3%)	24 (96.0%)	186 (97.4%)
1	3 (2.8%)	1 (1.7%)	1 (4.0%)	5 (2.6%)
smoking status				
0	39 (36.4%)	28 (47.5%)	10 (40.0%)	77 (40.3%)
1	41 (38.3%)	27 (45.8%)	14 (56.0%)	82 (42.9%)
2	27 (25.2%)	4 (6.8%)	1 (4.0%)	32 (16.8%)
Baseline diagnosis category				
Back pain only	8 (7.5%)	0 (0%)	0 (0%)	8 (4.2%)
Back and leg pain	25 (23.4%)	3 (5.1%)	1 (4.0%)	29 (15.2%)
Spinal stenosis	51 (47.7%)	20 (33.9%)	22 (88.0%)	93 (48.7%)

Others	23 (21.5%)	36 (61.0%)	2 (8.0%)	61 (31.9%)
Manual spine-related CPT				
No	0 (0%)	0 (0%)	11 (44.0%)	11 (5.8%)
Yes	107 (100%)	59 (100%)	14 (56.0%)	180 (94.2%)
Percutaneous spine-related CPT				
No	1 (0.9%)	0 (0%)	0 (0%)	1 (0.5%)
Yes	106 (99.1%)	59 (100%)	25 (100%)	190 (99.5%)
Spine image-related CPT				
No	0 (0%)	1 (1.7%)	0 (0%)	1 (0.5%)
Yes	107 (100%)	58 (98.3%)	25 (100%)	190 (99.5%)
Spine surgery-related CPT				
No	107 (100%)	59 (100%)	25 (100%)	191 (100%)
Yes	0 (0%)	0 (0%)	0 (0%)	0 (0%)

	Henry Ford (N=107)	**Kaiser (N=59)**	**Harvard Vanguard (N=25)**	**Total (N=191)**
RDQ score(baseline)				
Mean (SD)	17.6 (3.45)	14.6 (3.87)	14.7 (3.67)	16.3 (3.88)
Median [Min, Max]	18.0 [8.00, 23.0]	15.0 [7.00, 21.0]	15.0 [7.00, 20.0]	17.0 [7.00, 23.0]
NRS leg pain(baseline)				
Mean (SD)	8.18 (1.52)	7.05 (1.52)	6.16 (1.65)	7.57 (1.70)
Median [Min, Max]	8.00 [5.00, 10.0]	7.00 [3.00, 10.0]	6.00 [4.00, 9.00]	8.00 [3.00, 10.0]
NRS back pain(baseline)				
Mean (SD)	8.43 (1.36)	5.32 (2.49)	5.80 (2.69)	7.13 (2.46)

Median [Min, Max]	9.00 [4.00, 10.0]	6.00 [0, 10.0]	6.00 [0, 10.0]	8.00 [0, 10.0]
Brief Pain Inventory score(baseline)				
Mean (SD)	6.48 (1.86)	6.56 (2.54)	5.50 (1.98)	6.37 (2.12)
Median [Min, Max]	6.50 [1.00, 10.0]	7.00 [0, 10.0]	5.00 [1.50, 10.0]	6.50 [0, 10.0]
EQ5D score 0-100(baseline)				
Mean (SD)	78.5 (12.7)	61.3 (22.1)	59.4 (19.7)	70.7 (19.2)
Median [Min, Max]	80.0 [40.0, 100]	62.0 [5.00, 100]	62.0 [25.0, 95.0]	75.0 [5.00, 100]
EQ5D score 0-1(baseline)				
Mean (SD)	0.470 (0.178)	0.674 (0.164)	0.651 (0.170)	0.557 (0.198)
Median [Min, Max]	0.397 [0.165, 0.816]	0.761 [0.165, 0.827]	0.708 [0.271, 0.810]	0.597 [0.165, 0.827]
Times of fallen in past 3 weeks(baseline)				
Mean (SD)	0.0467 (0.253)	0.0339 (0.183)	0.360 (0.907)	0.0838 (0.402)
Median [Min, Max]	0 [0, 2.00]	0 [0, 1.00]	0 [0, 4.00]	0 [0, 4.00]
Patients' expectation in 3 months				
Mean (SD)	7.35 (1.64)	7.88 (2.05)	8.12 (1.92)	7.61 (1.83)
Median [Min, Max]	8.00 [3.00, 10.0]	8.00 [4.00, 10.0]	8.00 [4.00, 10.0]	8.00 [3.00, 10.0]

Figure 1.

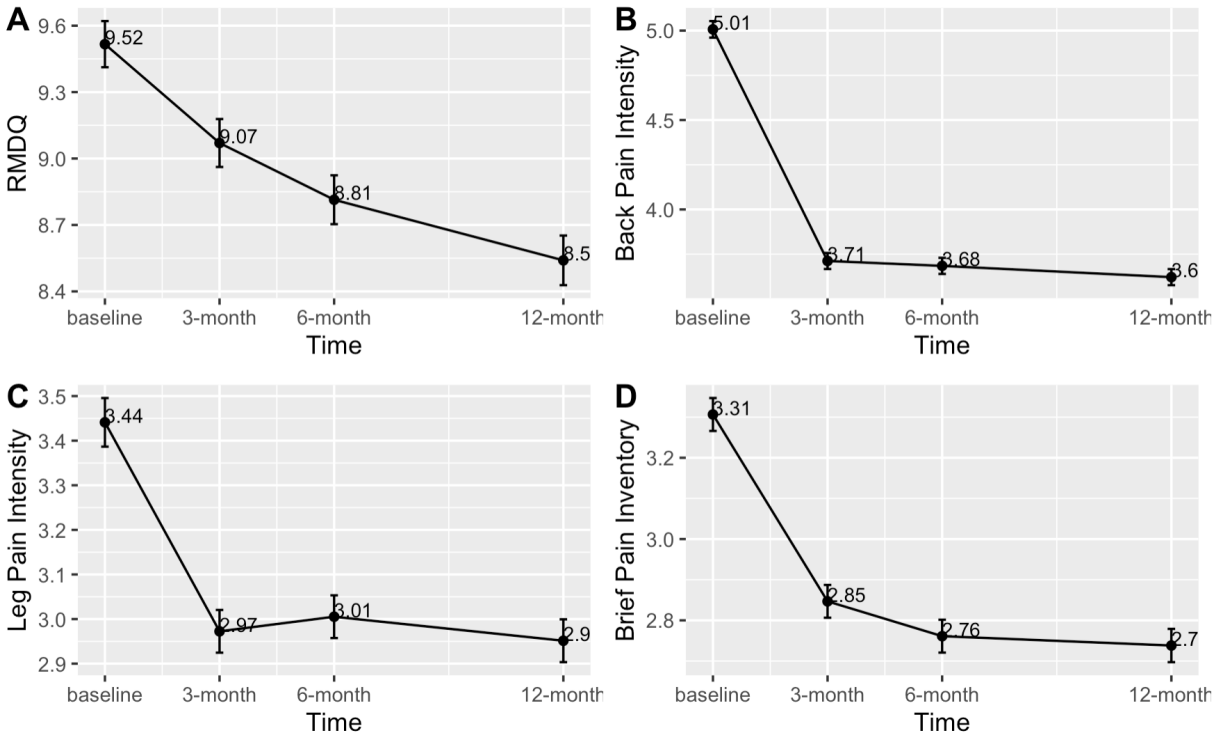

Fig. 1. (A–D) Unadjusted means and standard errors for patient-reported outcomes (PRO) at baseline, 3 month, 6 month and 12 month in BOLD dataset (A) The RMDQ means and SEs: the RMDQ at baseline and each follow-up timepoint. (B) Average back pain intensity means and SEs: the average back pain intensity (0–10 pain NRS score) at baseline and at each follow-up timepoint. (C) Average leg pain intensity means and SEs: the average leg pain intensity (0–10 pain NRS score) at baseline and at each follow-up timepoint. (D) Average BPI means and SEs: the average BPI Activity Interference Scale score at baseline and at each follow-up timepoint.

Figure 2.

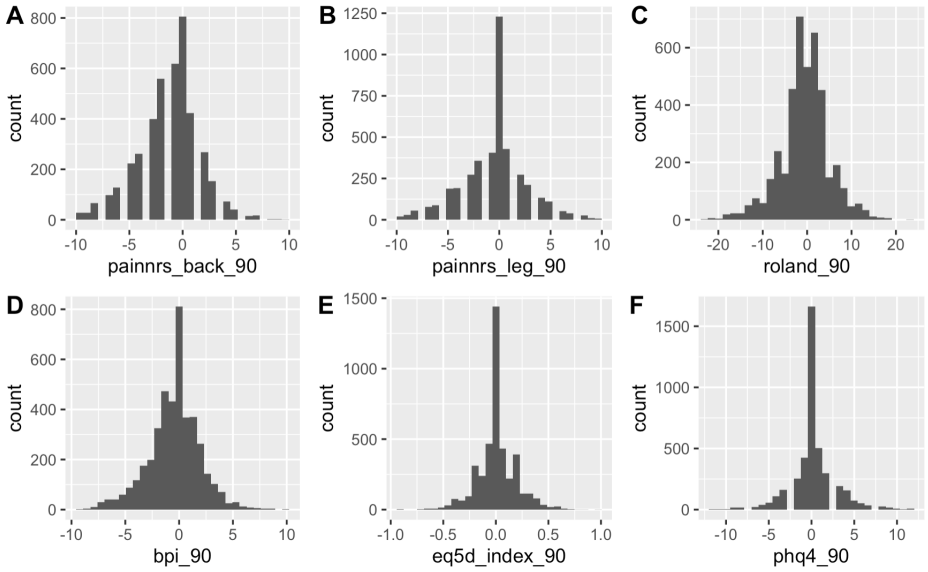

Fig. 2. Distribution of changes of PROs from 0 to 90 days in BOLD dataset (A-F). (A) Distribution of 3-month change in back pain intensity (B) Distribution of 3-month change in leg pain intensity (C) Distribution of 3-month change in RMDQ (D) Distribution of 3-month change in BPI score (E) Distribution of 3-month change in EQ5D_index (F) Distribution of 3-month change in PHQ4.

Limitation:
1.variables not consistent
2.2 cohort, 1 model
3.test on less (base on dataset)