Risk-Modeling Predictive Approaches to Treatment Effect Heterogeneity in a Randomized Trial of Epidural Corticosteroid Injections

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

Ziyu 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

1. Introduction

Lumbar spinal stenosis is a common cause of spine-related disability and functional limitations among older adults[1-2], leading to suffering back and leg pain, depression and weakness[3-4]. The prevention of such symptoms is difficult, and the effectiveness of a widely used treatment, epidural glucocorticoid injections, has been shown to be questionable. Previous research found that there were no significant between-group differences in either self-reported disability or self-reported back pain at 6 weeks[5]. Our goal for this project is to further understand how the effectiveness of epidural glucocorticoid injections for spinal stenosis may vary across patients - a concept described as Heterogeneity of Treatment Effects (HTE).

In traditional randomized controlled trials (RCTs), evaluating heterogeneity of treatment effects (HTE) is typically conducted through subgroup analysis contrasting effects in groups of patients defined one variable at a time (e.g. males versus females; old versus young). There are important limitations of these "one variable at a time" analyses, such as low statistical power and potential inflation of the type I error rate due to multiple statistical comparisons. Therefore, we are interested in adopting risk modeling approaches that account for multiple variables at the same time, one of the strategies suggested in the Predictive Approaches to Treatment effect Heterogeneity (PATH) statement[6-8].

A multidisciplinary technical expert panel developed the PATH statement using targeted literature reviews, simulations to characterize potential problems with predictive approaches, and a deliberative process engaging an expert panel. This statement recommended a promising approach, "risk modeling", in which treatment effects are estimated in strata of predicted risk (**Fig.1**)[9]. 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 "risk modeling" approach aims to provide patient-centered estimates of outcome risks with versus without the intervention, taking into account all relevant patient attributes simultaneously.

Our ultimate goal was to develop a framework for risk-based assessment of HTE for Lumbar Epidural Steroid Injections for Spinal Stenosis (LESS) trial following the PATH statement recommendations. To be more specific, our first aim was to develop and validate a model in an external dataset to generate risk scores. In our study, we chose the Back-Pain Outcomes Using Longitudinal Data (BOLD) Cohort[10-11], which consisted of rich

baseline measurements collected from patient-reported outcomes (PROs) and electronic health data (EHR). Our second aim was to create risk scores for patients in the LESS trial using the selected model in Aim 1. We treated the risk score as a pre-specified derived covariate to stratify those patients and explored risk-based variation in treatment effects by evaluating whether the effectiveness of epidural glucocorticoid injections is modified by the risk scores. This HTE analysis would provide valuable insight to comprehensively assess the effectiveness and safety of epidural glucocorticoid injections.

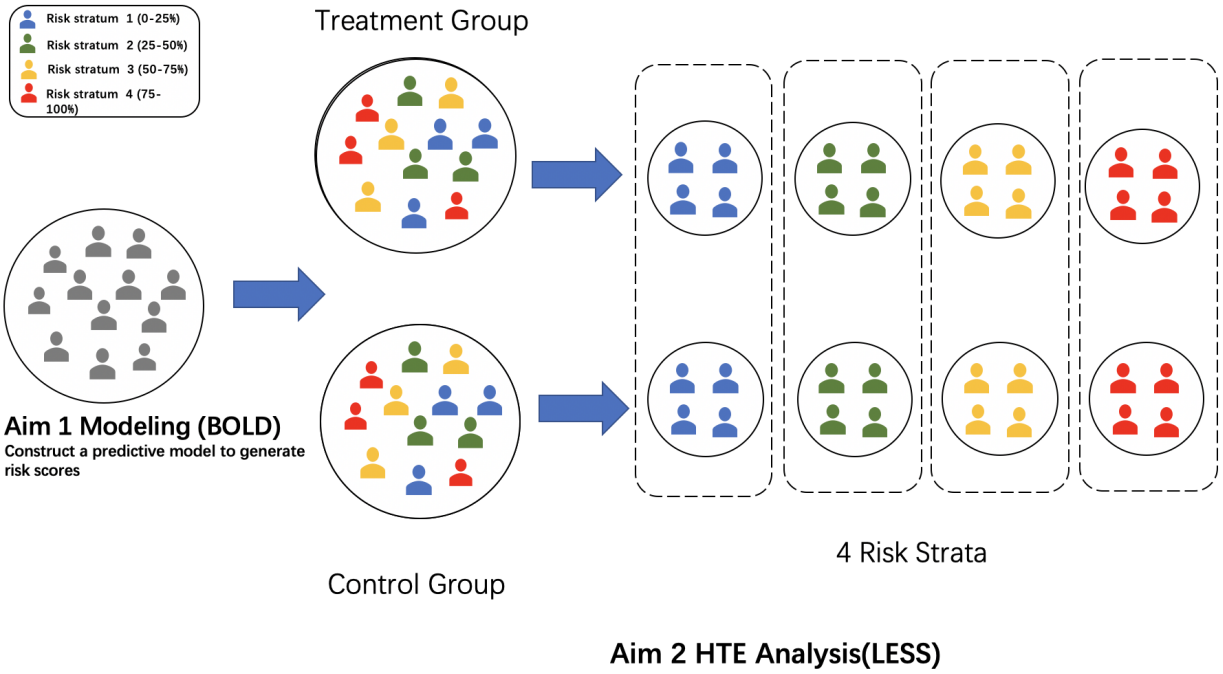

Fig. 1 The modeling process includes: 1) Developing a multivariable regression-based model that predicts risk for outcomes(usually select primary outcome) using BOLD data 2) Applying the model to stratify patients within the LESS trial and examine risk-based variation in treatment effects within each risk strata

2. Methods

2.1 Patients and settings

2.1.1 BOLD cohort

The BOLD cohort includes 5239 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 BOLD research team obtained sociodemographic characteristics (gender, race, ethnicity, education, etc.) at baseline and the PROs at baseline and then 3, 6, 12 and 24 months after enrollment, either by mailed questionnaire or by telephone follow-up. We considered the RMDQ (Roland-Morris Disability Questionnaire)[12] and back pain NRS scores (0–10 numerical rating scale) [13] as the assessment of back-related functional limitations/disability and back-related pain respectively, which were measured at baseline and longitudinally.

Electronic Health Records (EHR) data was also available beginning 12 months prior to baseline (the index clinical encounter for back pain) through the baseline assessment. Scores reflecting medical conditions and comorbidities[14], International Classification of Diseases, Ninth Revision (ICD-9) codes[15] and Current Procedural Terminology (CPT) codes[16] that reflect medical procedures or services performed were included in the EHR dataset.

2.1.2 LESS trial cohort

We investigated HTE in the LESS trial. This was a double-blinded, randomized controlled trial of epidural glucocorticoid injections for lumbar spinal stenosis, as compared to control epidural injections using local anesthetic only. 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. For the LESS data, we also had their PROs at baseline and 3 weeks after enrollment. EHR data including ICD-9 codes, CPT codes and comorbidity records were also available. Descriptive statistics of EHR data and PROs data in the BOLD and LESS cohort are listed in Table 1 and Table 2 respectively.

2.1.3 Inclusion criteria

Our primary inclusion and exclusion criteria were based on the baseline RMDQ and back pain NRS scores. There were 696 patients with low levels of functional limitations (RMDQ scores 0-2) in the BOLD cohort but none in the LESS trials. Similarly, there were 534 patients with low levels of back pain (NRS scores 0-1) in the BOLD cohort while the number was 15 in the LESS trials **(Fig.10 in appendix**). Due to this difference, we made adjustments based on the distribution of baseline RMDQ and back pain NRS scores to make the BOLD and LESS samples more comparable, by excluding patients with baseline RMDQ <= 2 and back pain NRS scores < 2 to adjust for such inconsistency.

Also, more than half of the LESS sample (221 patients) did not have EHR data. To account for this, we made two versions of the LESS trial data. The first version included all patients (LESS-1) and the second version included only those with both PROs and EHR data (LESS-2). We used complete case analysis and excluded those patients in our datasets who had any data missing.

2.2 Measures

2.2.1 Predictor of Interest

We chose baseline characteristics from the 12-month period prior to the index visit, including PRO and EHR data. All the baseline variables mentioned below were available in both BOLD and LESS. Predictors reflecting sociodemographic from PRO data were age, gender, BMI scores, ethnicity, education, employment status, marital status, having a lawyer involved in current pain or health, smoking status.

More detailed information about additional variables from PROs and EHR:

Roland-Morris Disability Questionnaire(RMDQ): 24-item Roland-Morris Disability Questionnaire (RMDQ) modified to specify disability related to either back or leg pain

Pain Numerical Rating Scale (NRS): Averaged back pain intensity and averaged leg pain intensity in the past week on 0–10 numerical rating scales

Health-related quality of life (HRQoL): European Quality of Life 5 Dimension (EQ5D)[17], including both the quality of life index (0–1, with 0 being death and 1 being perfect health) (European Quality of Life 5 Dimension Index [EQ5D-Index]) and the visual analog scale (0-100, with 0 being “a health state worse than death” and 100 being “perfect health”) (European Quality of Life 5 Dimension Visual Analog Scale [EQ5D-VAS])

Brief Pain Inventory score: The validated Brief Pain Inventory (BPI) Interference scale [18] measures pain interference with activities .The scale consists of 7 ratings (0–10) of how much back pain interferes with the following: gen- eral activity, mood, ability to walk, normal work, rela- tions with other people, sleep and enjoyment of life.

Falls: Number of falls in the past 3 weeks and how many resulted in injury, from the Behavioral Risk Factor Surveillance System (BRFSS) survey [19].

Psychological distress: The Patient Health Questionnaire-4 (PHQ-4) measure of anxiety and depressive symptoms [20].

Confidence from patients: Patients rated their confidence that their back and/or leg pain would be completely gone or much better in 3 months on a scale from 0 = 'not at all confident' to 10 = 'extremely confident'[21-22].

Body mass index (BMI): We calculated BMI using height and weight at baseline from the EHR data.

Quan comorbidity score: We calculated this score using EHR data for the 12 months before the index visit.

Baseline diagnosis: We used ICD-9-CM 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: Using Current Procedural Ter- minology (CPT) code and ICD-9-CM diagnosis code data , we assessed whether certain spine-related interventions occurred between the index visit and 12 months earlier by dividing back pain diagnostics and treatments into four categories (1) manual treatments, (2) diagnostic imaging, (3) injection treatments, or (4) spine surgery[23].

2.2.2 Outcome variables

Aim1: Develop Risk Scores in BOLD
Primary outcome: 3-month RMDQ (continuous variable)
Secondary outcomes: 3-month NRS back pain (continuous variable)

Aim 2: HTE analysis in LESS
Outcome: 1. Actual 6-week RMDQ; 2. Actual 6-week back pain NRS;

2.3 Statistical analysis

2.3.1 Aim 1: Methods for building the predictive model

We utilized the BOLD cohort to develop multivariable regression models that predict the risk of the outcome. Our primary outcome was the 3-month RMDQ (continuous variable). We developed predictive models using the least absolute shrinkage and selection (LASSO) regression method, a machine learning statistical model that penalizes the absolute value of the model coefficients. It selected a reduced set of the known covariates for use in a model.

Three sequential LASSO models were constructed in our analysis (**Table 1**) for each outcome: The candidate variables for Model 1 (the "baseline model") were the baseline variables including socio-demographics and baseline levels for pain, disability, and several other measures (**Table 3**).

The candidate variables for Model 2 (the "full model") were all the Model 1 variables plus baseline spine-related diagnostic and therapeutic interventions (ICD-9 codes and CPT codes respectively), categorical Quan comorbidity, and study sites. Finally, the candidate variables for Model 3 (the "refined full model") included the Model 1 variables plus ICD-9 code, study sites, more detailed categorization of CPT code and comorbidities.

Constructing the above three sequential models helped us to examine the tradeoff between performance and broad applicability. First, we aimed to build a more broadly applicable prediction model to be applied to LESS-1 (Model 1). Second, we were also interested in building a less broadly applicable model but using more variables provided in LESS-2, for which we developed Model 2 and Model 3. We then compared the prediction performance of these three models on the same subset LESS-2.

We first split the whole BOLD cohort into a training set and a testing set with a 4:1 ratio. Then, we did a 5-fold cross validation on the training set to choose the best model in this step, based on the highest R^2. We then evaluated the performance of the selected models on the testing set in BOLD and also on the subset of LESS data in which patients received the control treatment in order to understand the performance of the three models across the BOLD and LESS samples. We tested Model 1 on both LESS-1 and LESS-2. For Model 2 and Model 3, we tested them only on LESS-2.

Table 1. Three sequential models. Tables include variables, data source, and the number of patients in LESS with control treatment for testing.

Model	Variables	Data Source	Testing in LESS with control treatment
1. Baseline Model	age, sex, hispanic, education, marital, employment, lawyer, smoking status, falls, bpi, painexpect3mo, RMDQ at baseline, NRS back pain at baseline, NRS leg pain at baseline, PHQ-4, BMI, EQ_VAS_baseline, EQ_index_baseline	PRO	LESS-1
			LESS-2

2. Full Model	Model 1 variables + study_site, ICD-9 categories, comorbidity categories, CPT categories (manual, spine-image, percut, spine-surgery)	PRO+EHR	LESS-2
3. Refined Full Model	Model 1 variables + study_site, ICD-9 categories, refined* comorbidity categories, refined* CPT categories	PRO+EHR	LESS-2

*refined: sub-categories

In order to evaluate model prediction for the continuous outcomes, we calculated how well the models explain variations in outcomes using the coefficients of determination (R^2) and mean squared error (MSE). Here the R^2 was calculated using the square of Pearson correlation (R). Since one of our main purposes was to derive a risk score and use it to stratify patients by quartiles (rank of risk score) for further treatment effect modeling, we decided to calculate R^2 using the square of Pearson correlation and treated it as our primary criterion for choosing the model. Scatter plots comparing actual RMDQ with predicted RMDQ were also investigated to evaluate modeling performance. Our primary selection of models was based on the R^2 in the LESS data with the one having the highest R^2.

2.3.2 Aim2: Treatment Effect Modeling to Identify HTE

After selecting the model in Aim 1, we applied it in the LESS trial to generate risk scores. We first divided patients into 4 strata using 25%, 50% and 75% quartiles based on their risk scores. Then we evaluated the treatment effects within each strata using the model below.

$$6-week\ RMDQ = Risk\ strata + Treatment\ group + Risk\ strata\ \times Treatment\ group\ + Baseline\ RMDQ\ + Study\ sites$$

We estimated absolute treatment effects for 6-week RMDQ respectively to identify treatment effects. 95% confidence interval and P-value were reported for each stratum. We conducted the Wald test using the robust Huber Sandwich Estimator to evaluate the statistical significance of the interaction term between risk strata and treatment. Such strata-based analysis was for reporting purposes.

$$6-week\ RMDQ = Risk\ score + Treatment\ group + Risk\ score\ \times Treatment\ group\ + Baseline\ RMDQ\ + Study\ sites$$

Our primary analysis was to evaluate whether the association between treatment/control and outcome was the same for patients with different risk scores. We used the linear regression below with an interaction term between the continuous risk scores and treatment.We also conducted a Wald test to evaluate the statistical significance of the interaction term between the linear predictor of risk (continuous risk scores) and treatment. A similar process was used to model 6-week NRS scores.

3. Results

3.1 Descriptive statistics of different cohort

Baseline measures of 5243 patients were collected in the BOLD cohort and 400 patients in the LESS trial. After excluding any missingness and patients with baseline RMDQ ≤ 2 in the BOLD cohort, 3318 participants were included in analyses to develop and validate a predictive model for 3-month RMDQ that could provide a risk score for back-related functional limitations reflected by the 3-month RMDQ. When it comes to 3-month back NRS, 3480 patients in the BOLD are adopted. Table 2 summarizes the baseline descriptive statistics of demographic, disease diagnosis and CPT category variables for both the BOLD training set, BOLD testing set and LESS trial. Compared with patients in the BOLD cohort, patients in LESS trial were slightly younger, had higher BMI, and more likely to be non-working, non-married, smokers, and have more comorbidities. For diagnostic- and treatment-related variables, most patients in the LESS trial had spine-related manual care, spine-related percutaneous care and spine-related image, but no one had spine-related surgery.

Table 3 summarizes baseline PROs by cohort. Compared with patients in the BOLD cohort, patients in LESS had higher values for RMDQ scores, back and leg pain of NRS scores, recovery expectations, PHQ_4 and BPI scores. Besides, they had lower EQ_VAS_baseline scores and EQ_index_baseline scores.

In general, there is much more severe disability and pain in the LESS trial patients than in the BOLD cohort. There are also differences in the procedure, diagnosis as well as comorbidities between BOLD cohortand LESS patients, but otherwise, the differences between those groups are generally small.

Table 2. Demographic variables by cohort

	BOLD_Train (N=4580)	BOLD_Test (N=663)	LESS (N=400)
Age			
Mean (SD)	73.7 (6.80)	73.5 (6.71)	68.0 (9.98)

Median [Min, Max]	72.0 [65.0, 98.0]	72.0 [65.0, 101]	68.0 [50.0, 96.0]
Missing	507 (11.1%)	84 (12.7%)	0 (0%)
Sex			
Female	2665 (58.2%)	354 (53.4%)	221 (55.2%)
Male	1408 (30.7%)	225 (33.9%)	179 (44.8%)
Missing	507 (11.1%)	84 (12.7%)	0 (0%)
Body Mass Index			
Mean (SD)	29.1 (6.15)	28.9 (5.89)	30.4 (6.28)
Median [Min, Max]	28.3 [14.6, 64.2]	28.0 [17.1, 59.1]	29.3 [18.7, 61.5]
Missing	181 (4.0%)	28 (4.2%)	12 (3.0%)
Hispanic			
No	3793 (82.8%)	547 (82.5%)	378 (94.5%)
Yes	260 (5.7%)	30 (4.5%)	17 (4.2%)
Missing	527 (11.5%)	86 (13.0%)	5 (1.2%)
Race			
Black or African American	617 (13.5%)	99 (14.9%)	105 (26.2%)
White	2966 (64.8%)	406 (61.2%)	277 (69.2%)
Other	448 (9.8%)	65 (9.8%)	18 (4.5%)
Missing	549 (12.0%)	93 (14.0%)	0 (0%)
Working status			
No	226 (4.9%)	43 (6.5%)	93 (23.2%)
Working	220 (4.8%)	27 (4.1%)	35 (8.8%)
Missing	4134 (90.3%)	593 (89.4%)	272 (68.0%)
Education			
High School or less	1242 (27.1%)	172 (25.9%)	125 (31.2%)
Some college	1223 (26.7%)	169 (25.5%)	128 (32.0%)
Four year college graduate or more	1599 (34.9%)	235 (35.4%)	145 (36.2%)

Missing	516 (11.3%)	87 (13.1%)	2 (0.5%)
Marital			
No	1672 (36.5%)	219 (33.0%)	162 (40.5%)
Married/Living with partner	2389 (52.2%)	358 (54.0%)	237 (59.2%)
Missing	519 (11.3%)	86 (13.0%)	1 (0.2%)
Lawyer			
No	4040 (88.2%)	573 (86.4%)	380 (95.0%)
Yes	24 (0.5%)	4 (0.6%)	13 (3.2%)
Missing	516 (11.3%)	86 (13.0%)	7 (1.8%)
Smoking status			
Never Smoked	2237 (48.8%)	320 (48.3%)	172 (43.0%)
Quit smoking over a year ago	1573 (34.3%)	224 (33.8%)	169 (42.2%)
Current smoker, or quit less than a year ago	251 (5.5%)	33 (5.0%)	57 (14.2%)
Missing	519 (11.3%)	86 (13.0%)	2 (0.5%)
Quan comorbidity			
0	2703 (59.0%)	392 (59.1%)	92 (23.0%)
1	776 (16.9%)	108 (16.3%)	42 (10.5%)
2 or more	991 (21.6%)	143 (21.6%)	87 (21.8%)
Missing	110 (2.4%)	20 (3.0%)	179 (44.8%)
Baseline diagnosis category			
Back pain only	2765 (60.4%)	385 (58.1%)	8 (2.0%)
Back and leg pain	881 (19.2%)	128 (19.3%)	29 (7.2%)
Spinal stenosis	201 (4.4%)	27 (4.1%)	102 (25.5%)
Others	226 (4.9%)	39 (5.9%)	70 (17.5%)
Missing	507 (11.1%)	84 (12.7%)	191 (47.8%)
Manual spine-related CPT			
No	3969 (86.7%)	578 (87.2%)	14 (3.5%)
Yes	566 (12.4%)	77 (11.6%)	207 (51.8%)

Missing	45 (1.0%)	8 (1.2%)	179 (44.8%)
Percutaneous spine-related CPT			
No	4478 (97.8%)	646 (97.4%)	10 (2.5%)
Yes	57 (1.2%)	9 (1.4%)	211 (52.8%)
Missing	45 (1.0%)	8 (1.2%)	179 (44.8%)
Spine image-related CPT			
No	3256 (71.1%)	492 (74.2%)	2 (0.5%)
Yes	1279 (27.9%)	163 (24.6%)	219 (54.8%)
Missing	45 (1.0%)	8 (1.2%)	179 (44.8%)
Spine surgery-related CPT			
No	4517 (98.6%)	654 (98.6%)	221 (55.2%)
Yes	18 (0.4%)	1 (0.2%)	0 (0%)
Missing	45 (1.0%)	8 (1.2%)	179 (44.8%)

SD, standard deviations; CPT: Current Procedural Terminology codes.

Table 3. PROs and interventions by cohort

	BOLD_Train (N=4580)	**BOLD_Test (N=663)**	**LESS (N=400)**
RMDQ score(baseline)			
Mean (SD)	9.83 (6.31)	9.63 (6.47)	15.8 (4.37)
Median [Min, Max]	10.0 [0, 24.0]	9.00 [0, 23.0]	17.0 [5.00, 24.0]
Missing	507 (11.1%)	84 (12.7%)	0 (0%)
NRS back pain(baseline)			
Mean (SD)	5.06 (2.79)	5.02 (2.78)	6.66 (2.47)
Median [Min, Max]	5.00 [0, 10.0]	5.00 [0, 10.0]	7.00 [0, 10.0]
Missing	507 (11.1%)	84 (12.7%)	0 (0%)
NRS leg pain(baseline)			
Mean (SD)	3.53 (3.30)	3.42 (3.37)	7.20 (1.83)
Median [Min, Max]	3.00 [0, 10.0]	3.00 [0, 10.0]	7.00 [1.00, 10.0]
Missing	507 (11.1%)	85 (12.8%)	0 (0%)

How confident is the patient that the back of leg pain will be completely gone or will be much better 3 month from now?			
Mean (SD)	5.49 (3.71)	5.42 (3.58)	7.71 (1.85)
Median [Min, Max]	5.00 [0, 10.0]	5.00 [0, 10.0]	8.00 [2.00, 10.0]
Missing	516 (11.3%)	86 (13.0%)	7 (1.8%)
Patient Health Questionnaire 4(baseline)			
Mean (SD)	1.65 (2.51)	1.58 (2.46)	3.11 (3.04)
Median [Min, Max]	0 [0, 12.0]	0 [0, 12.0]	2.00 [0, 12.0]
Missing	522 (11.4%)	85 (12.8%)	15 (3.8%)
Times of fallen in past 3 weeks(baseline)			
No	3756 (82.0%)	538 (81.1%)	356 (89.0%)
One or more falls in the past e wk	316 (6.9%)	39 (5.9%)	42 (10.5%)
Missing	508 (11.1%)	86 (13.0%)	2 (0.5%)
Brief Pain Inventory score(baseline)			
Mean (SD)	3.39 (2.46)	3.30 (2.52)	6.10 (2.29)
Median [Min, Max]	3.14 [0, 10.0]	3.00 [0, 9.71]	6.50 [0, 10.0]
Missing	507 (11.1%)	84 (12.7%)	1 (0.2%)
EQ5D-VAS 0-100(baseline)			
Mean (SD)	74.4 (18.4)	73.6 (18.7)	66.4 (20.8)
Median [Min, Max]	80.0 [0, 100]	80.0 [0, 100]	70.0 [3.00, 100]
Missing	6 (0.1%)	3 (0.5%)	11 (2.8%)
EQ5D-Index 0-1(baseline)			
Mean (SD)	0.756 (0.174)	0.756 (0.183)	0.578 (0.203)
Median [Min, Max]	0.778 [-0.109, 1.00]	0.778 [-0.0384, 1.00]	0.597 [0.0494, 1.00]
Missing	11 (0.2%)	4 (0.6%)	1 (0.2%)

*SD, standard deviation; NRS, numerical rating scale; PHQ-4, Patient Health Questionnaire 4, question version asking about depression and anxiety; EQ5D-Index, European Quality of Life 5 Dimension Index;

EQ5D-VAS, European Quality of Life 5 Dimension Visual Analog Scale; PRO, patient reported outcome; RMDQ, Roland-Morris Disability Questionnaire.

3.2 Modeling for Aim 1: Develop and validate a predictive model in BOLD

3.2.1 Primary model: prediction of 3-month RMDQ (continuous variable)

Model 1 testing on LESS-1 containing 171 patients who had PRO variables as well as LESS-2 containing 94 patients who have both EHR and PRO had the highest R^2 and lowest MSE (**Fig. 3**). To be specific, the R^2 was 34% using the model 1 on LESS-1 and 22% on LESS -2, 21% using the model 2, and 19% using the model 3 (**Fig. 3**), indicating that the additional EHR variables including disease diagnosis and interventions did not provide additional prognostic information. Based on the result, we chose model 1 as the appropriate predictive model. Using the training dataset, this LASSO baseline model selected 21 out of 27 possible variables. The model excluded the variables hispanic ethnicity, having had a college education, marital status, having a lawyer involved in current pain or health, and having quit smoking over a year ago.

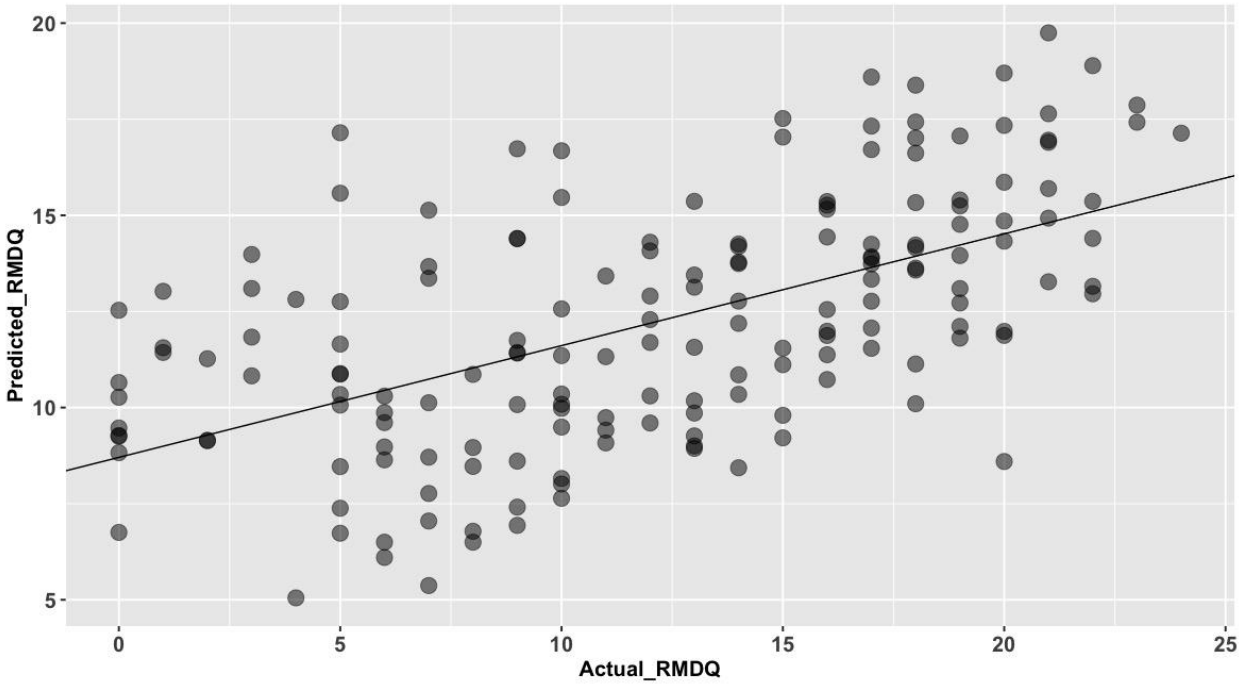

Fig. 2. Scatter plot comparing actual RMDQ versus predicted RMDQ generated by Model 1 among LESS patients with control treatment.

	MSE				*R^2*			
Models compared	BOLD_train	BOLD_test	LESS-1	LESS-2	BOLD_train	BOLD_test	LESS-1	LESS-2
1. Baseline Model	24.46	22.58	26.66	30.08	0.39	0.43	0.34	0.22
2. Full Model	23.68	24.61	/	33.64	0.41	0.37	/	0.21
3. Refined Full Model	23.48	24.76	/	33.05	0.41	0.36	/	0.19

Fig. 3. Prediction performance of continuous 3-month RMDQ as the outcome for each model. MSE and R^2 were calculated for the BOLD training set, BOLD testing set and LESS testing set.

3.2.2 Secondary models: Prediction of 3-month back pain NRS (continuous variable)

	MSE				*R^2*			
Models compared	BOLD_train	BOLD_test	LESS-1	LESS-2	BOLD_train	BOLD_test	LESS-1	LESS-2
1. Baseline Model	5.02	5.30	6.22	7.82	0.29	0.28	0.13	0.07
2. Full Model	5.06	4.97	/	8.82	0.29	0.29	/	0.08
3. Refined Full Model	5.06	4.95	/	8.18	0.29	0.30	/	0.07

Fig. 4. Prediction performance of continuous 3-month back pain NRS scores as the outcome for each model. MSE and R^2 were calculated for the BOLD training set, BOLD testing set and LESS testing set.

For continuous NRS back pain score, three sequential models all had similar performance on both BOLD sets and LESS testing sets. The R^2 was 13% using the model 1 on LESS-1 and 7% on LESS-2, 8% using the model 2, and 7% using the model 3 (**Fig. 4**). We can conclude that Model 2 had highest R^2 while Model 1 had lowest MSE among patients with control treatment and within LESS-2. However, neither of these three models provided sufficient predictive power for further evaluation of HTE in the LESS trial. Therefore, we decided not to use continuous back pain score as the endpoint in Aim 1.

3.3 Modeling for Aim 2: Treatment Effect Modeling to Identify HTE in LESS

We divided patients into four quartiles based on the risk scores generated from analysis of BOLD. For each risk strata from lowest quantile to the highest quantile, the range of their risk scores is (4.26, 10.22), (10.23, 12.79), (12.81, 15.19), (15.20, 20.11) (**Fig. 5**).

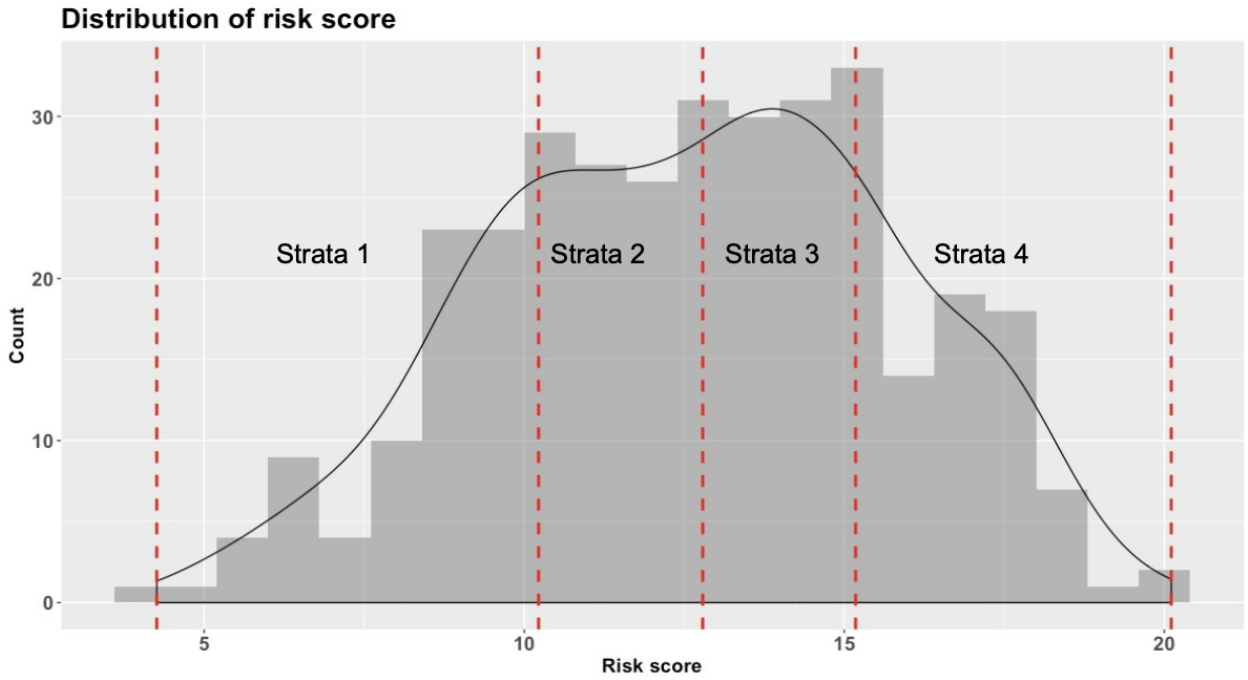

Fig. 5. Plot of distribution of risk scores four risk strata divided by quartiles. The red dotted line represents the range of each quartile.

Fig. 6 below showed the heterogeneity of epidural glucocorticoid injections effect among 4 risk strata for 6-week RMDQ and 6-week back pain NRS respectively. For 6-week continuous RMDQ, there was a decreasing trend from stratum 1 to stratum 3 indicating patients with higher predicted risk received more benefit from the epidural glucocorticoid injections. The treatment benefit was generally similar (although slightly smaller) in the stratum 4 (average treatment effect, -2.10 points; 95%CI, -4.42 to 0.22). We also identified that patients receiving epidural glucocorticoid injections within strata 3 had significant treatment benefit (average treatment effect, -2.36 points; 95%CI, -4.58 to -0.14). According to the Wald test with robust standard error, there was no statistically significant difference (p-value=0.19 > 0.05) of treatment effect across each strata (**Table 5**).

Table 4. Results of the strata analysis performed for continuous outcomes

Selected variables in aim2	RMDQ (continuous)	NRS (continuous)
tx=1	0.51 (-1.55, 2.57)	0.27 (-0.73, 1.27)
tx*risk2	-0.81 (-3.89, 2.26)	-0.39 (-1.82, 1.04)
tx*risk3	-2.87 (-5.91, 0.17)	-0.20 (-1.74, 1.34)
tx*risk4	-2.61 (-5.70, 0.48)	-1.19 (-2.73, 0.35)

Table 5. Modeling results of the strata analysis performed for continuous outcomes. Point estimate and 95% confidence interval were shown in the table for each covariates. The P-value testing significance of interaction terms was calculated using Wald test with robust Huber Sandwich Estimator for RMDQ and back pain NRS respectively.

For 6-week continuous NRS, the treatment effect also showed a decreasing trend from low risk group to high risk group, with stratum 4 having the most treatment benefit (average treatment effect, -0.92 points; 95%CI, -2.10 to 0.25). The P-value 0.48 (> 0.05) calculated using Wald test with robust standard error showed that there was no statistically significant difference of treatment effect across each strata when using 6-week back pain NRS as endpoints .

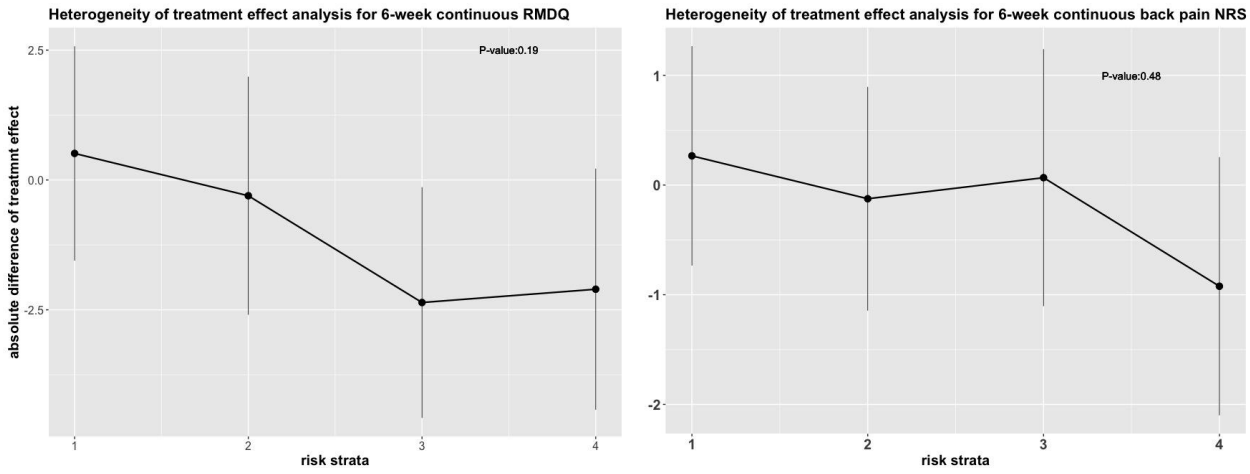

Fig. 6. Heterogeneity of epidural glucocorticoid injections effect among 4 risk strata. Point estimate and 95% confidence interval were shown in the plot. The P-value testing significance of interaction terms was calculated using Wald test with robust Huber Sandwich Estimator for RMDQ and back pain NRS respectively.

For our primary analysis, we evaluated the risk scores as continuous effect modifiers for the association between the 6-week RMDQ and treatment groups. As risk scores increased, the magnitude of estimated coefficient of treatment effects decreased linearly (**Fig. 7**), showing better treatment effects for patients with higher risk scores. However, there was no significance for the interaction term (P-value = 0.11) between treatment and continuous risk scores (**Table 5**). As for 6-week back NRS, the interaction term (P-value = 0.53) between treatment and continuous risk scores was also not statistically significant.

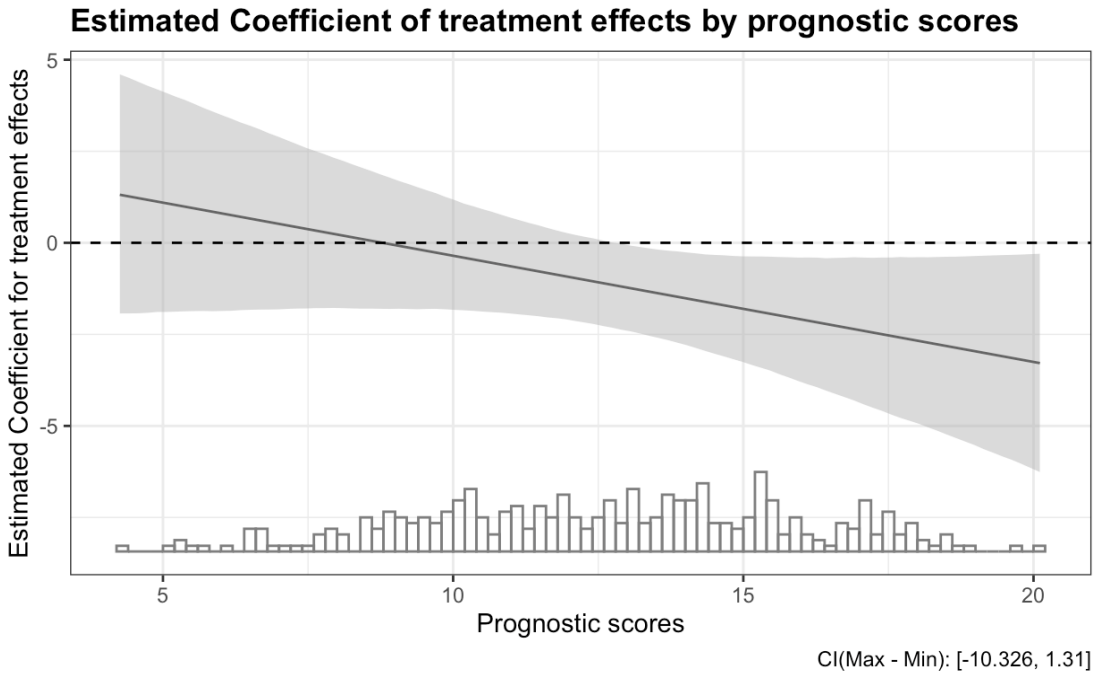

Fig. 7. Association of risk scores (along the x-axis) and the magnitude of the estimated coefficient of treatment effects by risk scores, with the distribution of risk scores at bottom. The caption at the right bottom corner shows the confidence intervals of the conditional effects corresponding to the grey area in the plot.

Table 5. Results of the overall HTE analysis for continuous RMDQ

Variable	Estimate (95% CI)	P-value
tx	2.58 (-2.08, 7.25)	0.28
risk scores	0.85 (0.42, 1.29)	0.00015 ***
Baseline RMDQ	0.33 (0.04, 0.61)	0.026 *
Site 2	2.31 (0.55, 4.06)	0.011 *
Site 3	0.76 (-0.83, 2.34)	0.35
Site 4	1.87 (0.33, 3.4)	0.018 *
tx* risk scores	-0.29 (-0.65, 0.064)	0.11

Table 5. Modeling results of the linear regression including continuous risk scores. Point estimate, 95% confidence interval and P-value were shown in the table for each covariates.

4. Discussion

To sum up, the risk modeling method based on the PATH statement showed that it was a promising method to evaluate heterogeneity of treatment effect. By applying an external model generated from the BOLD cohort, we observed a decreasing trend from low risk patients to high risk patients, although this was not statistically significant. We also identified one subgroup of patients receiving epidural glucocorticoid injections had significant treatment benefit. Moreover, while baseline patient characteristics were consistently valuable predictors of back disability and pain 3 months later, medical interventions and disease diagnosis generally didn't provide additional information.

We used a relatively novel approach PATH to evaluate HTE. PATH has two advantages compared with usual subgroup analysis. Firstly, PATH avoids low statistical power and multiplicity in "one-variable-at-a-time" analyses and provides patient-centered estimates, taking into account all related patient characteristics simultaneously. Secondly, PATH is feasible for analyzing data from different cohorts, by using external models. While there was no statistically significant difference of treatment effect across the four risk strata as well as the continuous risk score, the result was still informative in that we observed significant treatment effects in stratum 3. In the lower risk group, the epidural glucocorticoid injections did not improve RMDQ or back NRS though the treatment effectiveness increased with increasing risk stratum.

Limitations to our study include the following:

1. For Aim 1, the outcome for developing the risk score was 3-month RMDQ. However, in Aim 2, we chose the 6-week RMDQ as the outcome. There was a time difference between those two outcomes because we lacked the 6-week RMDQ data in the BOLD cohort. Therefore, we used the nearest RMDQ measurement, the 3-month RMDQs.
2. When we developed the model for the risk score, we only included the linear term instead of adding any interaction term or splines. We later performed sensitivity analyses by adding interaction terms. The resulting R^2 did not improve. Adding splines in the model made the results even worse. Therefore, we used only the variables in the linear term.
3. Variables were not completely consistent in these two cohorts. We only used the variables present in both cohorts to build the models. Since not all of the LESS cohort had EHR data, by including EHR in the model, we lost lots of LESS patients. Therefore, there was a trade-off between observations and variables that we could include in the models.
4. Some of the variables that we used were derived from previous studies. For example, the diagnosis of patients was derived from ICD-9 codes. Here, we categorized them into 4 groups based on Jerry,etc. paper in 2018 . We did the same for CPT codes, mean BPI, etc.

The different ways of categorization will impact the results. We used more detailed categorization for CPT code and ICD-9 code in model 3.

5. There were differences between BOLD and LESS cohorts, such as the degree of disability and pain measured. The baseline RMDQ and NRS distribution are different which is one explanation of why the models built on one cohort performed less well for the other one.
6. Model selection was based on results of testing on LESS, meaning that we used LESS data twice. However, it makes sense that in the clinical setting, our final goal is to develop an accurate risk score for the LESS cohort in order to have a convincing result for the HTE.

5. References

[1] Harrast MA. Epidural steroid injections for lumbar spinal stenosis. Curr Rev Musculoskel Med 2008;1:32-38

[2] Deyo RA, Mirza SK, Martin BI, Kreuter W, Goodman DC, Jarvik JG. Trends, major medical complications, and charges associated with surgery for lumbar spinal stenosis in older adults. JAMA 2010;303:1259-1265

[3] Daffner SD, Wang JC. The pathophysiology and nonsurgical treatment of lumbar spinal stenosis. Instr Course Lect 2009;58:657-668

[4] Englund J. Lumbar spinal stenosis. Curr Sports Med Rep 2007;6:50-55

[5] Friedly, Janna L., et al. "A randomized trial of epidural glucocorticoid injections for spinal stenosis." New England Journal of Medicine 371.1 (2014): 11-21.

[6] Kent, David M., et al. "The predictive approaches to treatment effect heterogeneity (PATH) statement." Annals of internal medicine 172.1 (2020): 35-45.

[7] 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

[8] 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

[9] 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).

[10] JarvikJG, Gold LS, Tan K, et al. Long-term outcomes of a large, prospective observational cohort of older adults with back pain.

[11] Jarvik JG, Comstock BA, Bresnahan BW, et al. Study protocol: the Back pain Outcomes using Longitudinal Data (BOLD) Registry. BMC Musculoskelet Disord 2012;13:64.

[12] Roland M, Morris R. A study of the natural history of back pain. Part 1: development of a reliable and sensitive measure of disability in low back pain. Spine 1983;8:141–4.

[13] Cleeland CS, Nakamura Y, Mendoza TR, Edwards KR, Douglas J, Serlin RC. Dimensions of the impact of cancer pain in a four country sample: new information from multidimensional scaling. Pain 1996;67:267–73.
[14] Quan H, Li B, Couris CM, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. Am J Epidemiol 2011;173:676– 82.
[15] StatisticsNCfH.InternationalClassificationofDiseases,NinthRevision (ICD-9). 2009. Available at: http://www.cdc.gov/nchs/icd/icd9.htm. Accessed February 12, 2018.
[16] AAMA (Internet). AMA—About CPT®. c1995-2013. Available at: http://www.ama-assn.org/ama/pub/physician-resources/solutions -managing-your-practice/coding-billing-insurance/cpt/about-cpt .page. Accessed February 12, 2018.
[17] Brooks R. EuroQOL: the current state of play. Health Policy (New York) 1996;37:53–72.
[18] Cleeland CS, Ryan KM: Pain assessment: global use of the Brief Pain Inventory. Ann Acad Med Singapore 1994, 23(2):129–138.
[19] Centers for Disease Control and Prevention (CDC). Self-reported falls and fall-related injuries among persons aged > or =65 years—United States, 2006. MMWR Morb Mortal Wkly Rep 2008;57:225–9.
[20] Kroenke K, Spitzer RL, Williams JB, Lowe B. An ultra-brief screening scale for anxiety and depression: the PHQ-4. Psychosomatics 2009;50:613–21.
[21] Iles RA, Davidson M, Taylor NF, O'Halloran P: Systematic review of the ability of recovery expectations to predict outcomes in non-chronic non-specific low back pain. J Occup Rehabil 2009, 19:25–40.
[22] Kongsted A, Vach W, Axo M, Bech RN, Hestbaek L: Expectation of recovery from low back pain: a longitudinal cohort study investigating patient characteristics related to expectations and the association between expectations and 3-month outcome. Spine (Phila Pa 1976) 2014, 39:81–90.
[23] Deyo RA, Bryan M, Comstock BA, et al. Trajectories of symptoms and function in older adults with low back disorders. Spine 2015;40:1352–62.

6. Appendix

6.1 Sensitive analysis

6.1.1 Generalized Additive Model (GAM) analysis for aim2

The results of Aim 2 in 3.3, especially the plot of estimated absolute treatment effects, suggest that there may be a non-linear association between each risk strata and the treatment effects,due to the non-monotonous trend. In order to explore whether there was a nonlinear relationship, we conducted GAM analysis in which the linear response variable depends linearly on unknown

smooth functions of some predictor variables, and interest focuses on inference about these smooth functions.

We fit a smooth function on risk scores grouped by treatment, and treated the rest categorical or binary variables as linear effects. We aimed to test the null hypothesis that whether there exists a linear trend between 6-week RMDQ scores and risk scores grouped by treatment. However, p-value ($p = 0.86$) of the smooth term showed no scientific significance.

Also, plots of GAM analysis also supported the results. For the smooth term, which was the risk scores grouped by treatment (**Fig.8**), the line was approximately u-shape, however, with too wide 95% confidence interval. Therefore, we should not reject that there was a flat line easily. And the GAM fit in Figure 9 was consistent with the conclusion drawn from the above linear interaction model in the lower range of risk scores, which were (4.26, 10.22) and (10.23, 12.79). But also shows the effect leveling off for higher risk scores. This is also consistent with the top left plot in Figure 10. Therefore, the above analysis showed that there was no non-linear relationship between the 6-week RMDQ scores and treatment groups among each risk strata, and the linear model fit in Aim 2 was appropriate.

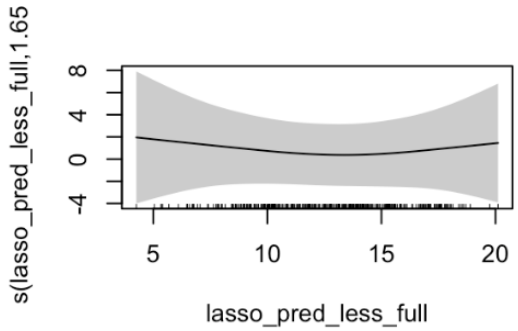

Fig. 8. Association of risk scores (along the x axis) and the magnitude of the estimated coefficient of treatment effects by risk scores, with the distribution of risk scores at bottom. The caption at the right bottom corner shows the confidence intervals of the conditional effects corresponding to the grey area in the plot.

6.1.2 Exploratory analysis of interaction terms in aim1

InAim 1 , we evaluated three sequential LASSO models and determined that model 1 that only used variables from PRO data was optimal. However, Model 1 did not consider non-linear relationships between variables so in the sensitivity analysis we also explored introducing interaction terms.

We believed that the baseline RMDQ score might interact with demographics variables such as age and sex, as well as patient reported outcomes such as NRS leg pain. We created interactions

in our LASSO model by multiplying the baseline RMDQ with other covariates. Then we fit a linear regression model among selected variables and evaluated whether interaction terms were significant by evaluating 95% confidence intervals and P-values (**appendix**). Interaction terms between EQ5D-index and baseline RMDQ (P-value 0.03) indicated that patients who had higher risk with higher baseline RMDQ will tend to have less improvement. Thus, we decided to further explore whether this interaction term could improve the prediction performance in Aim 1 and had a statistical significance in HTE analysis in Aim 2. The MSE and R^2 in Fig.13 shows that the modeling performance was almost the same between this interaction model and the model we selected in Aim 1 .

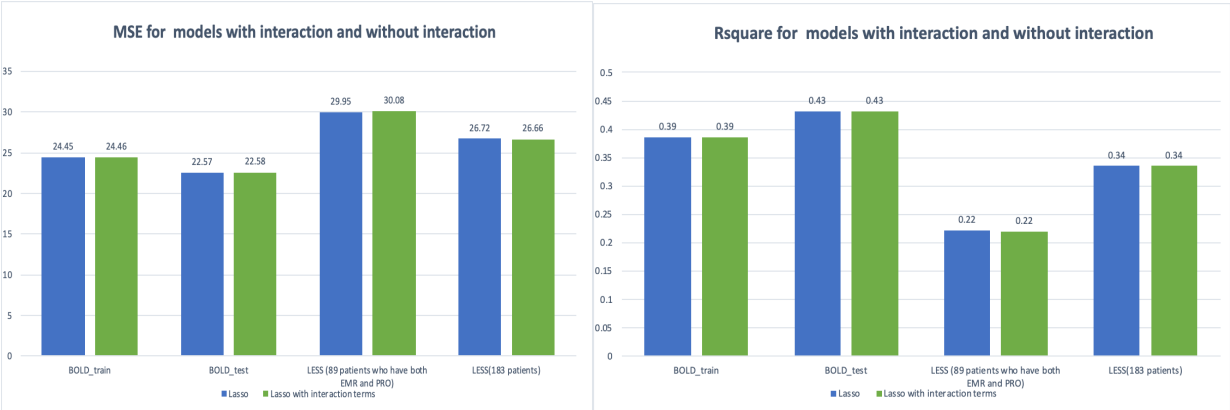

Fig. 9. Prediction performance of 3-month 30% back pain NRS scores improvement as the outcome for each model. MSE and R^2 were calculated for the BOLD training set, BOLD testing set and LESS testing set.

Fig.10 describes the heterogeneity of epidural glucocorticoid injections effectiveness among 4 risk strata for 6-week RMDQ. The monotonic decreasing trend from Strata 1 to 3 with a slight reverse in Stratum 4 was quite similar compared with the model we selected in Aim 1.The 95% confidence interval indicated that patients who received epidural glucocorticoid injections in Stratum 3 had significantly greater benefit compared with those in the control group. The Wald test of 3 degrees of freedom with robust standard error still showed there was no statistically significant difference (p-value=0.12 > 0.05) of treatment effect across each strata.

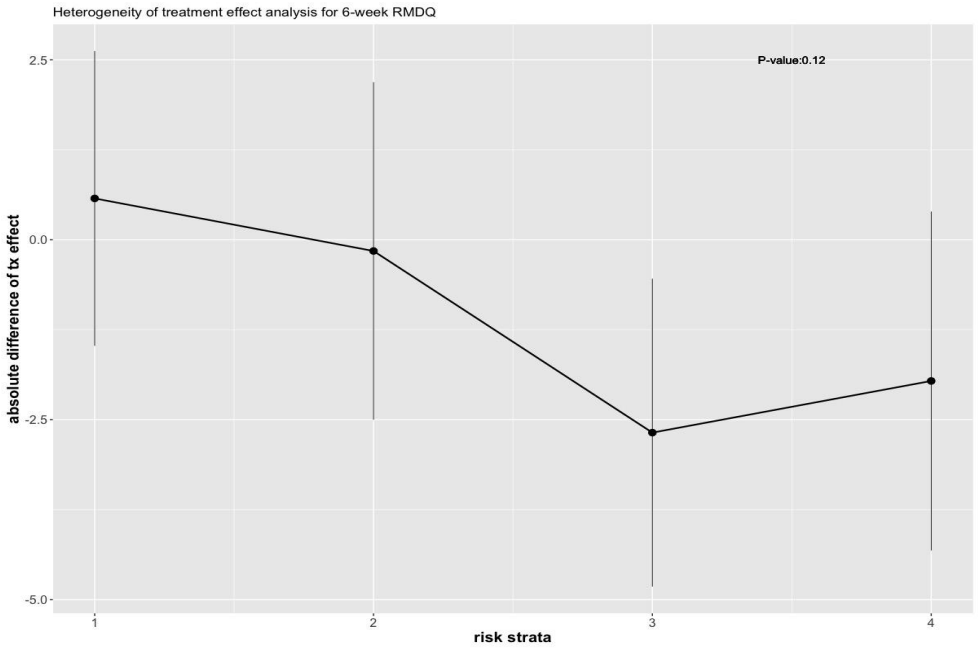

Fig. 10. Heterogeneity of epidural glucocorticoid injections effect among 4 risk strata using a model with interaction terms between baseline RMDQ and EQ5D scores(range from 0 to 1). Point estimate and 95% confidence interval were shown in the plot. The P-value 0.18 was calculated using Wald test with robust Huber Sandwich Estimator

6.3 Supplementary figures

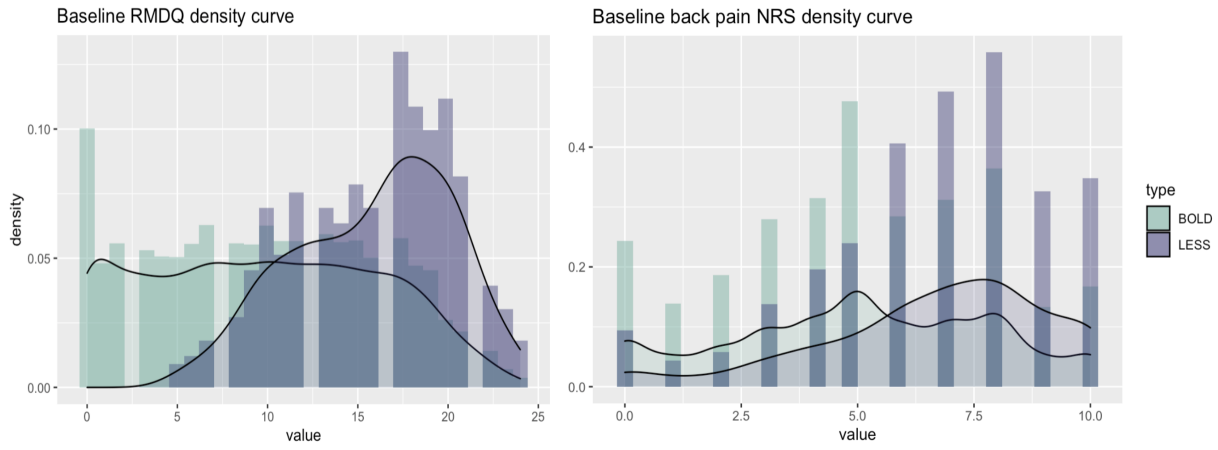

Fig. 11. Frequency density plot of baseline RMDQ and back pain NRS scores for the BOLD and LESS samples. Green bars represented patients within the BOLD cohort while purple bars represented patients within the LESS trial. Compared with the BOLD cohort, the LESS trial had fewer patients with low back pain and no patients with low functional limitation and disability.