Clinical/Basic Science Research Article

OPEN

Artificial neural networks outperform linear regression in estimating 9-month patient-reported outcomes after upper extremity fractures with increasing number of variables

Niels Brinkman, BS[a,*], Romil Shah, MD[a], Job Doornberg, MD, PhD[b,c], David Ring, MD, PhD[a], Stephen Gwilym, DPhil, FRCS(Orth)[d], Prakash Jayakumar, MD, PhD[a]

Abstract

Objective: To compare performance between linear regression (LR) and artificial neural network (ANN) models in estimating 9-month patient-reported outcomes (PROs) after upper extremity fractures using various subsets of early mental, social, and physical health variables.

Methods: We studied 734 patients with isolated shoulder, elbow, or wrist fracture who completed demographics, mental and social health measures, and PROs at baseline, 2–4 weeks, and 6–9 months postinjury. PROs included 3 measures of capability (QuickDASH, PROMIS-UE-PF, PROMIS-PI) and one of pain intensity. We developed ANN and LR models with various selections of variables (20, 23, 29, 34, and 54) to estimate 9-month PROs using a training subset (70%) and internally validated them using another subset (15%). We assessed the accuracy of the estimated value being within one MCID of the actual 9-month PRO value in a test subset (15%).

Results: ANNs outperformed LR in estimating 9-month outcomes in all models except the 20-variable model for capability measures and 20-variable and 23-variable models for pain intensity. The accuracy of ANN versus LR in the primary model (29-variable) was 83% versus 73% (Quick-DASH), 68% versus 65% (PROMIS-UE-PF), 66% versus 62% (PROMIS-PI), and 78% versus 65% (pain intensity). Mental and social health factors contributed most to the estimations.

Conclusion: ANNs outperform LR in estimating 9-month PROs, particularly with a larger number of variables. Given the otherwise relatively comparable performance, aspects such as practicality of collecting greater sets of variables, nonparametric distribution, and presence of nonlinear correlations should be considered when deciding between these statistical methods.

Keywords: patient-reported outcomes, capability, pain intensity, pain interference, psychosocial, estimating, prediction modeling, artificial neural network, machine learning, linear regression, model performance, increasing number of variables, comparison

1. Introduction

1.1. Background

Evidence suggests that psychosocial factors measured shortly after injury are strongly associated with levels of comfort and capability during recovery from injury and that comfort and capability have a surprisingly limited association with pathophysiology severity.[1–4] There is growing interest in sophisticated statistical models based on machine learning (ML), including artificial neural networks (ANNs), that might have the potential to provide more accurate estimations of patient outcomes and recovery trajectories during musculoskeletal care. Such estimates could help improve health by informing care strategies that address modifiable factors such as unhelpful thinking and feelings of distress early on after injury. Potential advantages of

R. Shah reports other from NSite; outside the submitted work. D. Ring reports other from Skeletal Dynamics, personal fees from Deputy Editor for Clinical Orthopaedics and Related Research, *personal fees from Universities and Hospitals, personal fees from Lawyers, personal fees from Health Services and Resource Administration and Department of Justice, personal fees from Premier Healthcare Solutions, personal fees from Wolters Kluwer Health, grants from National Institutes for Health, personal fees from Everus, other from MyMedicalHub; outside the submitted work. P. Jayakumar reports grants from Bass Family Foundation, grants from Kozmetsky Family Foundation, grants from West Health, personal fees from Proximie Ltd, other from Optum Health, other from University of Minnesota, other from AO Foundation; outside the submitted work. The remaining authors have no conflicts of interest to disclose.*

[a] Department of Surgery and Perioperative Care, The University of Texas at Austin, Dell Medical School, Austin, TX, [b] Department of Orthopaedic & Trauma Surgery, University Medical Center, Groningen, the Netherlands, [c] Department of Orthopaedic Trauma, Flinders University, Adelaide, Australia., [d] The Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, The University of Oxford, Oxford, United Kingdom

* Corresponding author. Address: 1601, Trinity St The Health Discovery Building, Building B, Austin, TX 78712. E-mail address: niels.brinkman@hotmail.com (N. Brinkman).

Source of funding: nil.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.otainternational.org).

OTAI (2023) e284

Received: 29 June 2023 / Accepted: 12 August 2023
Published online 22 December 2023

http://dx.doi.org/10.1097/OI9.0000000000000284

ANNs include greater suitability to nonparametric data[5] and nonlinear correlations[5–7] and better handling of colinearity among explanatory variables,[8] which is common with psychosocial factors.

1.2. Rationale

Despite these advantages, ANNs have not consistently outperformed standard logistic regression models in studies of musculoskeletal illness, and there remains a lack of clarity around their utility in research and clinical application.[9,10] A possible explanation is that ANNs might be most useful in data with high signal-to-noise ratio (SNR) (the proportion of the variation in an outcome that is explained by explanatory variables [R^2]),[11,12] whereas in human medical research, SNR is often relatively low. Another explanation could be the relatively high sample size needed to optimally train ML algorithms.[13] Some researchers claim that the advantage of ML might lie in better handling of large numbers of explanatory variables[13–16] and that ML is less likely to outperform regression models incorporating only a limited set of variables. However, the risk of model overfit may also increase when more variables are included. Model overfit is a phenomenon in which a model captures and "learns" from random noise and fluctuations in the training data to an extent that it negatively affects the performance of the model to new data.[17] Variable selection is a popular way to account for this limitation by only selecting a smaller subset of the most influential variables, thereby reducing the noise from redundant variables.[18] However, these statistical methods often select different variables for ML algorithms than regression models,[10] reducing their generalizability. Thoughtful preselection of potentially relevant variables might therefore be beneficial.[10]

The performance of ANNs compared with regression is less clear when the same variables or a larger number of potentially relevant variables, including psychosocial factors (eg, mindset and social circumstances) and early patient-reported outcomes (PROs) such as level of comfort and capability. We sought to better understand the performance of ANNs compared with linear regression (LR) in estimating longer-term PROs after arm fractures based on subsets of variables by addressing the following study questions: (1) Is there a difference in the performance between ANN and LR in estimating 9-month levels of capability, pain interference, and pain intensity after upper extremity fractures? (2) What is the difference in performance of ANN and LR models that incorporate different subsets of explanatory variables after upper extremity fracture?

2. Methods

2.1. Study design and participants

This is a secondary analysis of data from a previous research study involving the prospective enrollment of 775 adult patients with isolated shoulder, elbow, or wrist fracture at a single level 1 trauma center.[19] The Institutional Review Board approved the request to collect and analyze protected health information for research purposes (Integrated Research Application System [IRAS] Number 16/YH/0017) and specifically considered (1) the risk and anticipated benefits, (2) the selection of subjects, (3) the procedures for obtaining and documenting informed consent, (4) the safety of subjects, and (5) the privacy of subjects and confidentiality of the data. All methods were performed in accordance with relevant guidelines and regulations of our Institutional Review Board.

We analyzed a total of 734 patients from this data set who completed demographic data, clinical data, and PRO measures at baseline (within the first week after fracture), 2–4 weeks, and 6–9 months postinjury (Table 1). Only adults fluent in English were included, and those experiencing polytrauma, fracture dislocations, and periprosthetic fractures were excluded.

2.2. Measured variables and variable selection

We measured demographic information, clinical variables, mental health factors, social circumstances, PROs, and patient satisfaction at baseline and 2–4 weeks postinjury. We then created 5 subsets with various selections of an increasing number of explanatory variables (20, 23, 29, 34, and 54 variables) with the intention to compare ANN and LR models among different data subsets and

TABLE 1

Demographics

Variable	N (%)
All observations	734 (100)
Sex	
Men	236 (32)
Women	498 (68)
Marital status	
Divorced/widowed	203 (28)
Partner/married	365 (50)
Single	166 (23)
Social support (living situation)	
Alone	153 (21)
Partner/friend(s)/family	514 (70)
Full/part-time care	67 (9.1)
Work status	
Employed	326 (44)
Homemaker	328 (45)
Retired	46 (6.3)
Unemployed/workers' compensation	34 (4.6)
Fracture location	
Distal radius	378 (51)
Elbow	183 (25)
Proximal humerus	173 (24)
Broad injury class*	
Category 1	377 (51)
Category 2	131 (18)
Category 3	226 (31)
Injury on dominant side	363 (49)
High energy injury	119 (16)
Open injury	13 (1.8)
Neurovascular injury	38 (5.2)
Surgery	118 (16)
Prior fracture	
Dominant arm	137 (19)
Nondominant arm	79 (11)
Opioid use	252 (34)
Antidepressant use	174 (24)
	Mean ± SD
Age (years)	59 ± 20
Years of education	14 ± 3.0
	Median (IQR)
NRS pain intensity (6–9 months)	2 (4)
Quick-DASH (6–9 months)	16 (39)

IQR = interquartile range; SD = standard deviation.

* Injury class category 1 (greater tuberosity, radial neck/head, and OTA/AO type A distal radius fracture); category 2 (proximal humerus fractures with 2 or less fragments, extra-articular elbow fractures, OTA/AO type B distal radius fracture); category 3 (proximal humerus fractures with 3 or more fragments, intra-articular elbow fractures, OTA/AO type C distal radius fracture).

TABLE 2

Variables included in each data subset

	20-variable model	23-variable model	29-variable model	34-variable model	54-variable model
Variables included	Age Female CCI Marital status Social support Years of education Work status IMD factor Fracture location Dominant side High energy Neurovascular injury Open injury Surgery Prior fracture Prior fx dominant Prior fx nondom Prior fx ipsilateral Broad injury class Opioid use at 2 weeks	Every variable to the left plus: PROMIS Depression CAT (t = 0) PROMIS Anxiety CAT (t = 0) PCS-13 (t = 0)	Every variable to the left plus: PROMIS Depression CAT (t = 1) PROMIS Anxiety CAT (t = 1) PCS-13 (t = 1) TSK-11 (t = 0, t = 1) PROMIS IS	Every variable to the left plus: Complication Antidepressant use Grip injured side Grip uninjured side Grip injured/uninjured	Every variable to the left plus: Clinical satisfaction (t = 0, 1) Hospital satisfaction (t = 0, 1) PROMIS PF CAT (t = 0, 1) PROMIS UE CAT (t = 0, 1) PROMIS PI CAT (t = 0, 1) Quick-DASH (t = 0, 1) EQ-5D-3L (t = 0, 1) NPRS (t = 0, 1) PSEQ-2 (t = 0, 1) PROMIS ES PAM

Broad injury class = classification of mild, moderate, severe fracture based on OTA/AO classification; CAT = Computer adaptive test; CCI = Charlson Comorbidity Index; EQ-5D-3L = European Quality of Life 5 Dimensions 3 Level Version; fx dominant = fracture of dominant arm; fx ipsilateral = fracture of ipsilateral arm; fx nondom = fracture of nondominant arm; Grip = Grip Strength; IMD factor = Indices of Multiple Deprivation factor; NPRS = 11-point Numerical Rating Scale of Pain; PAM = Patient Activation Measure; PCS-13 = 13-item Pain Catastrophizing Scale; PROMIS = Patient-Reported Outcome Measurement Information System; PROMIS ES = PROMIS Emotional Support; PROMIS IS = PROMIS Instrumental Support; PROMIS PF CAT = PROMIS Physical Function CAT; PROMIS PI CAT = PROMIS Pain Interference CAT; PROMIS UE CAT = PROMIS Upper Extremity Physical Function CAT; PSEQ-2 = 2-item Pain Self-Efficacy Questionnaire; Quick-DASH = Quick Disabilities of Arm, Shoulder, and Hand; TSK-11 = 11-item Tampa Scale of Kinesiophobia.

sizes (Table 2). We selected the 29-variable model a priori as the primary model on which to test model performance in an effort to achieve a balance between (1) optimizing model performance, while limiting the risk for model overfit, and (2) selecting the variables that were deemed most clinically relevant (associated with PROs) based on current evidence. There are a wide range of statistical methods that can be used to reduce the risk for model overfit, which mostly revolve around selecting a smaller subset of the most influential explanatory variables. In this study, we opted to use a clinically oriented approach to select variables, recognizing the substantial evidence concerning modifiable risk factors related to mental health (eg, symptoms of depression and anxiety, negative pain thoughts, and fear of movement) and social health (eg, instrumental and emotional social support) that are dominantly associated with pain intensity and level of capability after injury.[1,3,4,20] Two senior authors selected the variables in a stepwise approach and arrived at 5 different subsets (Table 2).

2.3. Development of ANN models

First, we randomly split the entire data set into 3 data subsets: a training subset, validation subset, and test subset, each consisting of 70%, 15%, and 15% of the total observations, respectively. We then used the training subset (data from 514 patients, 70%) to develop a total of 20 ANN models (for 4 response variables using 5 different subsets of explanatory variables), by running data through the models to iteratively adjust the weights of the connections (more elaborate explanation of how ANNs work is included in Appendix A, http://links.lww.com/OTAI/A87). We proceeded to use the validation subset (data from 110 patients, 15%) to validate and adjust the developed ANN models. This process involved editing several hyperparameters in the model with the aim of achieving better model performance. Finally, the test subset (fresh data from 110 patients [15%] that had not been used previously) was used to test the primary 29-variable ANN model.

2.4. Development of linear regression models

Similarly, we developed 20 LR models for each of our 4 response variables using 5 different subsets of explanatory variables from the training and validation subsets with proprietary software (SAS Software, Cary, NC). The performance of primary 29-variable LR model was assessed in the test subset.

2.5. Response variables and model performance

The response variables (outcomes of interest) were levels of capability, pain interference (incapability specifically related to pain), and pain intensity 6–9 months after upper extremity fracture. The level of capability was measured using the Quick Disability of the Arm, Shoulder, and Hand (Quick-DASH) questionnaire and the Patient-Reported Outcome Measurement Information System (PROMIS) Upper Extremity Physical Function (UE) computer adaptive test (CAT). Pain interference was measured using the PROMIS Pain Interference (PI) CAT. Pain intensity was measured using the Numerical Pain Rating Scale (NPRS). We defined model performance as the accuracy of the estimated PRO value to be within the distribution-based minimal important difference (MID)[21] threshold (0.5 times the standard deviation of the change from baseline to follow-up[22]) of the actual score.

2.6. Variable importance

The definition of variable importance (eg, the contribution of an explanatory variable to the estimation of an outcome) is different between LR and ANN models. For LR, variable importance can be determined by assessing the (1) regression coefficient: the effect size of a correlation, (2) *P*-value: probability that a correlation is due to random chance, and (3) semipartial R^2: the proportion of the variation in the outcome variable that is explained by the explanatory variable. For ANN, while there is no gold standard for evaluating the relative importance of explanatory variables, a popular method is determining the average of "slopes" for each variable to define their level of importance within an ANN model.[23] In other words, this approach estimates the contribution of a variable to the estimation of an outcome but provides no information on how an explanatory variable is associated with an outcome. Owing to the fundamental differences in determining the variable importance within both models, it is not possible to directly compare them.

2.7. Software

Data preprocessing and ANN analysis were performed using Python and the Pytorch statistical package. LR analysis was performed in SAS (SAS Software, Cary, NC).

3. Results

3.1. Is there a difference in performance of ANN and LR models in estimating 9-month level of capability, pain interference, and pain intensity after upper extremity fracture?

ANNs outperformed LR in estimating 9-month Quick-DASH, PROMIS UE PF, and PROMIS PI scores in all models except the 20-variable models. ANN also outperformed LR in estimating 9-month NPRS scores in all models except the 20-variable and 23-variable models (Table 3). The accuracy of ANN versus LR in the primary 29-variable model of the test subset was 83% versus 73% for Quick-DASH, 68% versus 65% for PROMIS UE PF, 66% versus 62% for PROMIS PI, and 78% versus 65% for NPRS. Notably, mental and social health factors contributed relatively more to the estimation of the outcome than other types of variables in both models (ANN: Fig. 1 and LR: Table 4).

3.2. What is the difference in performance of ANN and LR models that incorporate different numbers of explanatory variables?

The performance of ANN and LR models in the estimation of all outcomes generally improved with an increasing number of explanatory variables included in the models (Table 3). The only exception was the decrease in performance between the 34-variable and 54-variable ANN model for PROMIS UE PF (from 74% to 71%). ANN models seem to benefit most from a higher number of included variables.

4. Discussion

Symptom intensity and level of capability after recovery from injury can be estimated by analyzing their relationship with mental, social, and pathophysiological factors soon after injury. We sought to compare the performance of ANN versus LR in estimating 9-month level of capability, pain interference, and pain intensity after upper extremity fracture. We also evaluated the differences in estimating performance of both models based on various subsets of explanatory variables. We found that ANNs outperformed LR in estimating 9-month PROs after upper extremity fracture, especially when a larger number of variables were included. LR performed better in models with a smaller number of included variables. Notably, mental and social health

TABLE 3

Accuracy of LR versus ANN in estimating PROs at 6–9 months using various data subsets

	Linear regression		Artificial neural network	
	Training subset	Validation subset	Training subset	Validation subset
Quick-DASH				
M-20	63%	61%	99%	64%
M-23	65%	58%	100%	65%
M-29	74%	72%	100%	80%
M-34	76%	75%	100%	81%
M-54	81%	77%	100%	85%
PROMIS UE				
M-20	60%	59%	83%	54%
M-23	62%	57%	85%	66%
M-29	64%	61%	99%	67%
M-34	65%	66%	97%	74%
M-54	74%	70%	100%	71%
NPRS				
M-20	62%	69%	99%	65%
M-23	62%	67%	91%	63%
M-29	68%	66%	99%	74%
M-34	70%	75%	99%	78%
M-54	77%	77%	100%	86%
PROMIS PI				
M-20	54%	47%	77%	40%
M-23	56%	49%	82%	49%
M-29	60%	63%	93%	65%
M-34	61%	65%	98%	67%
M-54	70%	69%	100%	72%

factors contributed relatively more to the estimation of the comfort and capability compared with other variables in both models.

4.1. Limitations

This study must be viewed in the light of several limitations. First, one could argue that variable selection based on consensus of senior authors is somewhat arbitrary.[10] However, this approach allowed us to compare ANNs and LR both with the same variables as well as with an increasing number of variables. Although our approach might not have resulted in the best performing models possible and different subsets of variables could have led to different results, thoughtful preselection of variables offers a practical means of achieving generalizability and supporting the feasibility of these models in clinical settings. Second, we used a distribution-based method for the estimation of MID (0.5 × standard deviation of delta score) rather than an anchor-based approach. Distribution-based methods are considered inferior compared to anchor-based MCID because they depend on a sample distribution rather than the patient's perspective,[24,25] but the distribution-based method may be adequate for this first step. Third, we combined both baseline and 2–4-week data points as explanatory variables within the models. While one might argue that estimation models could be more valuable when configured using data points as close as possible to injury, we believe an early window of recovery from baseline to 2–4 weeks better reflected a timeframe where patients at risk of prolonged pain and incapability may be identified, including those developing unhealthy mindsets and ongoing levels of distress and unhelpful thoughts that impede recovery. Notably, mental health interventions may be most effective during this period.[26,27]

4.2. Is there a difference in the performance between ANN and LR in estimating 9-month levels of capability, pain interference, and pain intensity after upper extremity fractures?

The observation that ANNs outperformed LR in the primary 29-variable model suggests that ANNs might be a better choice when a large number of potentially relevant variables are used including multiple psychosocial variables, which are often nonparametric and colinear. Previous comparisons of ANN and LR did not address psychosocial variables, whereas this data set included a large number of mental and social factors. These factors contributed most to the estimation of comfort and capability, which is consistent with the evidence regarding their strong association.[1,3,4,28–36] One might want to consider using ANN models as a favorable statistical approach for large nonparametric data sets with colinear variables and complex nonlinear correlations given the potential advantages over LR.[5–7] This is underlined by the observation that one variable had to be

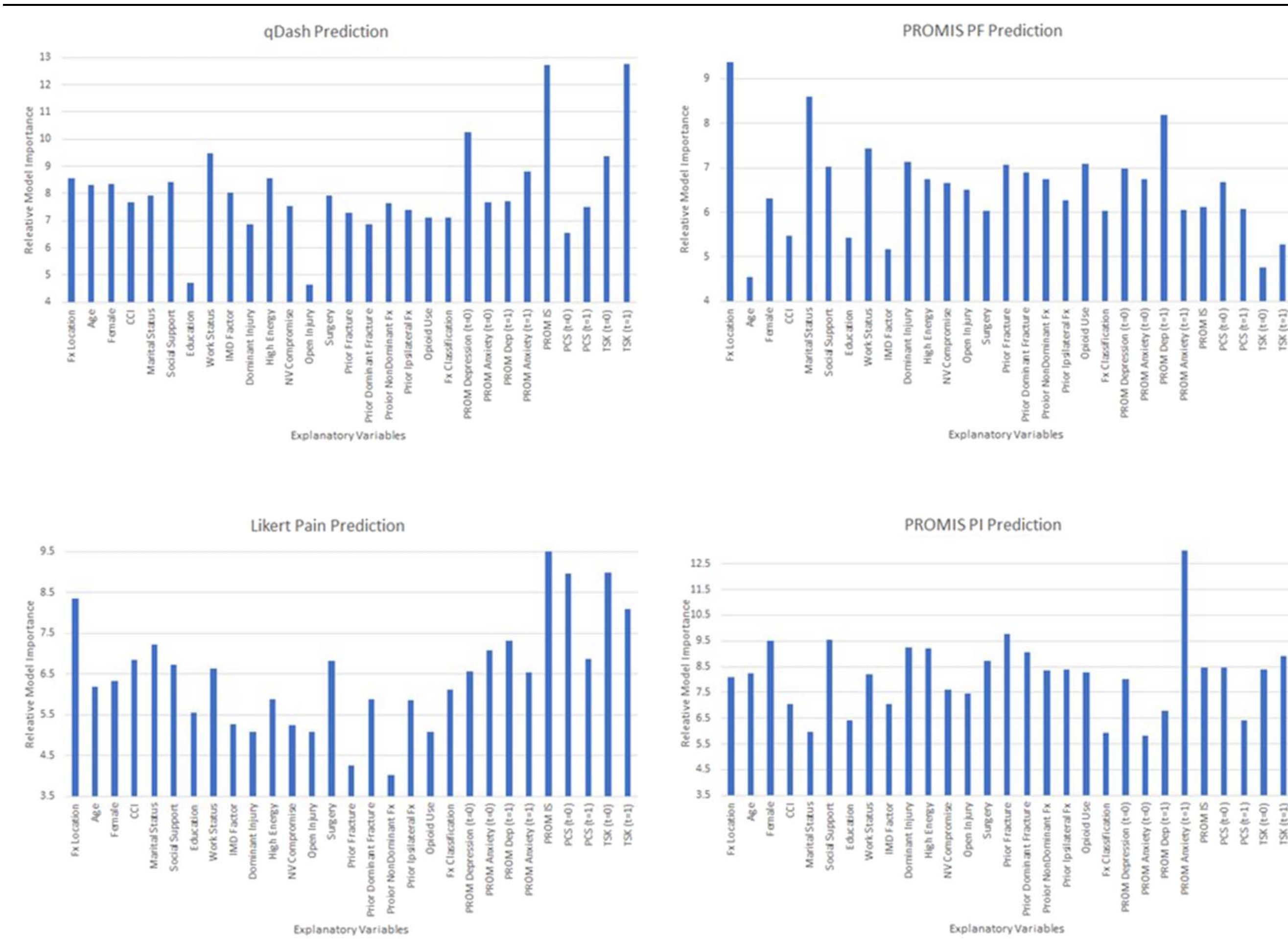

Figure 1. Relative weights for importance of variables in the primary 29-variable ANN model.

TABLE 4

Results of linear regression model used on the primary 29-variable subset

Variables	Quick-DASH		PROMIS UE PF		NPRS		PROMIS PI	
	Regression Coefficient	*P*	Regression Coefficient	*P*	Regression Coefficient	*P*	Regression Coefficient	*P*
Age	**0.14**	**0.014**	**−0.32**	**<0.001**	0.0097	0.16	0.047	0.17
Female	**4.3**	**<0.001**	**−1.8**	**0.0053**	**0.51**	**<0.001**	0.39	0.54
CCI	−0.26	0.69	**0.96**	**0.02**	0.0044	0.96	0.10	0.80
IMD factor	0.0042	0.83	0.012	0.34	−0.00024	0.92	−0.0042	0.73
Marital status								
Partner/married	*Reference value*		*Reference value*		*Reference value*		*Reference value*	
Single	0.86	0.52	0.27	0.75	−0.14	0.40	0.054	0.95
Separated	0.83	0.52	−0.43	0.60	**−0.43**	**0.006**	−0.54	0.50
Social status								
Partner/family	*Reference value*		*Reference value*		*Reference value*		*Reference value*	
Lives alone	**3.8**	**0.012**	−1.2	0.21	0.35	0.063	1.1	0.26
Part/full-time care	**4.7**	**0.0024**	**−2.0**	**0.039**	**0.56**	**0.0033**	1.4	0.14
Years of education	0.048	0.77	−0.13	0.22	0.020	0.31	0.051	0.62
Work status								
Working/homemaker	*Reference value*		*Reference value*		*Reference value*		*Reference value*	
Unemployed	**6.3**	**0.002**	**−6.3**	**<0.001**	**1.5**	**<0.001**	2.1	0.095
Retired	2.5	0.091	−1.8	0.051	**0.42**	**0.021**	**3.0**	**0.0012**
Workers comp	**10**	**<0.001**	**−5.6**	**<0.001**	**1.5**	**<0.001**	1.6	0.25
Fracture location								
Wrist	*Reference value*		*Reference value*		*Reference value*		*Reference value*	
Elbow/shoulder*	−0.53	0.36	**−1.1**	**0.0017**	**−0.19**	**0.0071**	−0.31	0.38
Injury on dominant side	1.1	0.17	−0.52	0.31	0.12	0.22	0.59	0.24
High energy injury	**3.5**	**0.0058**	−1.2	0.12	**0.64**	**<0.001**	**2.4**	**0.002**
Neurovascular injury	−0.23	0.9	−1.2	0.33	−0.032	0.9	0.17	0.89
Open injury	−0.49	0.88	1.1	0.59	−0.028	0.94	0.88	0.67
Surgery	2.1	0.10	−0.56	0.48	**0.30**	**0.05**	0.67	0.40
Prior fracture	−1.4	0.34	−0.46	0.63	0.13	0.47	0.32	0.73
Prior fracture of dominant arm	1.2	0.44	1.7	0.097	−0.062	0.75	−1.2	0.21
Prior fracture of ipsilateral arm	1.7	0.25	−1.0	0.28	0.035	0.85	0.49	0.60
Broad injury classification	1.7	0.85	−1.0	0.16	0.035	0.90	0.49	0.23
Opioid use at 2 weeks	−0.099	0.095	−0.46	0.88	−0.0081	0.22	−0.39	0.52
PROMIS Depression CAT (t = 0)	2.2	0.13	0.12	0.72	0.19	0.17	**0.52**	**<0.001**
PROMIS Depression CAT (t = 1)	**−0.087**	**0.0093**	**−0.013**	**0.011**	**−0.0096**	**0.024**	**−0.16**	**<0.001**
PROMIS Anxiety CAT (t = 0)	−0.15	0.28	**0.095**	**<0.001**	**−0.016**	**0.0029**	**−0.15**	**<0.001**
PROMIS Anxiety CAT (t = 1)	**0.11**	**<0.001**	−0.24	0.19	**0.036**	**0.0016**	**0.34**	**<0.001**
PCS-13 score (t = 0)	0.38	0.44	**−0.067**	**0.040**	0.031	0.64	0.23	0.40
PCS-13 score (t = 1)	**0.062**	**<0.001**	0.10	0.18	**−0.0046**	**<0.001**	**−0.042**	**0.040**
TSK-11 score (t = 0)	**−0.30**	**<0.001**	**0.071**	**0.0096**	**−0.055**	**<0.001**	**−0.11**	**<0.001**
TSK-11 score (t = 1)	**0.79**	**<0.001**	**−0.27**	**<0.001**	**0.14**	**<0.001**	**0.39**	**<0.001**
PROMIS IS	**1.4**	**<0.001**	−0.68	0.10	**0.15**	**<0.001**	0.60	0.14

Bold indicates statistical significance. The variable for prior fracture of the nondominant arm was omitted from the model because of multicollinearity with the other prior fracture variables.
Broad injury class = classification of low, moderate, high fracture complexity based on OTA/AO classification; higher scores indicate more complexity; CAT = Computer Adaptive Test; CCI = Charlson Comorbidity Index; IMD factor = Indices of Multiple Deprivation factor; PCS-13 = 13-item Pain Catastrophizing Scale; PROMIS = Patient-Reported Outcome Measurement Information System; PROMIS IS = PROMIS Instrumental Support; TSK-11 = 11-item Tampa Scale of Kinesiophobia.
* Shoulder and elbow fractures were grouped to provide more weight and statistical significance because wrist fractures had almost double the number of observations than shoulder/elbow individually. Proximal fracture locations likely have a similar effect on the level of capability while wrist likely has a different effect.

removed from the 29-variable LR model because of multicollinearity, while the ANN model had no problems processing this.

4.3. What is the difference in performance of ANN and LR models that incorporate different numbers of explanatory variables?

The observation that both ANN and LR models performed better with a larger number of variables suggests that both methods benefit from more data points. However, the risk that a model, whether developed using ANN or LR, will measure random variation (noise) rather than relevant variation (associations) (model overfit) also increases when more variables are included. The risk of model overfit is influenced by sample size (observations per variable), quality of the data, and importance of the included variables to the estimated outcome. The observed drop in performance between the 34-variable and 54-variable ANN models used to estimate PROMIS UE PF scores might be an example of model overfit.

ANNs performed better than LR in models where a higher number of variables were included (and improved more with a larger number of included variables than LR), compared to LR which performed better in models with a low number of included variables. This might reinforce the suggestion that ANNs are better in processing large nonparametric data sets with nonlinear correlations,[5–7], while LR might be better in processing smaller parametric data sets with linear correlations.[7]

Notably, there is some evidence that suggests that ANNs can perform as accurate as logistic regression while using less variables when variable selection methods are used.[10] This points to the potential advantage of using statistical variable selection methods to only include the most influential variables while omitting noise from redundant variables. Thoughtful preselection of potentially relevant and modifiable variables combined with statistical variable selection methods might be the best strategy to increase the generalizability and clinical relevance, while optimizing performance of the estimation models.

5. Conclusion

A better understanding of the optimal statistical method to estimate health outcomes with greater accuracy has the potential to inform the development of models that can be applied to enhance clinical decision-making and tailor more personalized care pathways through prognostic information generated early after injury. The observation that ANNs performed better in estimating 9-month PROs than LR, particularly with a larger number of variables, points to a potential benefit of ANNs. Given the relatively comparable performance of both ANNs and LR, other factors such as data distribution, multicollinearity, and complexity of the correlations should also be considered when deciding on statistical approach. Prioritizing the inclusion of modifiable variables such as psychosocial factors might be useful because they could help estimate and visualize the potential effects of certain treatments. Future studies should perform an external validation of our ANN algorithms and may also use anchor-based MCID estimation instead of distribution-based methods in the process. Studies continuing this line of research may help select an optimal method and set of variables to enable the estimation of health outcomes in real-time clinical practice.

REFERENCES

1. Jayakumar P, Teunis T, Vranceanu AM, et al. Early psychological and social factors explain the recovery trajectory after distal radial fracture. *J Bone Joint Surg Am*. 2020;102:788–795.
2. Alsalman A, Shah R, Thomas JE, et al. Symptoms of depression and catastrophic thinking attenuate the relationship of pain intensity and magnitude of incapability with fracture severity. *J Psychosom Res*. 2022; 158:110915.
3. Jayakumar P, Overbeek CL, Lamb S, et al. What factors are associated with disability after upper extremity injuries? A systematic review. *Clin Orthop Relat Res*. 2018;476:2190–2215.
4. Vranceanu AM, Bachoura A, Weening A, et al. Psychological factors predict disability and pain intensity after skeletal trauma. *J Bone Joint Surg Am*. 2014;96:e20.
5. Almeida JS. Predictive non-linear modeling of complex data by artificial neural networks. *Curr Opin Biotechnol*. 2002;13:72–76.
6. Lek S, Delacoste M, Baran P, et al. Application of neural networks to modelling nonlinear relationships in ecology. *Ecol Modell*. 1996;90: 39–52.
7. Akbilgic O, Davis RL. The promise of machine learning: when will it be delivered? *J Card Fail*. 2019;25:484–485.
8. Teunis T, Jayakumar P, Ring D. The problem of collinearity in mental health and patient reported outcome research. *J Hand Surg Am*. 2021;46: e1–e2.
9. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol*. 2019;110:12–22.
10. Oosterhoff JHF, Gravesteijn BY, Karhade AV, et al. Feasibility of machine learning and logistic regression algorithms to predict outcome in orthopaedic trauma surgery. *J Bone Joint Surg Am*. 2022;104:544–551.
11. Ennis M, Hinton G, Naylor D, et al. A comparison of statistical learning methods on the Gusto database. *Statistics Med*. 1998;17:2501–2508.
12. Mitchell TM. Does machine learning really work? *AI Mag*. 1997;18:11.
13. Deo RC, Nallamothu BK. Learning about machine learning: the promise and pitfalls of big data and the electronic health record. *Circ Cardiovasc Qual Outcomes*. 2016;9:618–620.
14. Goldstein BA, Navar AM, Carter RE. Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. *Eur Heart J*. 2017;38:1805–1814.
15. Deo RC. Machine learning in medicine. *Circulation*. 2015;132: 1920–1930.
16. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artif Intell Med*. 2001;23:89–109.
17. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med*. 2004;66:411–421.
18. Kernbach JM, Staartjes VE. Foundations of machine learning-based clinical prediction modeling: Part II-generalization and overfitting. *Acta Neurochir Suppl*. 2022;134:15–21.
19. Jayakumar P, Teunis T, Vranceanu AM, et al. Relationship between magnitude of limitations and patient experience during recovery from upper-extremity fracture. *JB JS Open Access*. 2019;4:e0002.
20. Wright MA, Adelani M, Dy C, et al. What is the impact of social deprivation on physical and mental health in orthopaedic patients? *Clin Orthop Relat Res*. 2019;477:1825–1835.
21. Wright A, Hannon J, Hegedus EJ, et al. Clinimetrics corner: a closer look at the minimal clinically important difference (MCID). *J Man Manip Ther*. 2012;20:160–166.
22. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care*. 2003;41:582–592.
23. Fish KE, Blodgett JG. A visual method for determining variable importance in an artificial neural network model: an empirical benchmark study. *J Target Meas Analysis Mark* 2003;11:244–254.
24. McGlothlin AE, Lewis RJ. Minimal clinically important difference: defining what really matters to patients. *JAMA*. 2014;312:1342–1343.
25. Lyman S, Lee YY, McLawhorn AS, et al. What are the minimal and substantial improvements in the HOOS and KOOS and JR versions after total joint replacement? *Clin Orthop Relat Res*. 2018;476:2432.
26. Archer KR, Mackenzie EJ, Bosse MJ, et al. Factors associated with surgeon referral for physical therapy in patients with traumatic lower-extremity injury: results of a national survey of orthopedic trauma surgeons. *Phys Ther*. 2009;89:893–905.
27. Pray JE. Responding to psychosocial needs: physician perceptions of their referral practices for hospitalized patients. *Health Soc Work*. 1991;16: 184–192.
28. Turk DC, Okifuji A. Psychological factors in chronic pain: evolution and revolution. *J Consult Clin Psychol*. 2002;70:678–690.
29. Shin YH, Yoon JO, Kim YK, et al. Psychological status is associated with symptom severity in patients with carpal tunnel syndrome. *J Hand Surg Am*. 2018;43:484.e1–484.e8.
30. Vissers MM, Bussmann JB, Verhaar JAN, et al. Psychological factors affecting the outcome of total hip and knee arthroplasty: a systematic review. *Semin Arthritis Rheum*. 2012;41:576–588.
31. Vranceanu AM, Barsky A, Ring D. Psychosocial aspects of disabling musculoskeletal pain. *J Bone Joint Surg Am*. 2009;91:2014–2018.
32. Talaei-Khoei M, Fischerauer SF, Lee SG, et al. Pain catastrophizing mediates the effect of psychological inflexibility on pain intensity and upper extremity physical function in patients with upper extremity illness. *Pain Pract*. 2017;17:129–140.
33. Talaei-Khoei M, Fischerauer SF, Jha R, et al. Bidirectional mediation of depression and pain intensity on their associations with upper extremity physical function. *J Behav Med*. 2018;41:309–317.
34. Talaei-Khoei M, Mohamadi A, Mellema JJ, et al. The direct and indirect effects of the negative affectivity trait on self reported physical function among patients with upper extremity conditions. *Psychiatry Res*. 2016; 246:568–572.
35. Fischerauer SF, Talaei-Khoei M, Bexkens R, et al. What is the relationship of fear avoidance to physical function and pain intensity in injured athletes? *Clin Orthop Relat Res*. 2018;476:754–763.
36. Teunis T, al Salman A, Koenig K, et al. Unhelpful thoughts and distress regarding symptoms limit accommodation of musculoskeletal pain. *Clin Orthop Relat Res*. 2022;480:276–283.