BMC Medical Informatics and Decision Making

REVIEW

Open Access

A narrative review of the use of PROMs and machine learning to impact value-based clinical decision-making

Michal Pruski[1,2], Simone Willis[3*] and Kathleen Withers[2,4]

Abstract

Purpose This review summarises the studies which combined Patient Reported Outcome Measures (PROMs) and Machine Learning statistical computational techniques, to predict patient post-intervention outcomes. The aim of the project was to inform those working in value-based healthcare how Machine Learning can be used with PROMs to inform clinical practice.

Methods A systematic search strategy was developed and run in six databases. The records were reviewed by a reviewer if they matched the review scope, and these decisions were scrutinised by a second reviewer.

Results 82 records pertaining to 73 studies were identified. The review highlights the breadth of PROMs tools investigated, and the wide variety of Machine Learning techniques utilised across the studies. The findings suggest that there has been some success in predicting post-intervention patient outcomes. Nevertheless, there is no clear best performing Machine Learning approach to analyse this data, and while baseline PROMs scores are often a key predictor of post-intervention scores, this cannot always be assumed to be the case. Moreover, even when studies looked at similar conditions and patient groups, often different Machine Learning techniques performed best in each study.

Conclusion This review highlights that there is a potential for PROMs and Machine Learning methodology to predict patient post-intervention outcomes, but that best performing models from other previous studies cannot simply be adopted in new clinical contexts.

Keywords Prudent healthcare, Decision-making, Value in health, Algorithms, Prediction, Patient reported outcomes

*Correspondence:
Simone Willis
WillisS5@cardiff.ac.uk
[1]School of Health Sciences, The University of Manchester, Manchester, UK
[2]CEDAR, Cardiff and Vale UHB, Cardiff, UK
[3]Specialist Unit for Review Evidence, Cardiff University, Cardiff, UK
[4]School of Engineering, Cardiff University, Cardiff, UK

Background

Wales is at the forefront of collecting patient reported outcome measures (PROMs) in clinical practice on a national level [1–4]. As the digital revolution progresses, novel technologies, such as machine learning (ML) and other artificial intelligence (AI) techniques, offer new possibilities of utilising healthcare data. This includes both facilitating big data research using routinely collected data, and the application of these findings to facilitate patient care.

Being a subset of AI, ML is a group of computational techniques which allows researchers to better scrutinise their data. While ML techniques utilise various levels of supervision (labelling the data), they are all potent tools that allow the identification of relationships in the data by creating models. Such models could potentially be used to make predictions for e.g. clinicians to prognosticate how patients might perform under two different treatment regimes. Therefore, one of the potential clinical uses of ML is to identify the best care pathways for individual patients. Consequently, within the context of this work, ML is best understood as a group of complex statistical modelling techniques.

Value-Based healthcare is concerned with the real-life impact of clinical decisions to achieve better outcomes and experiences for service users [5]. As such, it recommends assessing patients' ability to carry out every-day activities as meaningful measures, over changes in e.g. biochemistry markers. Value-Based healthcare also encourages prudent use of limited resources, and while this is a complex principle, it is clear that interventions need to make an impact to be considered valuable. PROMs, which often focus on measuring patients' symptom severity and ability to undertake everyday activities, can help to identify such impactful interventions.

In Wales, value-based healthcare application is directed by four principles of Prudent Healthcare [6]:

1. Achieve health and wellbeing with the public, patients and professionals as equal partners through co-production
2. Care for those with the greatest health need first, making the most effective use of all skills and resources
3. Do only what is needed, no more, no less; and do no harm
4. Reduce inappropriate variation using evidence-based practices consistently and transparently

The ability to predict which patients might benefit from a given intervention could help to bring these principles into practice. Having evidence-based predictions will allow patients and clinicians to make better informed decisions. Such predictive models will help identify those of greatest need. They could help ascertain which interventions potentially do not offer any benefit to specific patients. Finally, predictive models might help to usher in precision medicine to help highlight instances of appropriate variation of treatment recommendations.

While some reviews have attempted to address aspects of this topic, they did not utilise comprehensive literature search strategies and so painted only a limited picture of such applications [7, 8]. One of these reviews did specifically look at the ability of ML in combination with PROMs to predict patient outcomes [7]. However, this study only interrogated two databases (PubMed and Scopus) and discussed fifteen studies. The other review focused only on the use of PROMs in clinical AI trials and searched only the ClinicalTrials.gov register [8]. It did not specifically look at the ability of PROMs and ML to predict patient outcomes.

This review was undertaken with the objective to inform stakeholders, such as decision-makers and researchers working in value-based healthcare, about the current applications of ML techniques to PROMs data. It particularly aimed to inform stakeholders how PROMs data collected during routine clinical practice can be utilised in a value-based healthcare system. This information can be used to identify areas of interest for undertaking similar projects, as well as in identifying approaches which have historically not proven to be successful. It is hoped that this review will provide a quick reference guide for those looking to identify studies in their field of interest.

This review looked at published studies which combined PROMs and ML to predict patients' post-intervention outcomes (Table 1). The review included studies where PROMs were used either as outcome measures and/or as predictors in the ML models. A broad understanding of 'healthcare intervention' has been adopted, inclusive of such phenomena as a hospital stay in specialist care, surgery or psychological interventions (Table 1). Nevertheless, due to this already broad scope, the review did not consider other types of outcomes, such as costs or care-giver wellbeing. Moreover, due to the volume of identified studies we did not consider other predictor variables or outcome measures that might have been used in the identified studies.

Main text

Methods

As the review aimed to inform stakeholders on the topic, we conducted a narrative review. Whilst a scoping review would provide a more rigorous process (e.g. due to a more comprehensive search), we were limited due to time and resources. To increase the methodological rigour of the narrative review, we drew on systematic review methods by systematically searching databases, and having a

Table 1 Review scope

	Inclusion	Exclusion
Population	Patients receiving any healthcare intervention e.g. - Surgical intervention - Psychological intervention - Hospital stay - Pharmacological treatment Patients within any part of the healthcare system (e.g. primary care, secondary care)	
Phenomenon of interest	Combining ML with PROMs There are three scenarios where ML and PROMs could be combined: 1. PROMs data collected pre-intervention or phenomenon, followed by ML predicting clinical outcome (i.e. entering PROMs data and using ML for the prediction of clinical outcome) 2. Using ML for the prediction of post-intervention or phenomenon PROMs (ML can be using any kind of data) 3. Pre-intervention or phenomenon PROMs with ML predicting post-intervention PROMs PROMs can be combined with other data (e.g. clinical data, demographic) All validated PROMs ML is defined as statistical computational modelling techniques utilising various levels of user supervision.	PROMs that have not been validated. If only a single item scale such as one visual analogue scale or rating scale was used, e.g. a pain scale was used.
Outcome	Patient outcomes - Clinical outcomes - Mortality - PROMs Healthcare system - Prioritisation of patients - Waiting list - Cost savings - Cost effectiveness	Caregiver outcomes Financial outcomes Diagnosis prediction
Study design	Any study design (including quantitative, qualitative, mixed-methods, case study, protocols, systematic reviews, rapid reviews; including conference abstracts and registered trials)	Narrative reviews Letters Editorials
Year	2000 onwards	Studies published before 2000
Language	English language	Not English language

second reviewer check a proportion of abstracts and data extraction. Due to the narrative nature of the review, it was not registered on PROSPERO.

A search strategy (Supplementary File 1) was developed and run in Medline All (Ovid) to identify relevant records using a combination of free-text and indexed terms. The search included broad terms, such as 'AI' to account for the fact that some authors might have used this more general term, rather than specifically describing their techniques as ML. The search was adapted and run in the following five databases: Embase (Ovid), The Cochrane Library, Scopus, IEEE Xplore and ACM Digital Library. The searches were carried out on the 11th October 2023. Records were imported into EndNote 20 and deduplicated [9]. Two reviewers screened studies at title/abstract and full-text using Endnote. One reviewer assessed all records at title/abstract against the inclusion criteria (Table 1). Full texts were obtained, and assessed by one reviewer against the inclusion criteria. At both stages, a second reviewer checked all included records and 10% of excluded records, noting any discrepancies. Discrepancies were resolved through discussion between the two reviewers.

One reviewer extracted key information from each record into a table, which was checked by the second reviewer. Key information was summarised as a narrative and is presented below.

Results

Literature search results

The searches retrieved a total of 2,075 records, with 1,789 records remaining after deduplication. Following title/abstract screening, 167 records were assessed at full-text, of these, 82 records pertaining to 73 individual studies met the inclusion criteria. The reasons for the exclusion of the remaining 85 records are provided in Fig. 1.

Of the included records, 25 considered the application of ML and PROMs to patients with hip and knee conditions, 14 regarded studies of patients with spinal conditions, and 9 included patients with other musculoskeletal

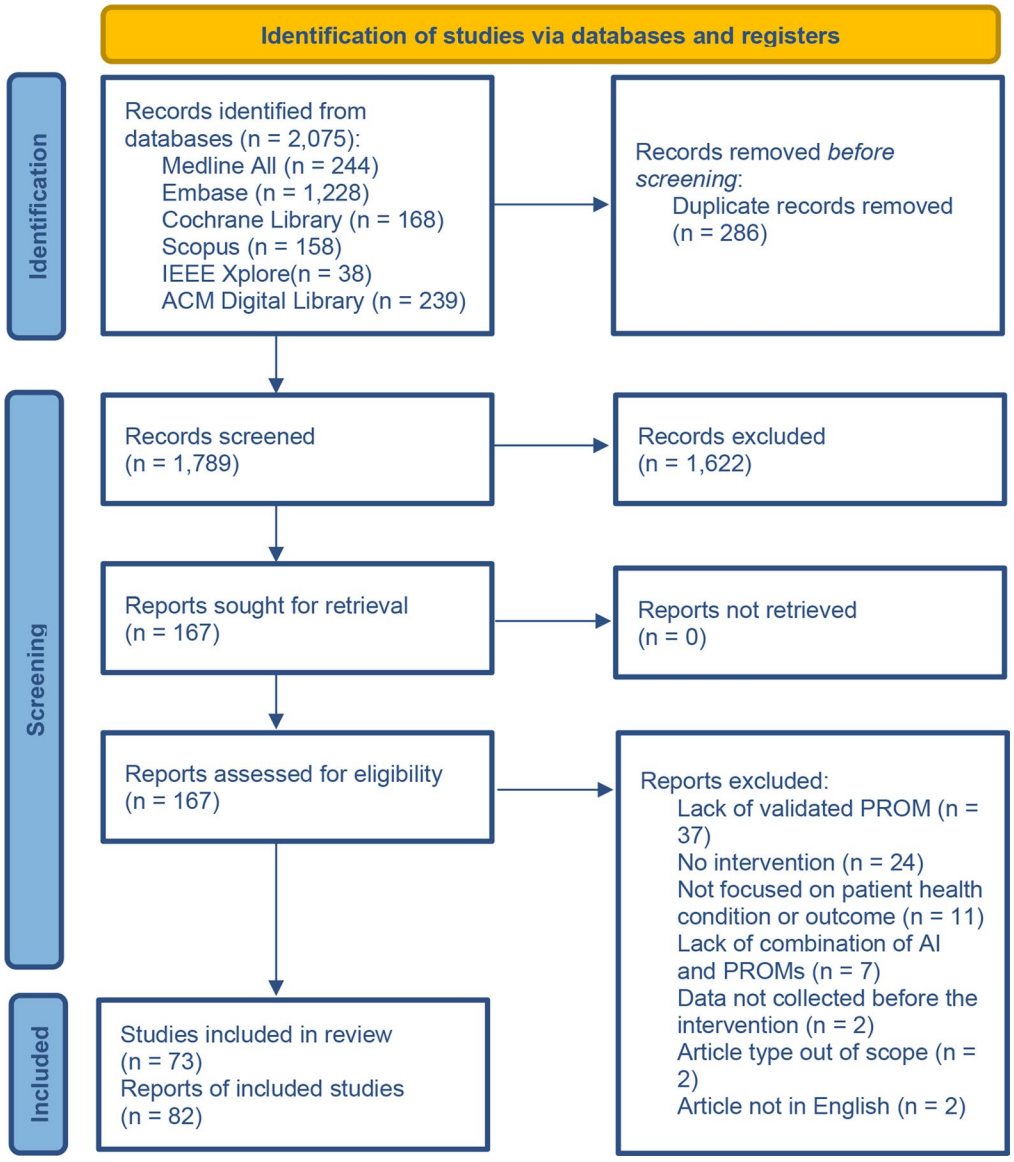

Fig. 1 Study selection flow diagram. Adapted from PRISMA. [10, 11]

conditions. There were 12 records that looked at cancer patients, four at patients with neurodegenerative conditions, and seven which focused on mental health. The remaining 11 records looked at patients with rheumatoid arthritis, COVID-19, cardiac and respiratory problems, stroke, snake bites, critical care, and the care of the elderly.

Findings from the literature

This section summarises the findings in the identified records, indicating the PROM tools and ML algorithms used. Only the PROM tools, and not specific sub-scores, are reported, and similar ML algorithms are grouped together, to allow for an overall narrative to be presented. These are challenging to summarise due to the heterogeneity in reporting of the studies' methodologies; for example, some studies described in detail which ML techniques they utilised for feature selection and which for classification when building their ML model, while some studies only stated the generic type of ML algorithm used. Briefly, the majority of records were published between 2019 and 2023 (Fig. 2). Across all studies (Table 2) 220 PROMs were used, of these there were 133 unique PROM tools, with different versions of the same core PROM tool, such as abbreviated versions, counted as different unique PROM tools. Many of the PROM tools were only investigated in single studies, while others, such as Short Form36, were used across a range of studies. Across all studies, 269 ML techniques were mentioned (Table 2). As noted, it is difficult to know

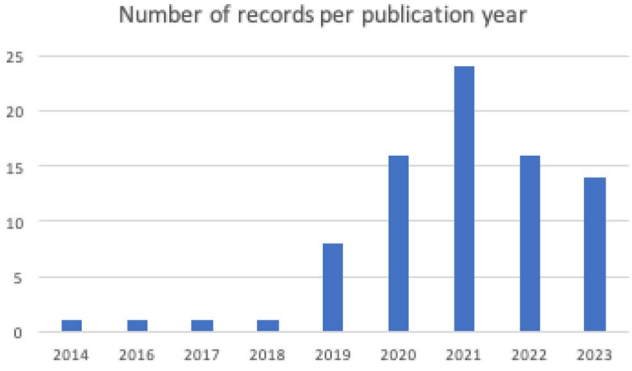

Fig. 2 Years of publication for the retrieved records

how many of these were unique instances due to lack of clarity in the reporting; Fig. 3 provides a summary of the ML techniques used. Of the techniques described in the identified studies, Boosting approaches were most popular, followed by Random Forest and Support Vector approaches, which have been historically 'three of the most powerful machine learning methods with demonstrated high predictive accuracies in many application domains' [12]. The following sections provide a brief overview of the included studies, and discussion of the main findings regarding the PROM tools used and the ML techniques employed.

Hips and knees

There were 25 records which described 20 studies relating to hip and knee conditions, out of which one was a systematic review and for three of these studies only protocols or registry entries were retrieved. There was one record of a systematic review looking at ML powered decision support systems for total hip and knee arthroplasty [26]. It listed twelve studies that considered PROMs as their prediction outputs out of a total 49 studies included in that systematic review. Of these, ten studies were not identified by the systematic search carried out as part of the present review, with the reasons for this discussed later on. Two records pertain to a study looking at the impact of the utilisation of a decision support tool in patients with knee osteoarthritis [17, 18]. One record described a study looking at using wearable sensor data to predict six-week postoperative outcomes in joint replacement patients and only utilised wearable sensor data as predictors in their models [13]. One record described a study that looked at both hip and knee total arthroplasty patients and the authors found all models to perform better than simple heuristics (rules of thumb) [14]. One record described a study that looked at predicting knee replacement surgery from symptomatic and radiographic data [16]. One record related to a study looking at pain and function outcomes 1-year after surgery [15]. One record described a study looking to predict 3-month postoperative outcomes [34]. Two records described a study looking at the capability of radiographic indices in predicting PROM scores [28, 29]. One record looked at the functional improvement in athletes with femoroacetubular impingement syndrome using a two year horizon [24]. Another study also looked at femoroacetubular impingement syndrome patient outcomes [25]. One study considered a visual analogue score for satisfaction as its outcome measure, and did not find PROMs to be important outcome predictors [22]. Conversely, another study found the baseline score of a PROM tool to be an important predictor measure of its outcome value [23]. Two records described a study looking at 1- and 2-year post-osteochondral allograft outcomes in knee cartilage defect patients [30, 31]. This study found diffident models to be the best performing for predicting different outcome PROM scores. Two records related to a study looking at patient willingness to undergo total knee arthroplasty when they had access to a prognostic tool, and utilised a 12-month follow-up window [36, 37]. One study looked to develop a precision medicine approach to managing knee osteoarthritis patients that are either overweight or obese and found different ML algorithms to be best at predicting different outcome measures they had considered [19]. Another study looked at predicting meaningful improvement after total knee arthroplasty [35]. One record described as study that looked to improve treatment decisions in hip and knee surgery patients [27]. Three of the records were of protocols, and for one of these a trial registry entry was also retrieved. One was a protocol looking at developing a decision support tool for patients undergoing total knee arthroplasty and adopting a 3-month postoperative horizon [32]. The second was a protocol of a study looking at osteoarthritis patients undergoing hip arthroplasty, focusing on the impact of traumatic experiences and mental conditions on postoperative outcomes [33]. The remaining two records related to a study looking at osteoarthritis patients with hip or knee problems [20, 21].

These studies utilised 42 unique PROM tools, with the most frequently utilised being the Short Form-36 and Knee Injury and Osteoarthritis Outcome Score (both used six times), followed by Hip Disability and Osteoarthritis Outcome Score, Hip Outcome Score, and Western Ontario and McMaster University Osteoarthritis Index, with each used five times (Table 2). Boosting, Random Forests and Neural Networks were the three most often explored algorithms (Table 2). Short Form-36 was the most studied PROMs outcome of interest, while the Hip Outcome Score was the most frequently identified PROMs tool that was a significant predictor of the studied outcomes (Fig. 4). Random Forests and Elastic-Net Penalised Logistic Regression were the most successful ML techniques studied (Fig. 4). Even though the studies

Table 2 Summary of PROM tools and ML techniques used in the included studies

Theme	First Author & Year	PROMs considered	ML techniques
Hips & Knees	Bini 2019 [13]	KOOS, HOOS, VR-12	Unsupervised ML for feature selection, K-Means analysis for cluster identification
	Fontana 2019 [14]	**KOOS, HOOS, SF-36**, EQ-5D, WOMAC	Logistic LASSO, RF, Linear SVM
	Harris 2021 [15]	**KOOS**, AUDIT, PHQ-2, EQ-5D-5 L	LR, LASSO, GBoost, QDA
	Heisinger 2020 [16]	KOOS, WOMAC	NN
	Jayakumar 2020 & 2021 [17, 18]	PROMIS, KOOS, PHQ −2 and -9, GAD −2 and −7, **Knee Osteoarthritis Decision Quality Instrument**	Not stated
	Jiang 2021 [19]	**WOMAC, SF-36**	Penalised Regression, Kernel Ridge Regression, RF, Reinforcement Learning Trees, List-Based Dynamic Treatment Regime, Residual Weighted Learning, Bayesian Additive Regression Trees, Zero-Order models
	Kastrup 2023 (protocol) & NCT04332055 (registry entry) [20, 21]	**OKS**, **OHS**, EQ-5D-3 L and the Shared Decision Making Questionnaires Questionniare-9	Not stated
	Kunze 2021 [22]	HOS, mHHS	Stochastic GBoost, RF, SVM, NN, ENPLR
	Kunze 2021 [23]	**HOS**, mHHS	StochasticGBoost, RF, SVM, NN, ENPLR
	Kunze 2021 [24]	**HOS**, mHHS, iHOT	Stochastic GBoost, RF, SVM, Adaptive GBoost, NN, ENPLR
	Kunze 2022 [25]	HOS, mHHS, **iHOT**	Stochastic GBoost, RF, SVM, XGBoost, NN, ENPLR
	Lopez 2021 (systematic review) [26]	HOOS, KOOS, SF-36	NN, Regression, Cluster Analysis, SVM, DT, Boosting, Bayesian Networks
	Milella 2022 [27]	**SF-12**	XGBoost, DT
	Ramkumar 2020 & 2021 [28, 29]	**HOS, mHHS, iHOT**	RF
	Ramkumar 2021 & Karnuta 2021 [30, 31]	**IKDC**, **KOS-ADL**, **SF-36**	LR, RF, GNB, XGBoost, Sigmoid XGBoost, Isotinic XGBoost, Top Three Ensemble methodology
	Ribbons 2023 (protocol) [32]	**WOMAC**, Depression Anxiety and Stress Scales-21, Pain Catastrophizing Scale, Brief Resilience Scale, Committed Action Questionnaire-8, Valued Living Scale, **SF-12 version 2**, Medical Outcome Study Social Support Survey, University of California Los Angeles Activity Scale, AUDIT	linear predictive models, DT, RF, GBoost, NN, Bayesian Soft Decision Trees
	Sergooris 2023 (protocol) [33]	**HOOS, SF-36, Global Perceived Effect Scale, Patient Specific Functioning Scale,** FACS-D, TSK-17, IEQ, Childhood Trauma Questionnaire, Mini International Neuropsychiatric Interview Simplified, Hospital Anxiety and Depression Scale, General Self-Efficacy Scale, Perceived Stress Scale	LASSO, DT, GBoost, recurrent NN
	Sniderman 2021 [34]	**HOOS**	LASSO
	Zhang 2022 [35]	WOMAC, SF-36	RF, XGBoost, SVM, LR, LASSO
	Zhou 2022 (protocol) & 2022 [36, 37]	**VR-12**, EQ-5D-3 L	LR, Classification Tree, XGBoost tree, RF
Spinal Conditions	Ames 2019 [38]	**ODI, Scoliosis Research Society-22, Optum SF-36v2 Health Survey**	Hierarchical clustering analysis
	Chan 2021 & 2022 [39, 40]	ODI, EQ-5D, North American Spine Society Satisfaction Questionnaire	K-Means Clustering
	Durand 2020 [41]	Scoliosis Research Society-22, ODI, SF-36	RF, ENR, LR, SVM with radial kernels, SVM with linear kernels
	Janssen 2021 [42]	SF-36, HADS, ODI, PCS	RF
	Liew 2020 [43]	**NDI**, **EQ-5D**, MSPQ, SES	Stepwise Regression, LASSO, Boosting, MuARS

Table 2 (continued)

Theme	First Author & Year	PROMs considered	ML techniques
	Merali 2019 [44]	**mJOA**, SF-36, **SF-6D**, NDI	RF, SVM and LR, Simple DT and NN; RF used for feature selection
	Muller 2021 [45]	COMI	LASSO cross validation, Ridge cross validation
	Rogers 2019 [46]	mJOA	RF, SVM, NN, DT
	Siccoli 2019 [47]	ODI	RF, XGBoost, Bayesian Generalised Linear Models, simple Generalised Linear Models, Boosted Trees, KNN, NN with a single hidden layer
	Sundararajan 2019 [48]	ODI	ENR
	Yagi 2022 [49]	**JOABPEQ**	Generalised Linear Regression, GLMM, LR, linear SVM, single-layer NN, Random Tree, 'linear-AS', 'tree-AS', XGBoost Linear, XGBoostTree, Chi-Squared Automatic Interaction Detection, classification trees, regression trees
	Zhang 2021 [50]	**mJOA**	DT, RF, ET, Adaboost, GBoost DT, Bernoulli Naïve Bayes, GNB, Passive Aggressive, QDA, Linear Discriminant Analysis, Linear SVM, SVM, KNN, SGD; feature processors: select percentile, select rate, Linear SVM pre-processor, ET pre-processor, Fast Independent Component Analysis, Feature Agglomeration, Principal Component Analysis
Other Musculoskeletal	Allaart 2023 (protocol) [51]	Constant-Murley Score, ASES, UCLA, OSS, WORC, DASH	Bayes Point Machine, Boosted DT, Penalised LR, NN, SVM
	Dipnall 2021 (protocol) [52]	WHODAS, EQ-5D-5 L	Linear Mixed Models, Generalised Linear Mixed Models, Longitudinal Multi-Level Factorization Machines Model, Longitudinal Support Vector Regression, Mixed Effects RF
	Kong 2020 [53]	**RMDQ**	LASSO
	Kumar 2020 [54]	SPADI, **ASES**, Simple Shoulder Test, **Constant-Murley Score**, **UCLA**	LR, XGBoost, Wide and Deep ML techniques
	Loos 2022 [55]	**MHQ**	LR, RF, GBoost
	Lu 2023 [56]	ASES, Single Assessment Numeric Evaluation, Constant-MurleyScore	RF
	Polce 2021 [57]	Single AssessmentNumeric Evaluation	Stochastic GBoost, RF, SVM, NN, ENPLR
	Verma 2021 [58]	**RMDQ**, **Work-ability index**, **PSEQ**, **FABQ**, **Global Perceived Effect Scale,** EQ-5D	XGBoost
	Verma 2022 [59]	HADS, Multidimensional Pain Inventory, Pain Disability Index, Psychological Inflexibility in Pain Scale, SF-36	For regression: LR, Passive Aggressive Regression, RF Regression, Stochastic Gradient Descent Regression, AdaBoost, Support Vector Regression, XGBoost Regression; for classification: balanced RF and Random Undersampling Boosting classifiers, both with DT as base estimators
	Vo 2023 (protocol) [60]	PROMIS, painDETECT, ODI, StarT Back Tool, Fear Avoidance Beliefs Questionnaire, Chronic Pain Acceptance Questionnaire, PCS, Interoception: Multidimensional Assessment of Interoceptive Awareness, PHQ, GAD-2, Perceived Stress Scale, Primary Care Post-traumatic Stress Disorder Screen, HEAL Positive Outlook, Global Physical Activity Questionnaire	NN, SVM
Cancer	Cunha 2021 [61]	Edmonton Symptom Assessment System, EORTC QLQ-C30	Feature selection: RF, XGBoost, Analysis of Variance F-Score, Recursive Feature Elimination with Cross-Validation fitted with a SVM, L1penalised Cox; classification: RF, KNN, XGBoost, LR, Voting Classifier (Ensemble)
	DeWees 2020 [62]	**CTCAE**	NN
	Golafshar 2020 [63]	CTCAE	NN

Table 2 (continued)

Theme	First Author & Year	PROMs considered	ML techniques
	Iivanainen 2020 & 2020 [64, 65]	CTCAE	XGBoost
	Nuutinen 2023 [66]	**EORTC QLQ-C30**, National Comprehensive Cancer Network distress thermometer	Variable selection: LR; classification: RF
	Qi 2017 [67]	**Expanded Prostate Cancer Index Composite – Short Form tool**	Deep NN
	Rossi 2021 [68]	MDASI	LR
	Savić 2021 [69]	**Prostate Symptom Score, International Index of Erectile Function – 5 item, Life Satisfaction Questionnaire – 11**	Naive Bayes, KNN, SVM, DT, RF; compared centralised and federated models
	Xu 2023 & Pfob 2023 & 2021 [70–72]	**BREAST-Q**	LR, XGBoost, NN (all noted as similarly well performing)
Neurodegenerative	Bougea 2023 [73]	PDQ-39	multivariate linear regression, Autoregressive Integrated Moving Average, Seasonal Autoregressive Integrated Moving Average, Long Short-Term Memory- recurrent NN
	Branco 2022 [74]	SF-36, Treatment Satisfaction Questionnaire for Medication, Fatigue Severity Scale, Beck Depression Inventory-II, Patient-Reported Indices in Multiple Sclerosis	LR, Linear Support Vector, GNB, RF
	Coratti 2023 (protocol) [75]	Spinal Muscular Atrophy Health Index	Not stated
	Rouleau 2020 [76]	**PDQ-39**	SVM
Mental Health	Bremer 2018 [77]	**QIDS**, **PHQ-9**, **EQ-5D**	Feature selection: LASSO; prediction: Linear Regression, Support Vector Regression, regression trees, ridge regression
	Camp 2022 [78]	QOLIE-10, **PHQ-9**	Multilayer Perceptron, RF, SVM, LR with Stochastic Gradient Descent, KNN, GBoost
	Chekroud 2016 [79]	**QIDS**, PHQ-9, EQ-5D	elastic net regularisation with GBoost
	Hufner 2022 [80]	**PHQ-4**	RF, Poisson Regression, KNN
	Kay 2021 [81]	QIDS	RF, GBoost, LR, Deep Learning, J48, Adaboost
	Manikis 2023 [82]	**HADS**, **EORTC QLQ-C30**, Positive and Negative Affect Schedule, Life Orientation Test, Mental Adjustment to Cancer scale, Sense of Coherence Scale, Connor-Davidson Resilience Scale, Mindful Attention Awareness Scale, Quality of Life Questionnaire – Breast Cancer Module, CancerBehaviorInventory – Brief and Post-Traumatic Growth Inventory, Fear of Cancer Recurrence Inventory, Revised Life Orientation Test, Cognitive Emotion Regulation Questionnaire - short, modified Medical Outcomes Study Social Support Survey	Balanced RF
	Martin 2019 [83]	Dimensional Anhedonia Rating Scale, Snaith Hamilton Pleasure Scale	Not stated
Other: Bariatric Surgery	Cao 2020 [84]	SF-36, obesity-related problems scale	Gaussian Bayesian Networks, multinomial discrete Bayesian Networks multivariable LR, Convolution NN
Other: Rheumatoid Arthritis	Curtis 2022 [85]	PROMIS, SF-36	RF, Elastic-Net Regularised Linear Model, XGBoost, SVM, DT
Other: Critical Care	Dias 2014 [86]	**EQ-5D**	Bayesian Networks
Other: Rheumatoid Arthritis	Duong 2022 [87]	HAQ	LASSO, RF
Other: Respiratory	Finnegan 2023 [88]	**Dyspnoea-12**, Centre for Epidemiologic Studies Depression Scale, Trait Anxiety Inventory, Fatigue Severity Scale, St George's Respiratory Questionnaire, Breathlessness Catastrophizing Scale and Vigilance Scale	Feature selection: elastic net procedure with ranked coefficients; classification: Support Vector Classifier with radial kernel

Table 2 (continued)

Theme	First Author & Year	PROMs considered	ML techniques
Other: Cardiac	Frodi 2021 (protocol) [89]	EQ-5D-5 L, Kansas City Cardiomyopathy Questionnaire	RF, K-Neighbors Classifier, Gradient Boosting Classifier, AdaBoost Classifier, Support Vector Classifier, Long Short-Term Memory NN
Other: COVID-19	Gentilotti 2023 [86]	**SF-36**	Principal Component Analysis, LR
Other: Snake Bites	Gerardo 2020 [90]	**Patient-Specific Functional Scale**	Bayesian Belief Network
Other: Stroke	Liao 2022 [91]	**SIS**, Motor Activity Log, Nottingham Extended Activities of Daily Living	RF, KNN, NN, SVM, LR
Other: Care of the Elderly	Stuckenschneider 2022 (protocol) [92]	short falls efficacy scale, Longitudinal Urban Cohort Ageing Study Functional Ability Index, Physical Activity Scale for the Elderly, Life Space Questionnaire, Depression in Old Age Scale, EQ-5D-3 L	Not stated
Other: Stroke	Thakkar 2020 [93]	SIS, Motor Activity Log	KNN, NN

Bold - main PROMs outcomes of interest; Underscore - significant outcome predictors or best performing ML techniques; Adaptive Boosting (Adaboost), American Shoulder and Elbow Surgeons score (ASES), Alcohol Use Disorders Identification Test (AUDIT), Core Outcome Measures Index (COMI), Common Terminology Criteria for Adverse Events (CTCAE), Disabilities of Arm, Shoulder and Hand score (DASH), Decision Trees (DT), Elastic-Net Penalised Logistic Regression (ENPLR), Elastic Net Regression (ENR), European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire– Core Questionnaire (EORTC QLQ-C30), EuroQoL 5-Dimension (EQ-5D), Fear Avoidance Belief Questionnaire (FABQ), Fear-Avoidance Components Scale (FACS-D), Generalised Anxiety Disorder (GAD), Gradient Boosting (GBoost), Generalised Linear Mixed Model (GLMM), Gaussian Naïve Bayes (GNB), Hospital Anxiety and Depression Scale (HADS), Health Assessment Questionnaire (HAQ), Hip Disability and Osteoarthritis Outcome Score (HOOS), Hip Outcome Score (HOS), Injustice Experience Questionnaire (IEQ), Hip Outcome Tool (iHOT), International Knee Documentation Committee (IKDC) subjective form, Japanese Orthopedic Association Back Pain Evaluation Questionnaire (JOABPEQ), K-Nearest Neighbor (KNN), Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Outcome Survey–Activities of Daily Living (KOS-ADL), Least Absolute Shrinkage and Selection Operator regression (LASSO), Logistic Regression (LR), MD Anderson Symptom Inventory (MDASI), Modified Harris Hip Score (mHHS), Michigan Hand outcomes Questionnaire (MHQ), modified Japanese Orthopedic Association scale (mJOA), Modified Somatic Perception Questionnaire (MSPQ), Multivariate Adaptive Regression Splines (MuARS), Neck Disability Index (NDI), Neural Networks (NN), Oswestry Disability Index (ODI), Oxford Hip Score (OHS), Oxford Knee Score (OKS), Oxford Shoulder Score (OSS), Pain Catastrophizing Scale (PCS), Patient Health Questionnaire (PHQ), Parkinson's Disease Questionnaire 39 (PDQ-39), Patient-Reported Outcomes Measurement Information System (PROMIS), Pain Self Efficacy Questionnaire (PSEQ), Quadratic Discriminant Analysis (QDA), Quick Inventory of Depressive Symptomatology (QIDS), Quality of Life in Epilepsy Inventory-10 (QOLIE-10), Random Forest (RF), Roland Morris Disability Questionnaire (RMDQ), Self Efficacy Scale (SES), Short Form (SF), Stochastic Gradient Descent (SGD), Stroke Impact Scale (SIS), Shoulder Pain and Disability Index (SPADI), Support Vector Machine (SVM), Tampa Scale for Kinesiophobia (TSK-17), University of California at Los Angeles shoulder score (UCLA), Veterans RAND 12-item Health Survey (VR-12), WHO Disability Assessment Schedule (WHODAS), Western Ontario and McMaster University Osteoarthritis Index (WOMAC), Western Ontario Rotator Cuff index (WORC), Extreme Gradient Boosting (XGBoost)

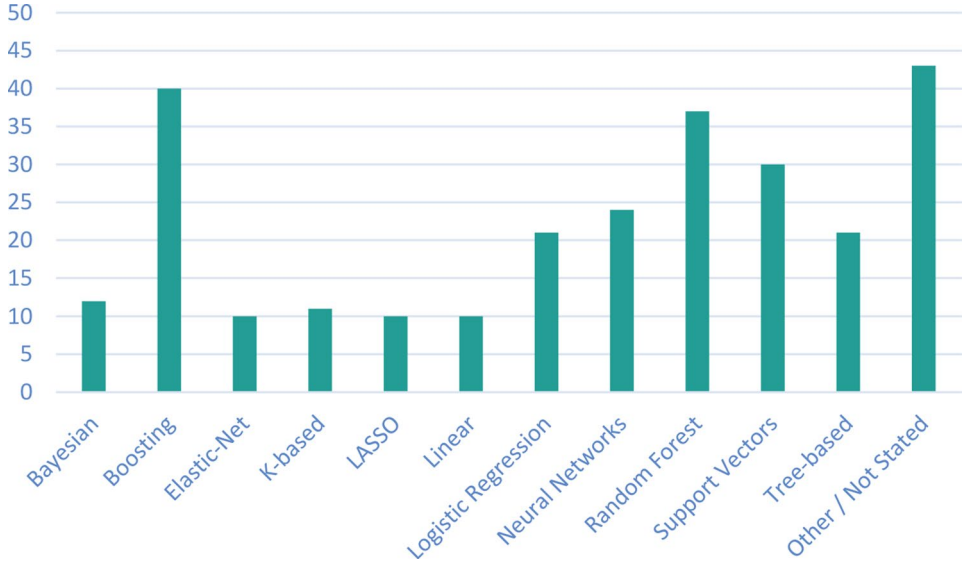

Fig. 3 The number of times an ML technique was investigated in the included studies

pertain to similar conditions their findings are very heterogeneous. For example, four studies that were undertaken by Kunze and colleagues utilised similar PROM tools and tested a similar selection of ML algorithms, but different algorithms were found to perform best in these studies [22–25]. Moreover, different models might perform best for predicting the minimal clinically important difference (the smallest improvement which would be considered worthwhile), patient acceptable syndrome state achievement (achieving a PROM outcome which patients deem acceptable), and substantial clinical benefit (achieving a PROM score change which patients

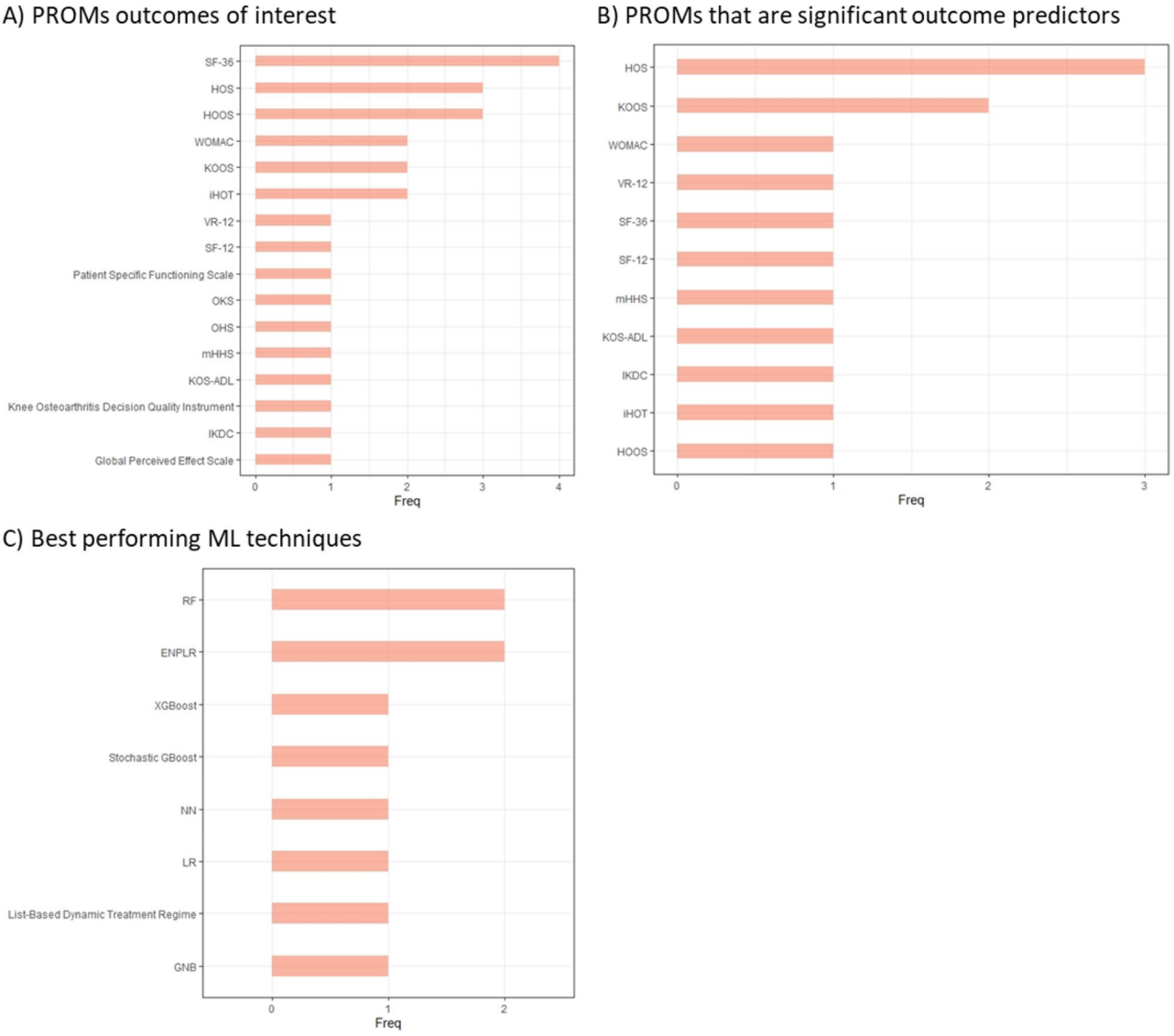

Fig. 4 Key PROMs and ML techniques in studies of hip and knee patients. (**A**) PROMs that studies identified as outcomes of interest (**B**) PROMs that studies identified as significant outcome predictors (**C**) ML techniques that studies highlighted as best performing when more than one ML technique was investigated. Elastic-Net Penalised Logistic Regression (ENPLR), Gaussian Naïve Bayes (GNB), Hip Disability and Osteoarthritis Outcome Score (HOOS), Hip Outcome Score (HOS), Hip Outcome Tool (iHOT), International Knee Documentation Committee (IKDC), Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Outcome Survey–Activities of Daily Living (KOS-ADL), Logistic Regression (LR), Modified Harris Hip Score (mHHS), Neural Networks (NN), Oxford Hip Score (OHS), Oxford Knee Score (OKS), Random Forest (RF), Short Form (SF), Veterans RAND 12-item Health Survey (VR-12), Western Ontario and McMaster University Osteoarthritis Index (WOMAC), Extreme Gradient Boosting (XGBoost)

deem significant), and sometimes these ML models do not perform better than simple PROM thresholds [25, 35]. Lastly, in one study which assumed a 4-year pre-operative horizon, the authors found that PROMs only indicated a significant worsening one year before surgery, while radiographic data provided earlier indications of deterioration [16]. However, another study did not find radiographic data to be able to predict PROMs outcomes [28, 29].

Spinal conditions

There were 13 records relating to 12 studies which fell into the spinal conditions category. Two studies looked at predicting patient outcomes one year after lumbar spinal stenosis surgery [47, 48]. One record described a study that used radiomics to predict patients' post-operative outcomes, and used fourteen ML classifiers and seven feature processors [50]. A second reported a study looking at predicting 3-months post-surgery quality of life [46]. A third record looked at 6-,12- and 24-month post-surgical outcomes [44]. Two records, related to a study

looking at the outcomes of lumbar spondylitis patients [39, 40]. Two records were of studies looking at adults with spinal deformity undergoing surgery. One study adopted a 2-year outcome horizon, and did not report on any notable outcome predictors [38]. The other study used a 1-year outcome horizon to predict operative vs non-operative management [41]. One record looked at 1- and 2-year postoperative outcomes after lumbar spinal fusion [42]. One record looked at cervical radiculopathy patients [43]. One record looked at the outcomes of patients who underwent decompression surgery for lumbar spinal canal stenosis [49]. One record looked at predicting post-surgical outcomes in patients with degenerative spinal disorders [45].

There were 15 different PROM tools that were investigated in these studies. The most frequently studied PROM tool was the Oswestry Disability Index (six times), followed by the Short Form-36 and modified Japanese Orthopaedic Association scale, both used three times (Table 2). Boosting, Random Forest, and Support Vector Machines were the most often tested ML approaches (Table 2). The modified Japanese Orthopedic Association scale was the most studied PROM outcome of interest, while the Short Form-36 and Oswestry Disability Index were the most frequently identified PROM tools that were significant predictors of the studied outcomes (Fig. 5). All of the ML techniques which were highlighted as being best performing were only mentioned once in the identified studies (Fig. 5). One study reported that various models had similar performance, so the authors highlighted the most parsimonious of these models [43]. For context, parsimonious models rely on a smaller number of variables, and as such are computationally more efficient and require less data collection. In another case, the authors created an ensemble of the five best models [49]. One study identified two phenotypes of patients: those of high and intermediate disease burden. It found that those with high disease burden demonstrated a greater 24-month horizon improvement on many measures compared to intermediate burden patients, though the higher burden patients had lower satisfaction [39, 40]. Another study found that Modified Somatic Perception Questionnaire and Self Efficacy Scale scores were important predictors of EuroQoL 5-Dimension (EQ-5D) scores, but the baseline EQ-5D score was not [43]. One study failed to make any reliable predictions [46].

Other musculoskeletal conditions

There were 10 records relating to 10 studies which described other musculoskeletal conditions. Out of these, three were protocols. Three studies looked at back pain patients. One record was of a study looking at predicting patients' response to acupuncture [53]. One looked at several outcome measures, but it contained some uncertainty regarding important predictors due to unexplained acronyms [58]. The last study, looked at a range of PROMs and considered seven regression algorithms and two classification algorithms [59]. Three studies looked at patients with shoulder problems. One looked at identifying subgroups of patients undergoing arthroscopic rotator cuff repair [56]. The other two studies looked at outcomes after shoulder arthroplasty. One of these looked at outcomes 2 years post-intervention [57]. The other study considered a range of time points (1 year, 2–3 years, 3–5 years, and more than 5 years post-intervention) for its outcomes [54]. One study looked at predicting outcomes after surgery for thumb carpometacarpal osteoarthritis [55]. One protocol described a study looking at predicting rotator cuff surgery outcomes [51]. The second protocol record described a study looking at predicting fracture outcomes [52]. The last was a protocol of a study to identify phenotypes of lower back pain [60].

As this is a heterogonous collection of studies, information on the most frequently used PROM tools and ML approaches is not provided. The one notable finding is that one study found the pre-intervention American Shoulder and Elbow Surgeons and Shoulder Pain and Disability Index scores, but not the Constant-Murley Score, to be predictive of the post-intervention Constant-Murley Score [54].

Cancer

There were 12 records which pertained to nine cancer studies. There were three records which related to two studies including cancer patients in general. One study looked at post-surgery complications in patients suffering from gastrointestinal and lung cancer [68]. Two records related to a study looking at immune-related adverse events in cancer patients receiving immune checkpoint inhibitor therapies [64, 65]. Five records looked specifically at patients suffering from breast cancer. Three of these records were published by members of the same group, utilising patient data from the Mastectomy Reconstruction Outcomes Consortium and looking at one and two year outcomes after breast surgery [70–72]. One record described a study that considered whether using a ML tool improved clinicians' ability to predict breast cancer patients' post-treatment quality of life [66]. One study looked at predicting adverse events in breast cancer patients [62]. Three records related to studies concerning prostate cancer patients. Out of these three studies, one study also looked at breast cancer patients, but did not report data relevant to this review with respect to the breast cancer patients [69]. One record looked at predicting adverse events [63]. Another record looked at outcomes in prostate cancer patients receiving stereotactic body radiation therapy [67]. One record, looked at

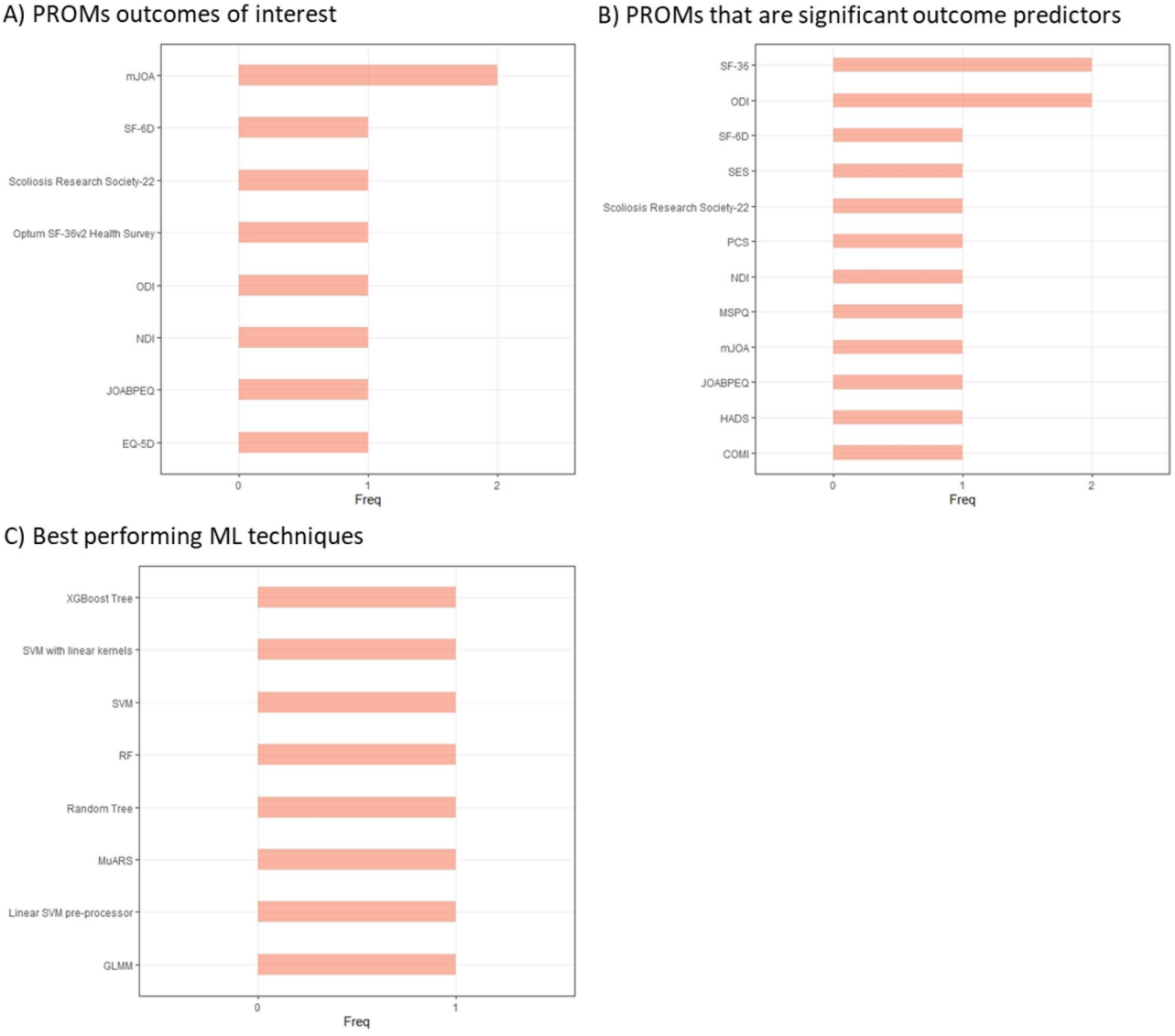

Fig. 5 Key PROMs and ML techniques in studies of spinal condition patients. (**A**) PROMs that studies identified as outcomes of interest (**B**) PROMs which studies identified as significant outcome predictors (**C**) ML techniques that studies highlighted as best performing when more than one ML technique was investigated. Core Outcome Measures Index (COMI), EuroQoL 5-Dimension (EQ-5D), Generalised Linear Mixed Model (GLMM), Hospital Anxiety and Depression Scale (HADS), Japanese Orthopedic Association Back Pain Evaluation Questionnaire (JOABPEQ), modified Japanese Orthopedic Association scale (mJOA), Modified Somatic Perception Questionnaire (MSPQ), Multivariate Adaptive Regression Splines (MuARS), Neck Disability Index (NDI), Oswestry Disability Index (ODI), Pain Catastrophizing Scale (PCS), Random Forest (RF), Self Efficacy Scale (SES), Short Form (SF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost)

predicting survival in patients suffering from metastatic non-small cell lung cancer [61].

Ten unique PROM tools were utilised in these studies, of these Common Terminology Criteria for Adverse Events was utilised three times and the European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire– Core Questionnaire twice; all other tools were only utilised once (Table 2). Boosting, Logistic Regression, Neural Networks and Random Forests were the most often investigated techniques, each utilised in four studies (Table 2). None of the PROMs which were used as predictors of outcome measures featured more prominently than others, while XGBoost was the most frequently mentioned best performing ML technique (Fig.6). One study, uniquely out of all the studies included in this review, compared centralised and federated models [69]. It found that centralised and federated models performed similarly in predicting short-term quality of life, but that centralised models performed better in making long-term predictions. Authors of one of the studies highlighted that various models performed comparatively [70–72]. Finally, one

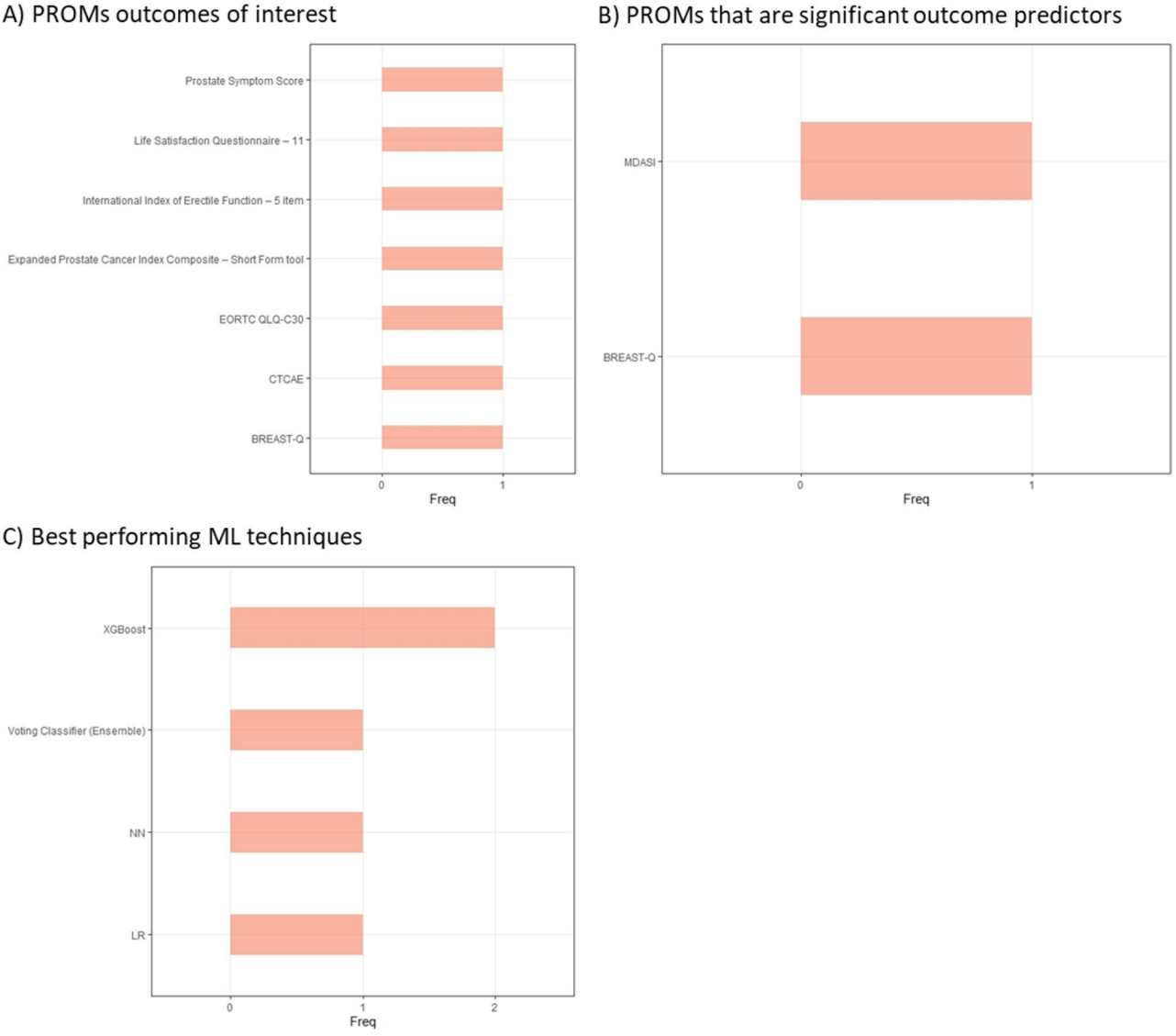

Fig. 6 Key PROMs and ML techniques in studies of cancer patients. (**A**) PROMs that studies identified as outcomes of interest (**B**) PROMs which studies identified as significant outcome predictors (**C**) ML techniques that studies highlighted as best performing when more than one ML technique was investigated. Common Terminology Criteria for Adverse Events (CTCAE), European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire– Core Questionnaire (EORTC QLQ-C30), Logistic Regression (LR), Neural Networks (NN), MD Anderson Symptom Inventory (MDASI), Extreme Gradient Boosting (XGBoost).

study reported higher accuracy in post-treatment quality of life with the aid of an ML technology, but they noted that the 95% confidence intervals do overlap between the aided and unaided groups [66].

Neurodegenerative conditions

There were four records which reported on four studies looking into neurodegenerative conditions, including one protocol. Two studies looked at patients with Parkinson's disease [73, 76]. One study looked at patients with multiple sclerosis and highlighted the utility of PROMs, but did not provide any measure of statistical certainty [74]. The one protocol was of a study looking at patients with spinal muscular atrophy [75].

Seven unique PROM tools were used in these studies, with the Parkinson's Disease Questionnaire 39 being used twice and all other tools only once (Table 2). The Parkinson's Disease Questionnaire 39 was the only PROM tool which was highlighted as an outcome measure of interest and a significant predictor of the studies' outcomes (Table 2). There was no ML technique that was more popular than any other, but Long Short-Term Memory- recurrent NN was the only technique which was highlighted as best performing (Table 2). There were no special observations relating to these studies.

Mental health

There were seven records relating to seven mental health studies. Two studies looked at patients with depression in general [77, 79]. One highlighted PROMs as important baseline features in predicting both quality of life and costs associated with usual and blended therapy treatments [77]. The other looked at data from two trials of depression treatment [79]. One study looked at depressive symptoms in epilepsy patients [78]. One study looked at the risk factors of poor mental health outcomes in outpatients managed for COVID-19 [80]. One study looked at depressive symptoms in patients with opioid use disorders with the aim of predicting re-admission [81]. One study, looked at potential responders to a pharmacological agent studied for its application in the care of patients with alcohol use or major depressive disorders [83]. One study looked at psychological resilience in breast cancer patients to develop a clinical decision support tool [82]. It is a sibling study of one of the studies discussed in the cancer section [66].

Across these studies 21 unique PROM tools were used, with Quick Inventory of Depressive Symptomatology and Patient Health Questionnaire-9 both used three times, and EQ-5D twice; all other tools were only used once (Table 2). Of the most frequently investigated ML methodologies, four studies investigated Random Forest methodology, three boosting, while K-Nearest Neighbor and Logistic Regression were both looked at twice (Table 2). Quick Inventory of Depressive Symptomatology and Patient Health Questionnaire-9 were the most investigated PROM outcome measures of interest and most frequently identified PROMs tools that were significant predictors of the studied outcomes (Fig. 7). No ML technique stood out as the most frequently best performing approach (Fig. 7). One study found that PROM scores were important predictors in some, but not all ML models predicting patient re-admission [81]. Another study noted patients with PROM scores indicating more depressive symptoms but better subjective quality of life responded best to treatment [78].

Other conditions

There were 11 records which related to 11 studies of a range of conditions that did not fit into any of the previously described categories, with two of these, being protocols. Two records pertained to studies looking at stroke patients undergoing such therapies as robot-assisted therapy and mirror therapy. One study described the range of explored therapies as 'contemporary task-oriented' [93]. The other looked at 'sensorimotor rehabilitation interventions' [91]. Two records described studies looking at the effect of different treatment regimens on patients suffering from rheumatoid arthritis. One study compared therapy with golimumab to therapy with infliximab, using data from a pragmatic trial to look at disease activity [85]. The second study looked at patient response to methotrexate treatment [87]. A single study looked at predicting breathlessness improvement using functional brain imaging [88]. One study explored the use of ML in the critical care setting and identified a range of risk factors for outcome after 6-weeks and 6-months [86]. One study looked at outcomes in patients suffering from post-COVID-19 syndrome [94]. A single study looked at predicting outcomes after bariatric surgery [84]. The authors did not provide any measure of statistical certainty, and as such it is not possible to comment how important any of these measures were as predictors. Finally, one study considered cytokine response and patient recovery after snake bites [90]. One protocol described a study on fall related emergencies in the care of the elderly [92]. The other protocol looked at a study predicting arrhythmic events and cardioverter-defibrillator therapy [89].

As this is a heterogonous collection of studies, information on the most frequently used PROM tools and ML approaches is not provided.

Discussion

ML approaches to PROMs

One notable observation from the identified studies was that there was no clear ML approach which appeared to be more effective at predicting outcomes. Consider for example Kunze's hip arthroscopy studies [22–25]. While they all evaluated data from patients with similar clinical indications and utilised similar PROM tools, a variety of ML approaches have been found to provide the best models in each study. Additionally, across the reported studies, researchers used a broad range of ML approaches, with studies often testing more than one approach and no one technique emerging as the preferred ML methodology across studies. These two observations suggest that researchers wishing to develop models for use in their own institutions should not solely rely on copying the approach which was reported to provide the best model in any previous study.

This review provides a summary of the ML techniques that have been previously used in combination with PROMs data. Researchers looking to apply ML techniques in their clinical settings can see from Table 2 which techniques proved to be most successful in the past in their clinical areas or with the PROMs that are currently collected at their institutions. This can help focus the research effort of those who only have the resources to investigate a limited range of ML techniques in their practice, but do not want to just copy a best-performing past approach. Nevertheless, it is likely that obtaining high quality reliable data is likely to be the biggest challenge when developing such ML models.

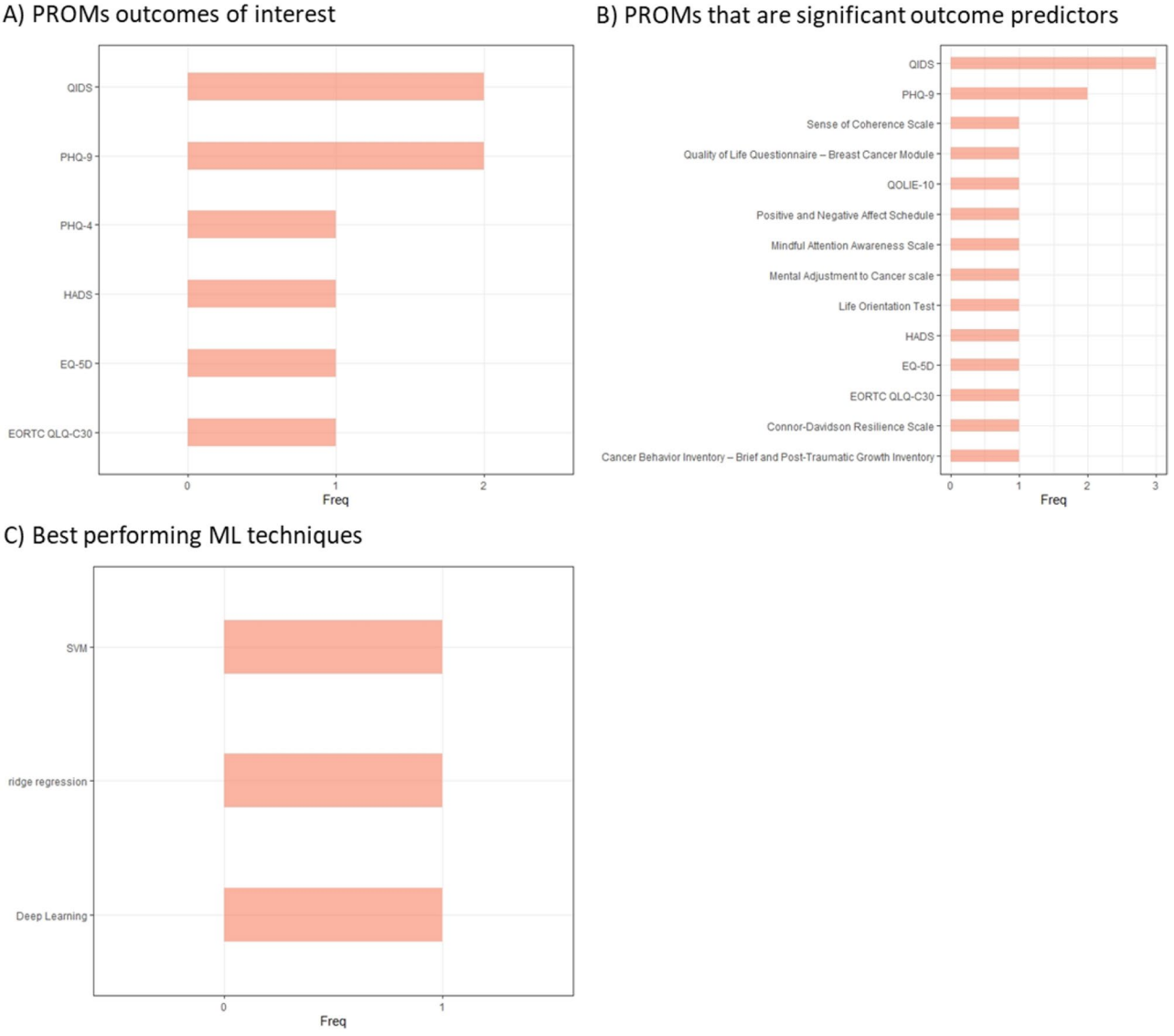

Fig. 7 Key PROMs and ML techniques in studies of mental health patients. (**A**) PROMs that studies identified as outcomes of interest (**B**) PROMs which studies identified as significant outcome predictors (**C**) ML techniques that studies highlighted as best performing when more than one ML technique was investigated. European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire– Core Questionnaire (EORTC QLQ-C30), EuroQoL 5-Dimension (EQ-5D), Hospital Anxiety and Depression Scale (HADS), Patient Health Questionnaire (PHQ), Quick Inventory of Depressive Symptomatology (QIDS), Quality of Life in Epilepsy Inventory-10 (QOLIE-10), Support Vector Machine (SVM).

This review did not present any specific performance metrics for the models described in the identified studies. There are several key reasons for this. The primary aim of the review was to identify what has been done previously in relation to PROMs and ML, rather than to perform an evidence synthesis to assess the specific performance of the identified models. It is also not clear what performance threshold is good enough for a model, and as such models are best considered within the context of specific clinical contexts, rather simple summaries. For example, where an ML model might be used to help decide whether to give a patient treatment X or treatment Y, the degree of confidence we might wish to have in a model will depend on the risk and benefit profiles of both interventions. When considering two medications with similar risk profiles and where treatment can be easily changed from one to another, a clinician might be content to accept the advice of a worse performing model than when deciding whether or not to amputate a limb.

Pre-operative PROMs

A theme which emerged amongst the included records was that often one of the most influential predictors of a post-intervention PROM score was the pre-intervention PROM score; of the 37 studies that used at least one PROM tool as its outcome score (i.e. excluding protocols

and studies that only utilised PROMs as predictors), 22 studies reported at least one post-intervention PROM score to have its pre-intervention counterpart as an important predictor. This suggests that the baseline well-being of a patient is the best predictor of their post-intervention wellbeing. Yet, there is a need to be careful not to conflate this with the impact of an intervention. For example, one study looking at lumbar spondylitis patients noted that the benefit of the intervention was greater in the subset of patients that were regarded as being in a worse health condition at the start of the study [39, 40]. The patients that might benefit most from the intervention might not be the same as those that will have the best PROM scores after it. Moreover, not all post-intervention PROM scores were best predicted by their pre-intervention counterparts. While this suggests that collection of pre-intervention PROM scores might be helpful when predicting patient outcomes, these scores will not always prove useful in such endeavours.

Study limitations

While this review utilised a comprehensive search of the literature, it is affected by a range of limitations. Abstracts often do not report all the variables assessed in a study. This means that studies might have been wrongly rejected during the abstract sift if the abstract omitted to indicate that PROMs had been utilised in a study. This might have particularly affected studies which utilised a range of variables, but did not report these in detail in the abstract. Considering the systematic search, a wide range of terms were used to identify relevant publications. However, some studies may have been missed where relevant concepts were described using alternative free text terms or controlled vocabulary. This might have been the case with the aforementioned review, though it is also possible that the outcomes reported by some of these publications did not align with our review scope [26]. Moreover, the review looked at the application of ML and PROMs to predicting post-intervention outcomes from pre-intervention data, where PROMs have been either a potential predictor and/or outcome measure. As such it excluded studies which looked at the diagnostic ability of ML applications utilising PROMs data, or predicting long-term outcomes from short-term post-intervention data. Due to the extensive nature of identified studies, the review did not report on other predictor or outcome factors. The review did also not report specific model performance, as this information is of little relevance without appreciating the broader clinical context of each ML model's use. Finally, to focus the review, we did not look at studies assessing care-giver wellbeing or where the outcome of interest was financial well-being.

Further research, through larger scale studies or meta-analysis might help to identify best performing ML techniques as well as PROMs that are most suitable for use in ML models. Nevertheless, the choice of PROMs tools might be dictated by other factors, such as their use in clinical practice or historical adoption reasons, and consequently the fact that studies might have reported success in using specific PROMs with ML techniques might not represent a strong enough incentive for adoption in clinical practice. Similarly, the choice of which ML techniques to use might also relate to whether one is interested in a regression problem or classification problem, and what data one has available to be used in a potential model. Consequently, such information can provide useful pointers to researchers, clinicians and healthcare decision-makers, but is unlikely to replace local evaluation of various models. Finally, it is important to remember that new PROMs and ML techniques might be developed, and such tools will need to be evaluated and considered in future research.

Conclusions

This review summarised 82 records describing 73 studies that predicted patient post-intervention outcomes using a combination of ML techniques and PROM tools, where PROMs were either used as a predictor or considered as an outcome. The biggest group of identified studies related to orthopaedics, particularly to hip and knee surgery. Even when authors studied patients with similar conditions, they often employed a range of PROM tools and ML techniques. The variety of approaches used and results of these studies, suggest that while it might be possible to develop clinically useful models, there is no one best ML technique. Those wishing to implement ML-based decision support tools should evaluate their data using a wide range of approaches to see which perform best, rather than simply replicating a published model.

Abbreviations

Adaboost	Adaptive Boosting
AI	Artificial Intelligence
ASES	American Shoulder and Elbow Surgeons score
AUDIT	Alcohol Use Disorders Identification Test
COMI	Core Outcome Measures Index
CTCAE	Common Terminology Criteria for Adverse Events
DASH	Disabilities of Arm, Shoulder and Hand score
DT	Decision Trees
ENPLR	Elastic-Net Penalised Logistic Regression
ENR	Elastic Net Regression
EORTC QLQ-C30	European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire: Core Questionnaire
EQ-5D	EuroQoL 5-Dimension
FABQ	Fear Avoidance Belief Questionnaire
FACS-D	Fear-Avoidance Components Scale
GAD	Generalised Anxiety Disorder
GBoost	Gradient Boosting
GLMM	Generalised Linear Mixed Model
GNB	Gaussian Naïve Bayes
HADS	Hospital Anxiety and Depression Scale
HAQ	Health Assessment Questionnaire
HOS	Hip Outcome Score
IEQ	Injustice Experience Questionnaire

iHOT	Hip Outcome Tool
IKDC	International Knee Documentation Committee subjective form
JOABPEQ	Japanese Orthopedic Association Back Pain Evaluation Questionnaire
KNN	K-Nearest Neighbor
KOOS	Knee Injury and Osteoarthritis Outcome Score
KOS-ADL	Knee Outcome Survey–Activities of Daily Living
LR	Logistic Regression
MDASI	MD Anderson Symptom Inventory
ML	Machine Learning
mHHS	Modified Harris Hip Score
MHQ	Michigan Hand outcomes Questionnaire
mJOA	modified Japanese Orthopedic Association scale
MSPQ	Modified Somatic Perception Questionnaire
MuARS	Multivariate Adaptive Regression Splines
NDI	Neck Disability Index
NN	Neural Networks
ODI	Oswestry Disability Index
OHS	Oxford Hip Score
OKS	Oxford Knee Score
OSS	Oxford Shoulder Score
PCS	Pain Catastrophizing Scale
PHQ	Patient Health Questionnaire
PDQ-39	Parkinson's Disease Questionnaire 39
PROMIS	Patient-Reported Outcomes Measurement Information System
PSEQ	Pain Self Efficacy Questionnaire
PROMs	Patient Reported Outcome Measures
QDA	Quadratic Discriminant Analysis
QIDS	Quick Inventory of Depressive Symptomatology
QOLIE-10	Quality of Life in Epilepsy Inventory-10
RF	Random Forest
RMDQ	Roland Morris Disability Questionnaire
SES	Self Efficacy Scale
SF	Short Form
SGD	Stochastic Gradient Descent
SIS	Stroke Impact Scale
SPADI	Shoulder Pain and Disability Index
SVM	Support Vector Machine
TSK-17	Tampa Scale for Kinesiophobia
UCLA	University of California at Los Angeles shoulder score
VR-12	Veterans RAND 12-item Health Survey
WHODAS	WHO Disability Assessment Schedule
WOMANC	Western Ontario and McMaster University Osteoarthritis Index
WORC	Western Ontario Rotator Cuff index
XGBoost	Extreme Gradient Boosting

Supplementary information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-025-03083-8.

Supplementary Material 1

Acknowledgements
We thank Robert Palmer, Andrew Brass and Frances Hooley for advice on this project. We would also like to thank Meg Kiseleva and Rebecca Hughes for their practical assistance.

Author contributions
MP conceived the idea of the project and discussed with KW. KW supervised the project. MP designed and executed the searches with help from SW. MP reviewed all the retrieved records, with SW acting as the second reviewer. MP extracted the data and SW checked its veracity. MP drafted the manuscript, with SW and KW commenting on the drafts.

Funding
This project is supported by the Welsh Value in Health Centre, now part of NHS Wales Performance and Improvement. MP is also supported by Health Education and Improvement Wales, as this project is undertaken as part of his specialist training and for a DClinSci award at the University of Manchester. Cardiff University provided funds for publishing this article Open Access. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 7 August 2024 / Accepted: 23 June 2025
Published online: 04 July 2025

References

1. World Economic Forum. Global Innovation Hub for Value in Healthcare, World Economic Forum. https://initiatives.weforum.org/global-coalition-for-value-in-healthcare/global-innovation-hub. Accessed 13 November 2024.
2. NHS Wales. Value in health puts Wales on the map as a global lead in health systems transformation. Value health. https://vbhc.nhs.wales/latest-news/latest-news/value-in-health-puts-wales-on-the-map-as-a-global-lead-in-health-systems-transformation/. Accessed 13 November 2024.
3. Withers K, Palmer R, Lewis S, et al. First steps in PROMs and PREMs collection in Wales as part of the prudent and value-based healthcare agenda. Qual Life Res. 2020;30:3157. https://doi.org/10.1007/s11136-020-02711-2.
4. de BD, van den BM, Ballester M, et al. Assessing the outcomes and experiences of care from the perspective of people living with chronic conditions, to support countries in developing people-centred policies and practices: Study protocol of the international survey of people living with chronic conditions (PaRIS survey). BMJ Open. 2022;12:e061424. https://doi.org/10.1136/bmjopen-2022-061424.
5. Hurst L, Mahtani K, Pluddemann A, et al. Defining value-based healthcare in the NHS. Centre for Evidence-Based Medicine; 2019.
6. Bevan commission. Prudent healthcare principles. Bevan Comm. https://www.bevancommission.org/about/prudent-principles/. Accessed 4 September 2022.
7. Verma D, Bach K, Mork PJ. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: A literature review. Informatics. 2021;8:56. https://doi.org/10.3390/informatics8030056.
8. Pearce FJ, Cruz Rivera S, Liu X, et al. The role of patient-reported outcome measures in trials of artificial intelligence health technologies: A systematic evaluation of ClinicalTrials.Gov records (1997-2022). Lancet Digit Health. 2023;5:e160–7. https://doi.org/10.1016/S2589-7500(22)00249-7.
9. The EndNote team. EndNote. 2013.
10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA, 2020 statement: An updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. https://doi.org/10.1136/bmj.n71.
11. Page MJ, Moher D, Bossuyt PM, et al. PRISMA, 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160. https://doi.org/10.1136/bmj.n160.
12. Ogutu JO, Piepho H-P, Schulz-Streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 2011;5:S11. https://doi.org/10.1186/1753-6561-5-S3-S11.
13. Bini SA, Shah RF, Bendich I, et al. Machine learning algorithms can use wearable sensor data to accurately predict six-week patient-reported outcome scores following joint replacement in a prospective trial. The J Arthroplasty. 2019;34:2242–47. https://doi.org/10.1016/j.arth.2019.07.024.
14. Fontana MA, Lyman S, Sarker GK, et al. Can machine learning algorithms predict which patients will achieve minimally clinically important differences

from total joint arthroplasty? Clin Orthop. 2019;477:1267–79. https://doi.org/10.1097/CORR.0000000000000687.
15. Harris AHS, Kuo AC, Bowe TR, et al. Can machine learning methods produce accurate and easy-to-use preoperative prediction models of one-year improvements in pain and functioning after knee arthroplasty? The J Arthroplasty. 2021;36:112–7.e6. https://doi.org/10.1016/j.arth.2020.07.026.
16. Heisinger S, Hitzl W, Hobusch GM, et al. Predicting total knee replacement from symptomology and radiographic structural change using artificial neural networks-data from the osteoarthritis initiative (OAI). J Clin Med. 2020;9:1298. https://doi.org/10.3390/jcm9051298.
17. Jayakumar P, Bozic KJ. Advanced decision-making using patient-reported outcome measures in total joint replacement. J Orthop Res. 2020;38:1414–22. https://doi.org/10.1002/jor.24614.
18. Jayakumar P, Moore MG, Furlough KA, et al. Comparison of an artificial intelligence-enabled patient decision Aid vs educational material on decision quality, shared decision-making, patient experience, and functional outcomes in adults with knee osteoarthritis: A randomized clinical trial. JAMA Netw Open. 2021;4:e2037107. https://doi.org/10.1001/jamanetworkopen.2020.37107.
19. Jiang X, Nelson AE, Cleveland RJ, et al. Precision medicine approach to develop and internally validate optimal exercise and weight-loss treatments for overweight and obese adults with knee osteoarthritis: Data from a single-Center randomized trial. Arthritis Care Res. 2021;73:693–701. https://doi.org/10.1002/acr.24179.
20. Kastrup N, Bjerregaard HH, Laursen M, et al. An AI-based patient-specific clinical decision support system for OA patients choosing surgery or not: Study protocol for a single-centre, parallel-group, non-inferiority randomised controlled trial. Trials. 2023;24. https://doi.org/10.1186/s13063-022-07039-5.
21. Nct. RCT measuring the effect of the ERVIN software. https://clinicaltrials.gov/show/NCT04332055. Published Online First: 2020.
22. Kunze KN, Polce EM, Rasio J, et al. Machine learning algorithms predict clinically significant improvements in satisfaction after hip arthroscopy. Arthrosc J Arthrosc Relat Surg. 2021;37:1143–51. https://doi.org/10.1016/j.arthro.2020.11.027.
23. Kunze KN, Polce EM, Nwachukwu BU, et al. Development and internal validation of supervised machine learning algorithms for predicting clinically significant functional improvement in a mixed population of primary hip arthroscopy. Arthrosc J Arthrosc Relat Surg. 2021;37:1488–97. https://doi.org/10.1016/j.arthro.2021.01.005.
24. Kunze KN, Polce EM, Clapp I, et al. Machine learning algorithms predict functional improvement after hip arthroscopy for femoroacetabular impingement syndrome in athletes. J Bone Jt Surg. 2021;103:1055–62. https://doi.org/10.2106/JBJS.20.01640.
25. Kunze KN, Polce EM, Clapp IM, et al. Association between preoperative patient factors and clinically meaningful outcomes after hip arthroscopy for femoroacetabular impingement syndrome: A machine learning analysis. Am J Sports Med. 2022;50:746–56. https://doi.org/10.1177/03635465211067546.
26. Lopez CD, Gazgalis A, Boddapati V, et al. Artificial learning and machine learning decision guidance applications in total hip and knee arthroplasty: A systematic review. Arthroplasty Today. 2021;11:103–12. https://doi.org/10.1016/j.artd.2021.07.012.
27. Milella F, Famiglini L, Banfi G, et al. Application of machine learning to improve appropriateness of treatment in an orthopaedic setting of personalized medicine. J Pers Med. 2022;12:1706. https://doi.org/10.3390/jpm12101706.
28. Ramkumar PN, Karnuta JM, Haeberle HS, et al. Radiographic indices are not predictive of clinical outcomes among 1735 patients indicated for hip arthroscopic surgery: A machine learning analysis. Am J Sports Med. 2020;48:2910–18. https://doi.org/10.1177/0363546520950743.
29. Ramkumar PN, Karnuta JM, Haeberle HS, et al. Radiographic indices are not predictive of clinical outcome among 1, 735 patients indicated for hip arthroscopy: A machine learning analysis. J ISAKOS. 2021;6:387–88. https://doi.org/10.1136/jisakos-2021-congress.16.
30. Ramkumar PN, Karnuta JM, Haeberle HS, et al. Effect of preoperative imaging and patient factors on clinically meaningful outcomes and quality of life after osteochondral allograft transplantation: A machine learning analysis of cartilage defects of the knee. Am J Sports Med. 2021;49:2177–86. https://doi.org/10.1177/03635465211015179.
31. Karnuta J, Haeberle H, Owusu-Akyaw K, et al. Pre-operative mental health predicts clinically meaningful outcomes after osteochondral allograft for cartilage defects of the knee: A machine learning analysis. Orthop J Sports Med. 2021;9. https://doi.org/10.1177/2325967121S00217.
32. Ribbons K, Johnson S, Ditton E, et al. Using presurgical biopsychosocial features to develop an advanced clinical decision-making support tool for predicting recovery trajectories in patients undergoing total knee arthroplasty: Protocol for a prospective observational study. JMIR Res Protoc. 2023;12:e48801. https://doi.org/10.2196/48801.
33. Sergooris A, Verbrugghe J, Matheve T, et al. Clinical phenotypes and prognostic factors in persons with hip osteoarthritis undergoing total hip arthroplasty: Protocol for a longitudinal prospective cohort study (HIPPROCLIPS). BMC Musculoskelet Disord. 2023;24:224. https://doi.org/10.1186/s12891-023-06326-9.
34. Sniderman J, Stark RB, Schwartz CE, et al. Patient factors that matter in predicting hip arthroplasty outcomes: A machine-learning approach. The J Arthroplasty. 2021;36:2024–32. https://doi.org/10.1016/j.arth.2020.12.038.
35. Zhang S, Lau BPH, Ng YH, et al. Machine learning algorithms do not outperform preoperative thresholds in predicting clinically meaningful improvements after total knee arthroplasty. Knee Surg, Sports Traumatol, Arthrosc. 2022;30:2624–30. https://doi.org/10.1007/s00167-021-06642-4.
36. Zhou Y, Weeden C, Patten L, et al. Evaluating willingness for surgery using the SMART choice (knee) patient prognostic tool for total knee arthroplasty: Study protocol for a pragmatic randomised controlled trial. BMC Musculoskelet Disord. 2022;23:179. https://doi.org/10.1186/s12891-022-05123-0.
37. Zhou Y, Schilling C, Dowsey M, et al. Development of the proto-knee tool using machine learning algorithms to predict clinical outcomes after total knee arthroplasty. Osteoarthr Cartil. 2022;30:S84. https://doi.org/10.1016/j.joca.2022.02.103.
38. Ames CP, Smith JS, Pellisé F, et al. Artificial intelligence based hierarchical clustering of patient types and intervention categories in adult spinal deformity surgery: Towards a New classification scheme that predicts quality and value. Spine. 2019;44:915–26. https://doi.org/10.1097/BRS.0000000000002974.
39. Chan AK, Wozny TA, Bisson EF, et al. Classifying patients operated for spondylolisthesis: A K-Means clustering analysis of clinical presentation phenotypes. Neurosurgery. 2021;89:1033–41. https://doi.org/10.1093/neuros/nyab355.
40. Chan W, Bisson, et al. 113 clinical presentation phenotypes of patients operated for lumbar spondylolisthesis: An analysis of the quality outcomes database. Neurosurgery. 2022;68:31–31. https://doi.org/10.1227/NEU.0000000000001880_111.
41. Durand WM, Daniels AH, Hamilton DK, et al. Artificial intelligence models predict operative versus nonoperative management of patients with adult spinal deformity with 86% accuracy. World Neurosurg. 2020;141:e239–53. https://doi.org/10.1016/j.wneu.2020.05.099.
42. Janssen ER, Osong B, Van Soest J, et al. Exploring associations of preoperative physical performance with postoperative outcomes after lumbar spinal fusion: A machine learning approach. Arch Phys Med Rehabil. 2021;102:1324–30.e3. https://doi.org/10.1016/j.apmr.2021.02.013.
43. Liew BXW, Peolsson A, Rugamer D, et al. Clinical predictive modelling of post-surgical recovery in individuals with cervical radiculopathy: A machine learning approach. Sci Rep. 2020;10:16782. https://doi.org/10.1038/s41598-020-73740-7.
44. Merali ZG, Witiw CD, Badhiwala JH, et al. Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy. PLoS One. 2019;14:e0215133. https://doi.org/10.1371/journal.pone.0215133.
45. Muller D, Haschtmann D, O'Riordan D, et al. Development of a machine-learning based model for predicting multidimensional outcome after surgery for degenerative disorders of the spine (9999). Swiss Med Wkly. 2021;151:17S.
46. Rogers S, Jacobs B, Bouchard J, et al. Machine learning to predict a single patient clinical course: How will your life change after a diagnosis of degenerative cervical myelopathy? CMAJ Can Med Assoc J. 2019;62:S82. https://doi.org/10.1503/cjs.010919.
47. Siccoli A, De Wispelaere MP, Schröder ML, et al. Machine learning-based preoperative predictive analytics for lumbar spinal stenosis. Neurosurg Focus. 2019;46:E5. https://doi.org/10.3171/2019.2.FOCUS18723.
48. Sundararajan K, Shestopaloff K, Lane K, et al. Development and validation of a surgical clinical decision support tool for lumbar spinal stenosis. CMAJ Can Med Assoc J. 2019;62:S74–5. https://doi.org/10.1503/cjs.010919.
49. Yagi M, Michikawa T, Yamamoto T, et al. Development and validation of machine learning-based predictive model for clinical outcome of decompression surgery for lumbar spinal canal stenosis. Spine J. 2022;22:1768–77. https://doi.org/10.1016/j.spinee.2022.06.008.
50. Zhang M, Ou-Yang H, Jiang L, et al. Optimal machine learning methods for radiomic prediction models: Clinical application for preoperative T_2 *-weighted images of cervical spondylotic myelopathy. JOR Spine. 2021;4:e1178. https://doi.org/10.1002/jsp2.1178.

51. Allaart LJH, Spanning SV, Lafosse L, et al. Developing a machine learning algorithm to predict probability of retear and functional outcomes in patients undergoing rotator cuff repair surgery: Protocol for a retrospective, multicentre study. BMJ Open. 2023;13:e063673. https://doi.org/10.1136/bmjopen-2022-063673.
52. Dipnall JF, Page R, Du L, et al. Predicting fracture outcomes from clinical registry data using artificial intelligence supplemented models for evidence-informed treatment (PRAISE) study protocol. PLoS One. 2021;16:e0257361. https://doi.org/10.1371/journal.pone.0257361.
53. Kong J-T, Tian L, Manber R, et al. Development and validation of a prediction model for response to acupuncture in treating back pain using machine-learning: Results from 2 independent clinical trials. Integr Med Res. 2020;9:100510. https://doi.org/10.1016/j.imr.2020.100510.
54. Kumar V, Roche C, Overman S, et al. What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? Clin Orthop. 2020;478:2351–63. https://doi.org/10.1097/CORR.0000000000001263.
55. Loos NL, Hoogendam L, Souer JS, et al. Machine learning can be used to predict function but not pain after surgery for thumb carpometacarpal osteoarthritis. Clin Orthop. 2022;480:1271–84. https://doi.org/10.1097/CORR.0000000000002105.
56. Lu Y, Berlinberg E, Patel H, et al. Unsupervised machine learning to identify clinically meaningful subgroups in patients undergoing arthroscopic rotator cuff repair. Orthop J Sports Med. 2023;11. https://doi.org/10.1177/23259671233S00165.
57. Polce EM, Kunze KN, Fu MC, et al. Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty. J Shoulder Elb Surg. 2021;30:e290–9. https://doi.org/10.1016/j.jse.2020.09.007.
58. Verma D, Bach K, Mork PJ. Using automated feature selection for building case-based reasoning systems: An example from patient-reported outcome measurements. 2021;282–95. Cambridge, United Kingdom: Springer-Verlag.
59. Verma D, Jansen D, Bach K, et al. Exploratory application of machine learning methods on patient reported data in the development of supervised models for predicting outcomes. BMC Med Inf Decis Mak. 2022;22:227. https://doi.org/10.1186/s12911-022-01973-9.
60. Vo NV, Piva SR, Patterson CG, et al. Toward the identification of distinct phenotypes: Research protocol for the low back pain biological, biomechanical, and behavioral (LB3P) cohort study and the BACPAC mechanistic research Center at the University of Pittsburgh. Pain Med. 2023;24:S36–47. https://doi.org/10.1093/pm/pnad009.
61. Cunha M, Borges AP, Carvalho V, et al. OA02.02 development of machine learning Model to estimate overall survival in patients with advanced NSCLC and ECOG-PS > 1. J Thorac Oncol. 2021;16:S850. https://doi.org/10.1016/j.jtho.2021.08.038.
62. DeWees TA, Golafshar MA, Bhangoo RS, et al. Artificial neural networks utilizing standardly collected electronic healthcare data provide clinically interpretable predictions of patient-reported adverse events for breast cancer. Int J Radiat Oncol Biol Phys. 2020;108:e766–7. https://doi.org/10.1016/j.ijrobp.2020.07.206.
63. Golafshar MA, Bhangoo RS, Petersen M, et al. Clinically interpretable predictions of patient-reported adverse events (PRO-CTCAE) for prostate cancer utilizing artificial neural networks. Int J Radiat Oncol Biol Phys. 2020;108:e911. https://doi.org/10.1016/j.ijrobp.2020.07.540.
64. Iivanainen SME, Ekstrom J, Kataja V, et al. 1841P predicting the onset of immune-related adverse events (irAes) in immune checkpoint inhibitor (ICI) therapies using a machine learning (ML) model trained with electronic patient-reported outcomes (ePros) and lab measurements. Ann Oncol. 2020;31:S1057. https://doi.org/10.1016/j.annonc.2020.08.1488.
65. Iivanainen SME, Ekstrom J, Kataja V, et al. 1876P a combination model of electronic patient-reported outcomes (ePros) and lab measurements in prediction of immune related adverse events (irAes) and treatment response of immune checkpoint inhibitor (ICI) therapies. Ann Oncol. 2020;31:S1068. https://doi.org/10.1016/j.annonc.2020.08.1523.
66. Nuutinen M, Hiltunen A-M, Korhonen S, et al. Aid of a machine learning algorithm can improve clinician predictions of patient quality of life during breast cancer treatments. Health Technol. 2023;13:229–44. https://doi.org/10.1007/s12553-023-00733-7.
67. Qi X, Neylon J, Santhanam A. Dosimetric predictors for quality of life after prostate stereotactic body radiation therapy via deep learning network. Int J Radiat Oncol Biol Phys. 2017;99:S167. https://doi.org/10.1016/j.ijrobp.2017.06.384.
68. Rossi LA, Melstrom LG, Fong Y, et al. Predicting post-discharge cancer surgery complications via telemonitoring of patient-reported outcomes and patient-generated health data. J Surg Oncol. 2021;123:1345–52. https://doi.org/10.1002/jso.26413.
69. Savić M, Kurbalija V, Ilić M, et al. Analysis of machine learning models predicting quality of life for cancer patients. Proceedings of the 13th International Conference on Management of Digital EcoSystems. Virtual Event. Tunisia: ACM; 2021: 35–42.
70. Xu C, Pfob A, Mehrara BJ, et al. Enhanced surgical decision-making tools in breast cancer: Predicting 2-year postoperative physical, sexual, and psychosocial well-being following mastectomy and breast reconstruction (INSPiRED 004). Ann Surg Oncol. 2023;30:7046–59. https://doi.org/10.1245/s10434-023-13971-w.
71. Pfob A, Mehrara BJ, Nelson JA, et al. Towards patient-centered decision-making in breast cancer surgery: Machine learning to predict individual patient-reported outcomes at 1-year follow-up. Ann Surg. 2023;277:e144–52. https://doi.org/10.1097/SLA.0000000000004862.
72. Pfob A, Mehrara BJ, Nelson JA, et al. Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPiRED-001). The Breast. 2021;60:111–22. https://doi.org/10.1016/j.breast.2021.09.009.
73. Bougea A, Derikvand T, Efthymiopoulou E, et al. Artificial neural network predicts sex differences of patients with advanced Parkinson's disease under levodopa-carbidopa intestinal gel. medRxiv. Published Online First: 2023. doi: https://doi.org/10.1101/2023.06.26.23291833.
74. Branco D, Martino BD, Esposito A, et al. Machine learning techniques for prediction of multiple sclerosis progression. Soft Comput. 2022;26:12041–55. https://doi.org/10.1007/s00500-022-07503-z.
75. Coratti G, Antonaci L, Masciocchi C, et al. Map the SMA protocol: A machine-learning based algorithm to predict therapeutic response in spinal muscular atrophy. Neuromuscul Disord. 2023;33:S86. https://doi.org/10.1016/j.nmd.2023.07.099.
76. Rouleau E, Potters W, Pina Fuentes D, et al. Machine learning predicts clinically important improvement in quality of life after STN-DBS in patients with parkinson's disease. Mov Disord. 2020;35:S630–1. https://doi.org/10.1002/mds.28268.
77. Bremer V, Becker D, Kolovos S, et al. Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: Data-Driven Analysis. J Med Internet Res. 2018;20:e10275. https://doi.org/10.2196/10275.
78. Camp EJ, Quon RJ, Sajatovic M, et al. Supervised machine learning to predict reduced depression severity in people with epilepsy through epilepsy self-management intervention. Epilepsy Behav. 2022;127:108548. https://doi.org/10.1016/j.yebeh.2021.108548.
79. Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: A machine learning approach. Lancet Psychiatry. 2016;3:243–50. https://doi.org/10.1016/S2215-0366(15)00471-X.
80. Hufner K, Tymoszuk P, Ausserhofer D, et al. Who is at risk of poor mental health following coronavirus disease-19 outpatient management? Front Med. 2022;9:792881. https://doi.org/10.3389/fmed.2022.792881.
81. Kay B, Maloney M, et al. Predicting factors of Re-hospitalization after medically managed intensive inpatient services in opioid use disorder. 2021.
82. Manikis GC, Simos NJ, Kourou K, et al. Personalized risk analysis to improve the psychological resilience of women undergoing treatment for breast cancer: Development of a machine learning-driven clinical decision support tool. J Med Internet Res. 2023;25:e43838. https://doi.org/10.2196/43838.
83. Martin B, Gao Q, Liu Y, et al. Explainable AI approach reveals treatment responders in a randomized controlled trial of BTRX-246040, a potent and selective NOP receptor antagonist. Neuropsychopharmacology. 2019;44:448. https://doi.org/10.1038/s41386-019-0547-9.
84. Cao Y, Raoof M, Szabo E, et al. Using bayesian networks to predict long-term health-related quality of life and comorbidity after bariatric surgery: A study based on the scandinavian obesity surgery registry. J Clin Med. 2020;9:1895. https://doi.org/10.3390/jcm9061895.
85. Curtis JR, Su Y, Black S, et al. Machine learning applied to patient-reported outcomes to classify physician-derived measures of rheumatoid arthritis disease activity. ACR Open Rheumatol. 2022;4:995–1003. https://doi.org/10.1002/acr2.11499.
86. Dias CC, Granja C, Costa-Pereira A, et al. Using probabilistic graphical models to enhance the prognosis of health-related quality of life in adult survivors of critical illness. 2014 IEEE 27th International Symposium on Computer-Based Medical Systems. New York, NY, USA: IEEE; 2014: 56–61.

87. Duong SQ, Crowson CS, Athreya A, et al. Clinical predictors of response to methotrexate in patients with rheumatoid arthritis: A machine learning approach using clinical trial data. Arthritis Res Ther. 2022;24:162. https://doi.org/10.1186/s13075-022-02851-5.
88. Finnegan SL, Browning M, Duff E, et al. Brain activity measured by functional brain imaging predicts breathlessness improvement during pulmonary rehabilitation. Thorax. 2023;78:852–59. https://doi.org/10.1136/thorax-2022-218754.
89. Frodi DM, Kolk MZH, Langford J, et al. Rationale and design of the SafeHeart study: Development and testing of a mHealth tool for the prediction of arrhythmic events and implantable cardioverter-defibrillator therapy. Cardiovasc Digit Health J. 2021;2:S11–20. https://doi.org/10.1016/j.cvdhj.2021.10.002.
90. Gerardo CJ, Silvius E, Schobel S, et al. Association of a cytokine response network with functional recovery from snakebite envenoming. Toxicon. 2020;182:S15–6. https://doi.org/10.1016/j.toxicon.2020.04.041.
91. Liao W-W, Hsieh Y-W, Lee T-H, et al. Machine learning predicts clinically significant health related quality of life improvement after sensorimotor rehabilitation interventions in chronic stroke. Sci Rep. 2022;12:11235. https://doi.org/10.1038/s41598-022-14986-1.
92. Stuckenschneider T, Koschate J, Dunker E, et al. Sentinel fall presenting to the emergency department (SeFalled) - protocol of a complex study including long-term observation of functional trajectories after a fall, exploration of specific fall risk factors, and patients' views on falls prevention. BMC Geriatr. 2022;22:594. https://doi.org/10.1186/s12877-022-03261-7.
93. Thakkar HK, Liao W, Wu C, et al. Predicting clinically significant motor function improvement after contemporary task-oriented interventions using machine learning approaches. J Neuroeng Rehabil. 2020;17:131. https://doi.org/10.1186/s12984-020-00758-3.
94. Gentilotti E, Górska A, Tami A, et al. Clinical phenotypes and quality of life to define post-COVID-19 syndrome: A cluster analysis of the multinational, prospective ORCHESTRA cohort. eClinicalmedicine. 2023;62:102107. https://doi.org/10.1016/j.eclinm.2023.102107.

Publisher's Note