frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 05 November 2024
DOI 10.3389/frai.2024.1477447

Check for updates

OPEN ACCESS

EDITED BY
Fabrizio Riguzzi,
University of Ferrara, Italy

REVIEWED BY
Sheng-Chieh Lu,
University of Texas MD Anderson Cancer Center, United States
Yuxuan Jin,
Cleveland Clinic, United States

*CORRESPONDENCE
Jeremy A. Balch
✉ Jeremy.balch@surgery.ufl.edu

RECEIVED 24 August 2024
ACCEPTED 18 October 2024
PUBLISHED 05 November 2024

CITATION
Balch JA, Chatham AH, Hong PKW, Manganiello L, Baskaran N, Bihorac A, Shickel B, Moseley RE and Loftus TJ (2024) Predicting patient reported outcome measures: a scoping review for the artificial intelligence-guided patient preference predictor.
Front. Artif. Intell. 7:1477447.
doi: 10.3389/frai.2024.1477447

Predicting patient reported outcome measures: a scoping review for the artificial intelligence-guided patient preference predictor

Jeremy A. Balch[1,2]*, A. Hayes Chatham[2], Philip K. W. Hong[1], Lauren Manganiello[3], Naveen Baskaran[4], Azra Bihorac[4], Benjamin Shickel[4], Ray E. Moseley[4] and Tyler J. Loftus[1]

[1]Department of Surgery, University of Florida, Gainesville, FL, United States, [2]Department of Health Outcomes and Biomedical Informatics, University of Florida, Gainesville, FL, United States, [3]College of Medicine, University of Florida, Gainesville, FL, United States, [4]Department of Medicine, University of Florida, Gainesville, FL, United States

Background: The algorithmic patient preference predictor (PPP) has been proposed to aid in decision making for incapacitated patients in the absence of advanced directives. Ethical and legal challenges aside, multiple practical barriers exist for building a personalized PPP. Here, we examine previous work using machine learning to predict patient reported outcome measures (PROMs) for capacitated patients undergoing diverse procedures, therapies, and life events. Demonstrating robust performance in predicting PROMs for capacitated patients could suggest opportunities for developing a model tailored to incapacitated ones.

Methods: We performed a scoping review of PubMed, Embase, and Scopus using the PRISMA-ScR guidelines to capture studies using machine learning to predict PROMs following a medical event alongside qualitative studies exploring a theoretical PPP.

Results: Sixty-eight studies used machine learning to evaluate PROMs; an additional 20 studies focused on a theoretical PPP. For PROMs, orthopedic surgeries ($n = 33$) and spinal surgeries ($n = 12$) were the most common medical event. Studies used demographic ($n = 30$), pre-event PROMs ($n = 52$), comorbidities ($n = 29$), social determinants of health ($n = 30$), and intraoperative variables ($n = 124$) as predictors. Thirty-four different PROMs were used as the target outcome. Evaluation metrics varied by task, but performance was overall poor to moderate for the best reported scores. In models that used feature importance, pre-event PROMs were the most predictive of post-event PROMs. Fairness assessments were rare ($n = 6$). These findings reinforce the necessity of the integrating patient values and preferences, beyond demographic factors, to improve the development of personalized PPP models for incapacitated patients.

Conclusion: The primary objective of a PPP is to estimate patient-reported quality of life following an intervention. Use of machine learning to predict PROMs for *capacitated* patients introduces challenges and opportunities for building a personalized PPP for *incapacitated* patients without advanced directives.

KEYWORDS

machine learn, palliative care, patient reported outcome, fairness, clinical decision support

Introduction

Machine learning and artificial intelligence-based algorithms are predicting our preferences on a daily basis. Using aggregated data from past actions—such as a purchase, click, or prolonged gaze—we are continuously offered things to buy, watch, and experience. In advertising, their accuracy can exceed 95% in some instances (Assistant et al., 2023). However, algorithmic preference predictors have not yet extended to the more somber, consequential domain of patient medical decision making.

The Patient Preference Predictor (PPP) for incapacitated patients has been debated in the literature for over 10 years (Rid, 2014). Defined as a tool to help clinicians and surrogate decision makers decided on life-sustaining treatment decisions, several authors have more recently proposed using artificial intelligence to gauge patients preferences when they are unable to make decisions for themselves (Biller-Andorno and Biller, 2019; Wendler et al., 2016). The psychosocial, ethical, and legal implications of using static, statistical evidence to predict end-of-life choices are substantial and complex. While it has been shown that they may provide a better indication of patient preferences than estranged family, friends, and court-designated surrogates–whose decisions are unfortunately often no better than chance (Rid and Wendler, 2010; Shalowitz et al., 2006)–these models would still leave myriad concerns related to loss of autonomy, fairness, lack of trust, and reproducibility (Rid, 2014; Jardas et al., 2021; Rid and Wendler, 2014a; Sharadin, 2018).

Many had hoped that widespread adoption of advanced directives would improve end-of-life decision making. Unfortunately, these documents, in addition to being sparsely available, are frequently too limited in scope for highly morbid interventions. They typically describe preferences for cardiopulmonary resuscitation (Do-Not-Resuscitate, DNR), intubation (Do-Not-Intubate, DNI), or hospitalization (Do-Not-Hospitalize, DNH), but fail to account for complex choices around feeding tube placement, prolonged mechanical ventilation, artificial cardiopulmonary support, or any procedure that leads to substantial change in quality of life (Fagerlin and Schneider, 2004; Detering et al., 2010). Moreover, patient preferences are protean. In the case of survival, they are subject to hindsight bias (Becerra Pérez et al., 2016), and in the case of death, are without a ground truth to know whether the patient received the care they wanted (Rid, 2014). The current practice is to hold the last stated desires as that ground truth (Rid, 2014; Jardas et al., 2021).

Artificial intelligence is currently being studied in thousands of predictive tasks in health care (Rajkomar et al., 2018; Rajpurkar et al., 2022). While these include complications and medical outcomes of interest, they are also increasingly focused on predicting Patient Reported Outcomes Measures (PROMs). PROMs reflect patient quality-of-life in a numeric form and may be a more personalized metric, unlike mortality or a complication defined by a diagnostic code (McGlothlin and Lewis, 2014; Weinfurt and Reeve, 2022). There is a small but rapidly growing interest in using pre-intervention variables, including quality-of-life metrics, to predict post-intervention patient perceptions of their care. PROMs are also expanding their presence in national databases, providing rich data sources for predictive tasks (Temple et al., 2024). We consider the PPP to be, at its core, a task of predicting patient-reported outcomes. Therefore, inclusion of PROMs for *capacitated* patients represent a potential ground truth for researchers interested in the feasibility and fairness of predicting preferences of *incapacitated* patients. In other words, if we know with reasonable certainty how a patient of certain characteristics and perceptions of their current quality of life would assess their life post-intervention, we can know whether or not they would prefer the intervention.

In this scoping review, we reconcile the philosophical and ethical debates of predicting *incapacitated* patient preferences with the current applications of machine learning in the real world for *capacitated* ones.

Materials and methods

We searched PubMed, Embase, and Scopus from January 1, 2019 to May 30, 2024 for terms related to machine learning for predicting PROMs to capture the most recent modeling techniques. Since PPPs for incapacitated patients are still theoretical, articles debating the ethical and practical issues of such models were reviewed separately. Search terms are shown in Supplementary file 1 and the PriSMA-ScR checklist is shown in Supplementary file 2. We identified 621 abstracts in the literature, which were reviewed by JAB, AHC, PH, LM, and NB. Cohen Kappa inter agreement scores ranged from 0.35–0.59. Disagreements were reviewed and resolved between the first author and the individual rater without need for arbitration. 115 full texts were reviewed by the first author. Twenty-seven studies were excluded leaving 88 studies for extraction. Eligible studies employed machine learning for a distinct, health-related event (surgical intervention, medical treatment, therapy secession, or diagnosis), and omitted post-event variables to predict the PROM in the outcome analysis. Twenty were theoretical discussions of patient preference predictors and 68 used machine learning to predict PROMs. Article flow is shown in Figure 1.

We extracted separate variables for the two sets of studies. For PROM studies, we gathered information on the study's main findings, independent and dependent variables, data origin (intraoperative, in-patient, out-patient), data quality assessments, data source (single institution, multi-institution, national database, etc.), machine learning techniques, population characteristics, participant count, performance metrics, fairness metrics, and explainability techniques. Data quality was judged according to the TRIPOD+AI guidelines for machine learning tasks; studies were considered "excellent" if TRIPOD-AI or CONSORT-AI guidelines were followed, "good" if the methods described data preprocessing steps, handling missingness, and adjusting for class imbalance, and "fair" if they missed one or more of those qualities. Studies were excluded if methods failed to describe data cleaning, validation, and model development. For ethical studies, we performed a narrative thematic analysis for the

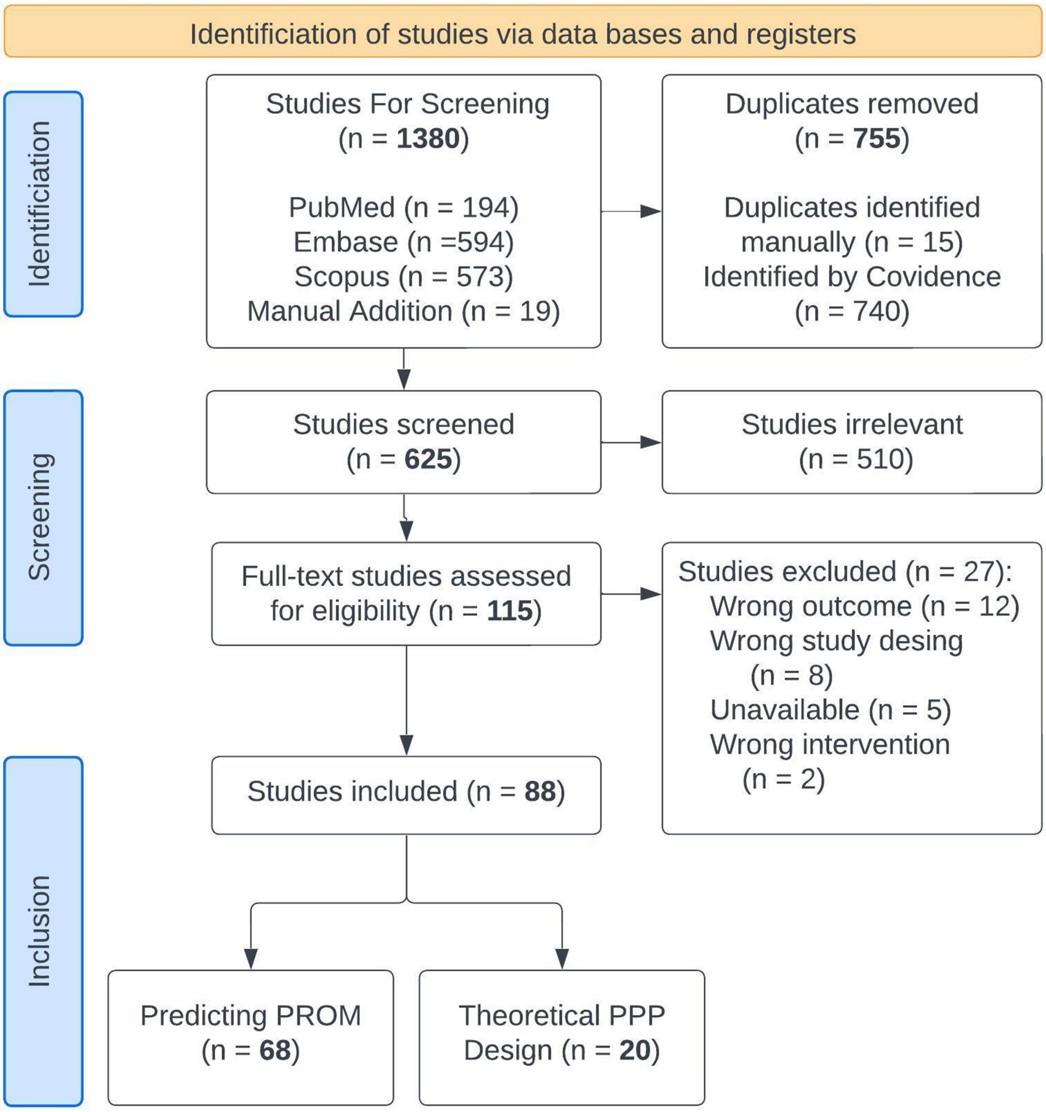

FIGURE 1
Article flow diagram.

ethical and legal principles identified, theoretical model inputs, fairness metrics, and proposed evaluation methods.

We employed Covidence® (Melbourne, Australia) software to manage multiple reviewers. Elicit® (Oakland, California) was used for initial data extraction, followed by manual confirmation and extraction of additional information (Elicit: The AI research Assistant, 2023).

Results

Study characteristics

Sixty-eight studies used machine learning to evaluate PROMs. All studies were retrospective, though three developed a web or smart phone based application (Karhade et al., 2021; Martin et al., 2022; Polce et al., 2021) and one study performed an external validation (Simmons et al., 2024). No studies examined how their findings altered clinical practice. The number of participants ranged from 22[23] to 130,945 (Zrubka et al., 2022). Studies were performed either at a single hospital ($n = 36$), multiple hospitals ($n = 23$), or employed regional or national registries ($n = 8$). As shown in Figure 2, most studies were related to extremity orthopedic surgeries ($n = 33$) or spinal surgeries ($n = 11$), followed by oncology ($n = 8$ for breast; $n = 7$ for head and neck, prostate, and general), and psychotherapy ($n = 5$). Clinical events for before and after comparisons included invasive procedures ($n = 48$), medical and psychological therapy ($n = 7$), diagnoses ($n = 6$), physical therapy ($n = 2$), and a medical device ($n = 1$), with some studies examining surgical and adjuvant therapy for cancer ($n = 3$). A substantial amount of research has been performed in predicting PROMs following total knee arthroplasty (TKA), with 17 studies examining this question alone and an additional 13 examining other extremity joint surgeries. 22% ($n = 15$) followed either the TRIPOD-AI or CONSORT-AI guidelines, 35.3% ($n = 24$) were ranked as “good,” and 41.2% ($n = 28$) were “fair.”

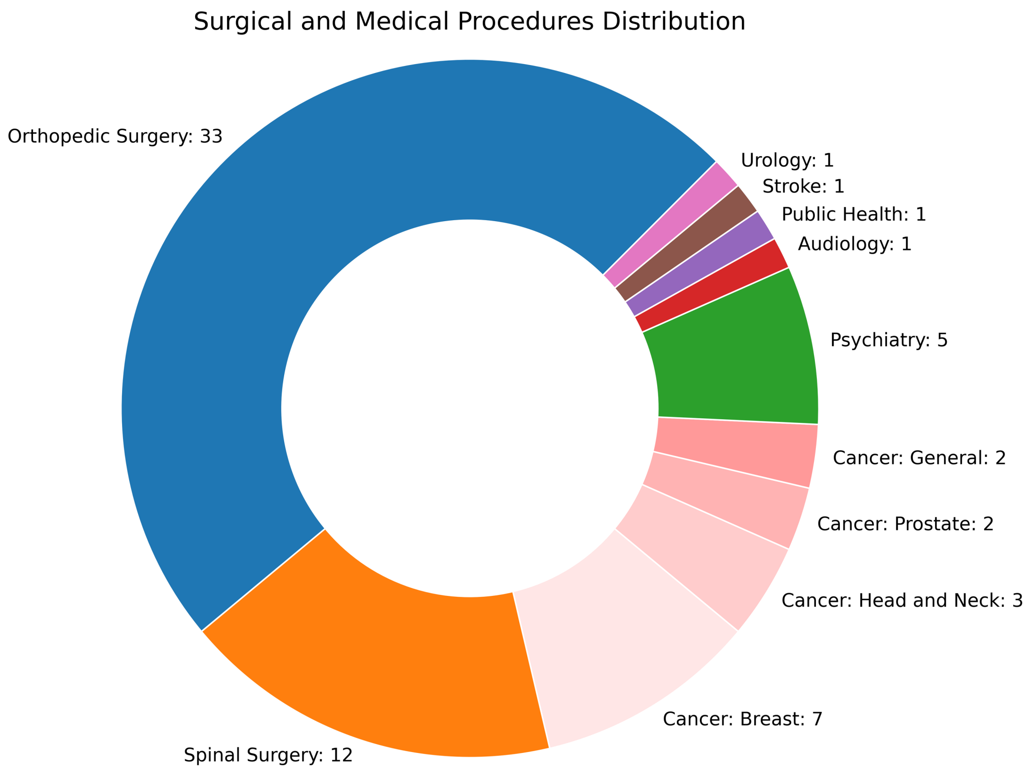

FIGURE 2
Distribution of studies by field.

Outcomes

Over a dozen PROM instruments were found in this literature review and are listed in Table 1. No one score predominated. The orthopedic studies focused on validated orthopedic PROM metrics. These include the ASES and ESES scores (American and European Shoulder and Elbow Surgeons) (Kumar et al., 2020; Alaiti et al., 2023; Taneja et al., 2024), COMI (Core Outcome Measures Index) (Halicka et al., 2023), Global Perceived Effect (Verma et al., 2023; Verma et al., 2022), KOOS (Knee Injury and Osteoarthritis Outcome Score) (Martin et al., 2022; Harris et al., 2021; Katakam et al., 2022; Ramkumar et al., 2021; Klemt et al., 2023; Fontana et al., 2019; Twiggs et al., 2021), Lysholm functional protocol (Ye et al., 2022), Oswestry Disability Index (Staartjes et al., 2019; Siccoli et al., 2019), QuickDASH (Quick Disabilities of the Arm, Shoulder, and Hand) (Brinkman et al., 2023; Harrison et al., 2022), Oswestry Disability Index (Staartjes et al., 2019), iHOT (International Hip Outcome Tool) (Pettit et al., 2023), HOS (Hip Outcome Score) (Kunze et al., 2021), HOOS (Hip Disability and Osteoarthritis Outcome Score) (Klemt et al., 2023; Fontana et al., 2019; Sniderman et al., 2021), IKDC (International Knee Documentation Committee) (Ramkumar et al., 2021; Ye et al., 2022; Ramkumar et al., 2021), MHQ (Michigan Hand outcomes Questionnaire) (Loos et al., 2022), Q score (Oxford Hip and Knee Score) (Huber et al., 2019), SRS-22r (Sociolois Research Society) (Ames et al., 2019; Nnamdi et al., 2023), and the WOMAC (Western Ontario and McMaster Universities Osteoarthritis Index) (Munn et al., 2022; Tschuggnall et al., 2021; Zhang et al., 2022; Zhou et al., 2023). These scores capture pain, symptoms, mobility/functionality, activities of daily living, and quality of life metrics related to the joint of interest, including the spine. Cancer-related tools included the BREAST-Q (Pfob et al., 2021; Pfob et al., 2023; Xu et al., 2023), Cancer Related Fatigue (Beenhakker et al., 2023), Lee Fatigue Scale (Kober et al., 2023; Kober et al., 2021), EORTC QLQ-C3 (European Organization for Research and Treatment of Cancer quality-of-life questionnaires) (Lee et al., 2020a), MDADI (MD Anderson Dysphagia Inventory) (Paetkau et al., 2024), IPSS (International Prostate Symptom Score) (Ghoreifi et al., 2023), EPIC 26 (Expanded Prostate Cancer Index Composite 26) (Agochukwu-Mmonu et al., 2022), and THYCA-QoL (Thyroid Cancer Quality of Life) (Lian et al., 2023). These score focus on measures related to symptoms following treatment, such as mastectomy results, fatigue, dry mouth, erectile dysfunction, etc. More universal quality-of-life PROMs included instruments and sub-instruments of the COST (COmprehensive Score for financial Toxicity) (Sidey-Gibbons et al., 2021), HAQ (Health Assessment Questionnaire) (Tschuggnall et al., 2021), EQ-5D-3L (EuroQol 5-Dimension 3-Level) (Zrubka et al., 2022; Harrison et al., 2022; Huber et al., 2019; Tschuggnall et al., 2021), PHQ-9 (Patient Health Questionnaire-9) (Coley et al., 2021; Bone et al., 2021), PASS (Patient Acceptable Symptom State) (Twiggs et al., 2021), PROMIS (Patient-Reported Outcomes Measurement Information System) (Karhade et al., 2021; Klemt et al., 2023; Brinkman et al., 2023; Hunter et al., 2024; Reps et al., 2022), and several versions of SF (Short Form Survey) (Ramkumar et al., 2021; Fontana et al., 2019; Ramkumar et al., 2021; Munn et al., 2022; Zhang et al., 2022; Zhou et al., 2023; Lian et al., 2023). In addition, several studies employed more basic instruments, capturing Visual Analogue Scores of Pain (Kumar et al., 2020; Halicka et al., 2023; Harris et al., 2021; Staartjes et al., 2019; Dolendo et al., 2022; Finkelstein et al., 2021; Park et al., 2023), numeric pain scores (Siccoli et al., 2019), and patient satisfaction scores on a Likert scale (Polce et al., 2021; Kumar et al., 2020; Munn et al., 2022; Farooq et al., 2020; Kunze et al., 2021; Kunze et al., 2020; Nam et al., 2023; Ulivi et al., 2023; Wang et al., 2023; Werneburg et al., 2023). Psychological measurements included GAD-7 (Generalized Anxiety Disorder) (Bone et al., 2021; Reps et al., 2022). One studied used the COSI (Client Oriented Scale of Improvement) for audiology (Suresh et al., 2023). Overall, these scores capture both cognitive, pain-related, and functional aspects of quality of life. 20 studies used minimally clinical important difference (MCID) on before and after scores of the PROMs to create a binary classification task.

TABLE 1 Patient reported outcome measure (PROM) instruments.

Orthopedic PROM	
ASES and ESES scores (American and European shoulder and elbow surgeons) (Kumar et al., 2020; Alaiti et al., 2023; Taneja et al., 2024) COMI (Core outcome measures index) (Halicka et al., 2023) Global perceived effect (Verma et al., 2023; Verma et al., 2022) KOOS (Knee injury and osteoarthritis outcome score) (Martin et al., 2022; Harris et al., 2021; Katakam et al., 2022; Ramkumar et al., 2021; Klemt et al., 2023; Fontana et al., 2019; Twiggs et al., 2021) Lysholm functional protocol (Ye et al., 2022) Oswestry disability index (Staartjes et al., 2019; Siccoli et al., 2019) QuickDASH (Quick disabilities of the arm, shoulder, and hand) (Brinkman et al., 2023; Harrison et al., 2022) Oswestry disability index (Staartjes et al., 2019)	iHOT (International hip outcome tool) (Pettit et al., 2023) HOS (Hip outcome score) (Kunze et al., 2021) HOOS (Hip disability and osteoarthritis outcome score) (Klemt et al., 2023; Fontana et al., 2019; Sniderman et al., 2021) IKDC (International knee documentation committee) (Ramkumar et al., 2021; Ye et al., 2022; Ramkumar et al., 2021) MHQ (Michigan hand outcomes questionnaire) (Loos et al., 2022) Q score (Oxford hip and knee score) (Huber et al., 2019) SRS-22r (Sociolois research society) (Ames et al., 2019; Nnamdi et al., 2023) WOMAC (Western ontario and mcmaster universities osteoarthritis index) (Munn et al., 2022; Tschuggnall et al., 2021; Zhang et al., 2022; Zhou et al., 2023)
Oncologic PROM	
BREAST-Q (Pfob et al., 2021; Pfob et al., 2023; Xu et al., 2023) Cancer related fatigue (Beenhakker et al., 2023) Lee Fatigue Scale (Kober et al., 2023; Kober et al., 2021) EORTC QLQ-C3 (European organization for research and treatment of cancer quality-of-life questionnaires) (Lee et al., 2020a; Lee et al., 2020b)	MDADI (MD Anderson dysphagia inventory) (Paetkau et al., 2024) IPSS (International prostate symptom score) (Ghoreifi et al., 2023) EPIC 26 (Expanded prostate cancer index composite 26) (Agochukwu-Mmonu et al., 2022) and THYCA-QoL (Thyroid cancer quality of life) (Lian et al., 2023)
General PROM	
COST (Comprehensive score for financial Toxicity) (Sidey-Gibbons et al., 2021) HAQ (Health assessment Questionnaire) (Tschuggnall et al., 2021) EQ-5D-3 L (EuroQol 5-dimension 3-Level) (Zrubka et al., 2022; Harrison et al., 2022; Huber et al., 2019; Tschuggnall et al., 2021) PHQ-9 (Patient health questionnaire-9) (Coley et al., 2021; Bone et al., 2021)	PASS (Patient acceptable symptom state) (Twiggs et al., 2021) PROMIS (Patient-reported outcomes measurement information system) (Karhade et al., 2021; Klemt et al., 2023; Brinkman et al., 2023; Hunter et al., 2024; Reps et al., 2022) SF (Short form survey) (Ramkumar et al., 2021; Fontana et al., 2019; Ramkumar et al., 2021; Munn et al., 2022; Zhang et al., 2022; Zhou et al., 2023; Lian et al., 2023) GAD-7 (Generalized anxiety disorder) (Bone et al., 2021; Reps et al., 2022)
Pain, visual analogue score (Kumar et al., 2020; Halicka et al., 2023; Harris et al., 2021; Staartjes et al., 2019; Dolendo et al., 2022; Finkelstein et al., 2021; Park et al., 2023)	
Patient satisfaction scores, Likert scale (Polce et al., 2021; Kumar et al., 2020; Munn et al., 2022; Farooq et al., 2020; Kunze et al., 2021; Kunze et al., 2020; Nam et al., 2023; Ulivi et al., 2023; Wang et al., 2023; Werneburg et al., 2023)	

Input data

Nearly all studies included demographic data as model features (n =65). The three remaining studies examined unstructured text (Lian et al., 2023; Wang et al., 2023; Matsuda et al., 2023). Twenty-nine studies included medical comorbidities. Thirty studies included sociodemographic data, including marital status, employment status, insurance information, drug use, and zip-code level income and education indices. Five studies assessed health care resource utilization, including hospitalizations and emergency room visits. Twenty-three of 36 studies involving surgeries included intraoperative characteristics, including surgeon, technical approach, types of implants used, characteristics of the tumor, and operative time. Fifty-two studies employed pre-event PROMs and included both the PROM outcome of interest alongside addition PROM metrics. Proportions of input variables are visualized in Figure 3.

Machine learning models

A range of machine learning techniques were employed. The majority (n =47) included some logistic or linear regression task as a base of comparison. All common machine learning models were employed, including linear and logistic regression, naive bayes, support vector machines, decision trees, random forest, and ensemble methods. These are shown in Figure 4. Only three studies employed large language models, both using the Bidirectional Encoder Representations from Transformers (BERT) architecture (Lian et al., 2023; Wang et al., 2023; Matsuda et al., 2023). Of note, 50 studies had some mention of data quality assessment. The majority addressed methods for handling missing data, largely through imputation or exclusion. Ten studies mentioned methods of handling class imbalance (Alaiti et al., 2023; Taneja et al., 2024; Ramkumar et al., 2021; Staartjes et al., 2019; Siccoli et al., 2019; Ramkumar et al., 2021; Ames et al., 2019; Zhang et al., 2022; Zhang et al., 2021; Chen et al., 2023).

Evaluation metrics

Overall, models performed poorly or moderately well, with few models approaching excellent discriminative capacity of AUROC exceeding 0.9. AUROC results ranged from 0.42 to 0.94 for any binary prediction task, with a mean of 0.78 and median of 0.77

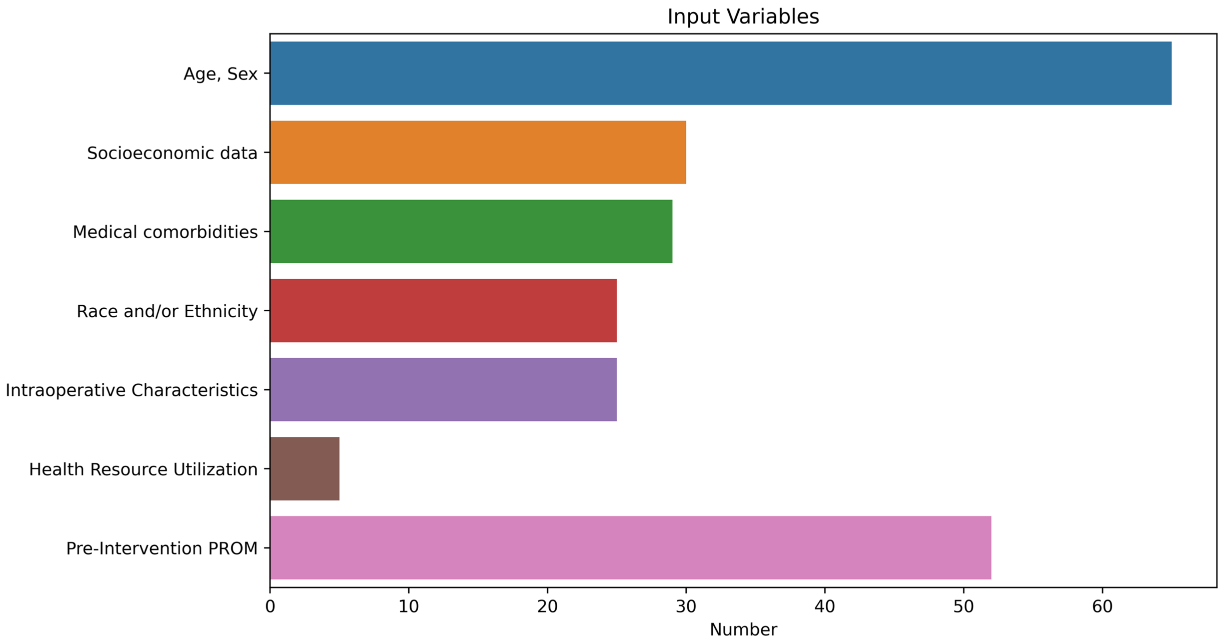

FIGURE 3
Input variables for patient reported outcome measure (PROM) prediction tasks.

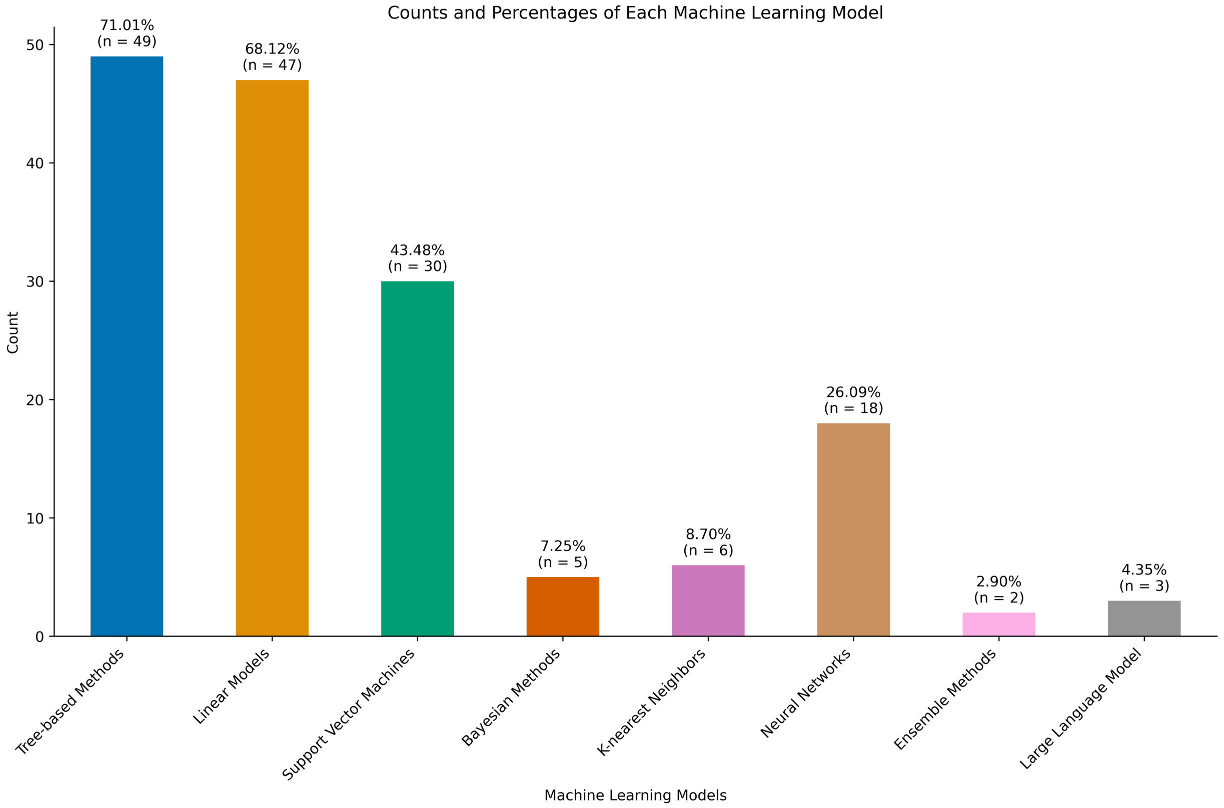

FIGURE 4
Machine learning models employed for patient reported outcome measure (PROM) prediction tasks.

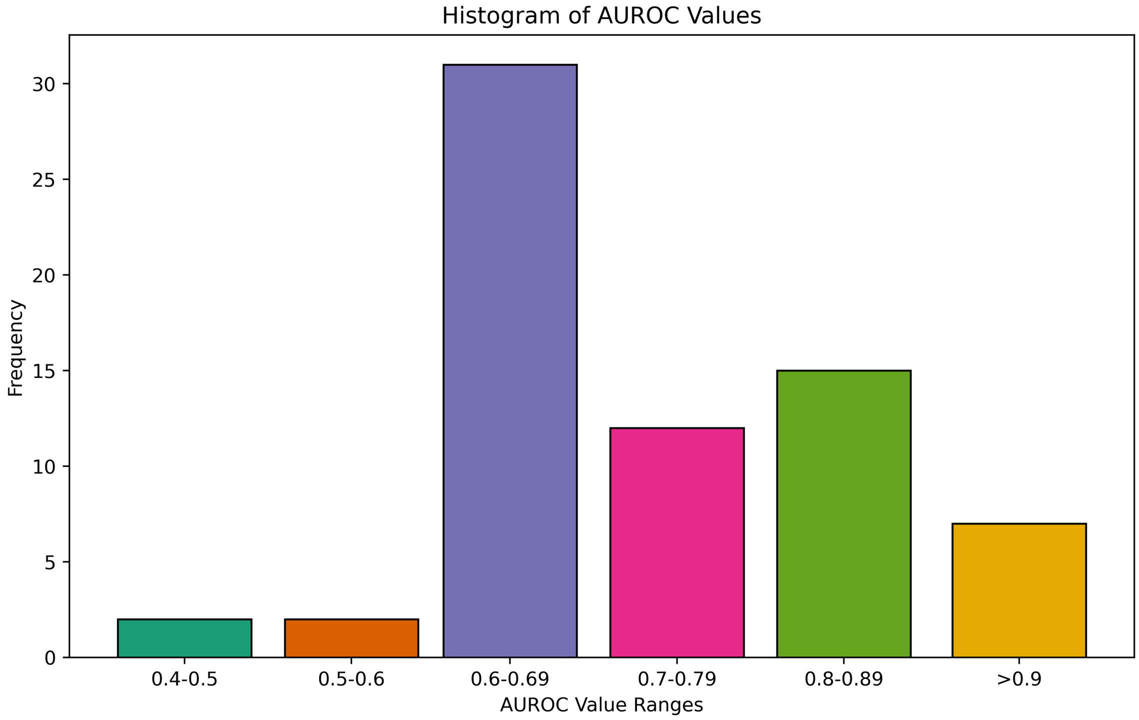

FIGURE 5
Area under the receiver operating curve (AUROC) performance metric distribution.

among all best-performing AUROCs reported. Several studies concluded that no meaningful relationship exists between pre-event variables and PROMs in their feature space, suggesting a need to collect more data or different variables (Ghoreifi et al., 2023; Halicka et al., 2023; Verma et al., 2023; Pettit et al., 2023; Loos et al., 2022; Beenhakker et al., 2023; Coley et al., 2021; Ulivi et al., 2023). A histogram of performance is shown in Figure 5. A few studies, however, found high discriminative performance, including for predicting MCID for improvement in back pain following lumbar disectomy (Staartjes et al., 2019) and hip pain following total hip arthroplasty (Kunze et al., 2021) as well as satisfaction with outcomes following mastectomy for cancer (Pfob et al., 2021). Other evaluation metrics included MSE and R (Rid, 2014), again with moderate performance at best (Ghoreifi et al., 2023; Verma et al., 2022; Agochukwu-Mmonu et al., 2022; Finkelstein et al., 2021; Ulivi et al., 2023; Suresh et al., 2023). There was no association between model type and performance. We also assessed calibration, which quantifies how much a model over or underestimates the probability of an event, an often overlooked, but no less important, metric (van den Goorbergh et al., 2022; Van Calster et al., 2019). 35.3% ($n = 24$) of studies evaluated the calibration of their models. The calibration was overall good, with excellent calibration metrics (intercepts $\leq \pm 0.1$ and slopes between 0.9 to 1.1) in 21% (4/19) of models that reported intercept and slope (Karhade et al., 2021; Halicka et al., 2023; Agochukwu-Mmonu et al., 2022; Ziobrowski et al., 2021). Other papers used Brier (Harris et al., 2021; Siccoli et al., 2019), Hosmer-Lemeshow (Martin et al., 2022; Lee et al., 2020a), and Speigelhatler (Xu et al., 2023) tests to prove calibration, noting acceptable performance.

Fairness and importance testing

While almost all models collected demographic information and mentioned need for external validation as a limitation to their generalizability, only six studies (8.9%) explicitly mentioned fairness or methods to mitigate bias (Simmons et al., 2024; Zrubka et al., 2022; Pfob et al., 2021; Pfob et al., 2023; Xu et al., 2023; Ziobrowski et al., 2021). Ziobrowski et al. examined model performance across age, sex, race/ethnicity, and income by estimating variations in the association of predicted risk with observed outcome using robust Poisson regression (Ziobrowski et al., 2021). In both their studies, Pfob et al. tested their models with and without sociodemographic and ethnic variables (fairness through unawareness) and obtained similar model performance (Pfob et al., 2021; Pfob et al., 2023). Zruboka, Simmons, and Xu evaluated prediction errors across different health statuses and demographics according to the PROM, with only the latter finding improve statistical performance for the African American group (Simmons et al., 2024; Zrubka et al., 2022; Xu et al., 2023). Simmons used the "four-fifths" legal guideline from the US Equal Employment Opportunity Commission to state that a "fair" model performs within 20% on any evaluation metric between demographic groups (Simmons et al., 2024). They found that ethnicity was rarely, but most frequently, outside this tolerance threshold, which the authors attributed to under-representation in the dataset.

Several studies employed potential fairness mitigation efforts without clear mention. One study employed inverse probability weighting to minimize the effects of missing data or under-represented groups, a potential marker of fairness but one that was not explicitly stated (Martin et al., 2022). Synthetic Minority Oversampling Technique (SMOTE) creates artificial data points that are plausibly close to actual data points and can be used as a fairness technique (Zhou et al., 2023). Several studies employed SMOTE create more balanced datasets in terms of their outcome of interest, and while this theoretically may improve representation of other minority classes, no study specifically examined this. However, we note that changing the overall prevalence of data classes through synthetic means may negatively impact model calibration (van den Goorbergh et al., 2022).

Importance testing was performed in 50 studies. Where performed, pre-event PROMs were either the largest or second largest contributor of post-event PROMs in all model predictions (Polce et al., 2021; Zrubka et al., 2022; Verma et al., 2023; Staartjes et al., 2019; Pettit et al., 2023; Munn et al., 2022; Pfob et al., 2021; Pfob et al., 2023; Xu et al., 2023; Kober et al., 2023; Kober et al., 2021; Park et al., 2023; Nam et al., 2023; Ulivi et al., 2023; Zhang et al., 2021). However, the correlations were not necessarily directly proportional: strong negative correlations of low PROMs sometimes predicted larger improvements in orthopedic studies, while other times demonstrated that PROMS in mobility, satisfaction rates, and narcotic use are unchanged after an event. Other top features trailed the PROMs, but included age, sex, BMI, patient anatomy, and comorbidities. Except for one study examining financial toxicity, where African American race was found to be predictive of toxicity (Sidey-Gibbons et al., 2021), no studies that measured ethnic or socioeconomic information reported its appearance in the top 5 predictive factors.

Narrative thematic analysis of theoretical PPPs

Patient preference predictors have been discussed in the literature since a series of publications in the *Journal of Medicine and Philosophy* in 2014 (Rid, 2014; Rid and Wendler, 2014a; Kim, 2014; Rid and Wendler, 2014b). With the growing prevalence of machine learning in medicine, the issue was re-visited in second series in 2022 in the *Journal of Medical Ethics* (Jardas et al., 2021; Earp, 2022; Ferrario et al., 2022; Schwan, 2022; Mainz, 2023). As technologies advance, the debates are becoming increasingly pertinent. In our analysis, we address key themes such as ethical considerations, the selection of model inputs, fairness in predictions, and the evaluation of model efficacy.

Ethical considerations

Patient autonomy is of utmost concern in the PPP, however, autonomy can be defined in both a primary sense ("I would not want CPR done") as well as a higher-order sense ("A decision was made for reasons I do not endorse") (Earp, 2022). Identifying what an incapacitated patient would want might also involve knowledge of how they prefer to make decisions. A second concern is the legal problem of using "naked statistical evidence." (Sharadin, 2018; Earp, 2022; Mainz, 2023; Ditto and Clark, 2014) Legal verdicts cannot be based on statistical correlations alone, as they do not imply causation, and the same may be said for a PPP. A third involves the lack of explainability with erosion of trust (Ferrario et al., 2022) mandating an alternative to "black box" models. Fourth, there potential for conflicting outputs by different PPPs (Sharadin, 2018) and even whether or not the patient would consent to the use of a PPP (Mainz, 2023). Deployment of these models would require extensive generalizability testing and buy-in from the public. Nevertheless, these articles acknowledge that a theoretical PPP has a low bar for improving decision making for incapacitated patients: human surrogate decisions, when analyzed retrospectively, are only slightly better than chance (Rid and Wendler, 2010; Shalowitz et al., 2006).

Model inputs

The discussions are fairly similar in their desired inputs for such a model and include: demographics, religious affiliations (Jardas et al., 2021; Rid and Wendler, 2014a; Sharadin, 2018; Rid and Wendler, 2014b; Earp, 2022; Ditto and Clark, 2014; Earp et al., 2024), level of risk taking (Jardas et al., 2021; Ditto and Clark, 2014), past treatment decisions (Rid and Wendler, 2014b; Mainz, 2023; Ditto and Clark, 2014; Earp et al., 2024; Benzinger et al., 2023), and baseline comorbidities (Mainz, 2023; Ditto and Clark, 2014). However, others call for more detailed examinations of attitudes toward death (Rid and Wendler, 2014a), personal experience with health care (Rid and Wendler, 2014a; Earp et al., 2024), and psychological and emotional functioning (Rid and Wendler, 2014a). Several argue for nation-level surveys to assess preferences and build more accurate models (Rid and Wendler, 2014a; Kim, 2014; Ditto and Clark, 2014), design forecasting scenarios of possible treatment outcomes (Ferrario et al., 2022), or even scraping publicly available information (Earp et al., 2024). As these studies focused more broadly on end-of-life decisions and not on specific operations or outcomes, none suggested intraoperative details or patient anatomy as a predictive measure.

Fairness

Model inputs are driven by the desire to build not only accurate models, but fair and just ones. Several papers warn that AI models may perpetuate social injustice (Rid, 2014; Biller-Andorno and Biller, 2019; Ferrario et al., 2022; Benzinger et al., 2023). In addition to incorporating various demographic and socioeconomic features, the perspectives of both the ill and the healthy must be incorporated to not unduly bias models toward one class of patients over another (Rid and Wendler, 2014b). Additionally, several authors mention that the just models would likely have to also understand what variables matter to the patient, i.e., whether or not to include religion, race, or education level as a factor (Wendler et al., 2016; Sharadin, 2018; Ferrario et al., 2022; Mainz, 2023). PROMs may potentially capture this variability, as they reflect direct, subjective patient expressions of their well-being. However, PROMs are not directly discussed by any of the cited articles.

Evaluation

Curiously, how to evaluate the accuracy of such models is also not often discussed (Rid and Wendler, 2014a). Many authors assume that surveying patients and their family members regarding decisions in hypothetical cases is sufficient to determine the accuracy of such models, however, given that patient preferences can often change radically in response to illness and end-of-life events, we ultimately lack a ground truth once incapacity has occurred. We know that interviewing survivors introduces a hindsight bias in treatment and that patients experience regret in only a minority of cases (Becerra Pérez et al., 2016; Rid and Wendler, 2014b). While several de-biasing strategies exist, no studies of predicting PROMs adjusted for hindsight bias in their analysis (Roese and Vohs, 2012).

Importantly, nearly all studies cautioned that the use of a patient preference predictors should only complement, and never replace, the provider or surrogate in making decisions for patients.

Discussion

This scoping review takes a novel approach to the theoretical development of the fair patient preference predictor but hypothesizing that the PPP would function essentially as PROM predictor. We show how current machine learning techniques predict PROMs for capacitated patient undergoing healthcare-related interventions might translate to predicting PROMs as a surrogate metric for incapacitated patient. We show that models had poor to moderate performance in predicting PROMs, the most important input variables were often from a pre-event PROM survey, and that few investigators directly assessed the fairness of their models.

There has been one previous review of using machine learning to predict PROMs (Verma et al., 2021) and another has called for placing them at the forefront of clinical AI research (Cruz Rivera et al., 2023). This has several implications to building the patient preference predictor. First, we see that demographics, social determinants, and even medical comorbidities rarely feature in the top 10 feature importance graphs, despite their inclusion in the majority of studies. Second, we see that baseline surveys of pain, functionality, and satisfaction are highly correlated with future PROMs. Third, fairness assessments on sociodemographic variables were rare, but when performed, were often reassuring. Given that sociodemographic variables were less predictive than pre-intervention PROM scores, it is possible that building a patient preference predictor incorporating these variables of functionality and wellbeing would be fair. This, however, does not negate the need the perform fairness testing. Fourth, we find a robust system of measuring patient satisfaction in place for select medical subspecialties (orthopedics). Documenting before and after changes in PROMs to establish MCID may benefit the future development of a patient preference predictor. We see promising developments with the incorporation of PROMs into the National Surgical Quality Improvement Program (NSQIP) (Temple et al., 2024). Finally, we find that large language models are showing potential for extracting this kind of information from unstructured textual data (Lian et al., 2023; Wang et al., 2023; Matsuda et al., 2023).

We noted several limitations to the included studies. Existing models have overall small numbers compared to the thousands to millions of examples machine learning models benefit from. This limits their generalizability but also highlights the difficulty of collecting quality of life metrics on patients, which are unfortunately limited to burdensome survey or interview data. Likely because of this, model performance is poor to moderate with AUROC's rarely exceeding 0.90. Second, nearly half of the studies were focused on extremity joint surgeries, which may limit generalizability, but remains informative based on the individual study's choice of model inputs, architectures, and evaluation metrics. Third, we note the wide variety of PROM metrics used. While these are helpful to hyper-specific outcomes, we would like to see more generalizable and wildly used PROM metrics to facilitate generalizability. Fourth, few studies report evaluation metrics outside AUROC, including AUPRC and F1 scores, which may be better at capturing rare events. It is up to the individual specialty to determine the appropriate threshold for clinical use, but models that aid in predicting life and death decisions for incapacitated patients would likely require a higher bar.

Conclusion

This review highlights many of the issues discussed in machine learning predictions of patient-centered outcomes. There are numerous practical, legal, and ethical barriers to using statistical evidence to fairly anticipate a decision in the incapacitated patient. Although machine learning models typically have poor to moderate performance in predicting PROMs, they often compare favorably with human surrogate decisions, which are only slightly better than chance.

Author contributions

JB: Conceptualization, Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. AC: Writing – original draft, Writing – review & editing. PH: Data curation, Writing – review & editing. LM: Data curation, Writing – review & editing. NB: Data curation, Writing – review & editing. AB: Funding acquisition, Supervision, Writing – review & editing. BS: Funding acquisition, Supervision, Writing – review & editing. RM: Conceptualization, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing. TL: Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. JAB was supported by the NIH T32 GM008721-24; TJL was supported by NIH awards K23 GM140268 and R01 GM149657.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1477447/full#supplementary-material

References

Agochukwu-Mmonu, N., Murali, A., Wittmann, D., Denton, B., Dunn, R. L., Montie, J., et al. (2022). Development and validation of dynamic multivariate prediction models of sexual function recovery in patients with prostate Cancer undergoing radical prostatectomy: results from the MUSIC statewide collaborative. *Eur. Urol. Open Sci.* 40, 1–8. doi: 10.1016/j.euros.2022.03.009

Alaiti, R. K., Vallio, C. S., Assunção, J. H., Andrade e Silva, F. B., Gracitelli, M. E. C., Neto, A. A. F., et al. (2023). Using machine learning to predict nonachievement of clinically significant outcomes after rotator cuff repair. *Orthop. J. Sports Med.* 11:6180. doi: 10.1177/23259671231206180

Ames, C. P., et al. (2019). Development of predictive models for all individual questions of SRS-22R after adult spinal deformity surgery: a step toward individualized medicine. *Eur. Spine J.* 28, 1998–2011. doi: 10.1007/s00586-019-06079-x

Assistant, P., et al. Predicting consumer behaviour with artificial intelligence. (2023).

Becerra Pérez, M. M., Menear, M., Brehaut, J. C., and Légaré, F. (2016). Extent and predictors of decision regret about health care decisions: a systematic review. *Med. Decis. Mak.* 36, 777–790. doi: 10.1177/0272989x16636113

Beenhakker, L., Wijlens, K. A. E., Witteveen, A., Heins, M., Korevaar, J. C., de Ligt, K. M., et al. (2023). Development of machine learning models to predict cancer-related fatigue in Dutch breast cancer survivors up to 15 years after diagnosis. *J. Cancer Surviv.* doi: 10.1007/s11764-023-01491-1

Benzinger, L., Ursin, F., Balke, W. T., Kacprowski, T., and Salloch, S. (2023). Should artificial intelligence be used to support clinical ethical decision-making? A systematic review of reasons. *BMC Med. Ethics* 24:48. doi: 10.1186/s12910-023-00929-6

Biller-Andorno, N., and Biller, A. (2019). Algorithm-aided prediction of patient preferences — An ethics sneak peek. *N. Engl. J. Med.* 381, 1480–1485. doi: 10.1056/nejmms1904869

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., et al. (2021). Dynamic prediction of psychological treatment outcomes: development and validation of a prediction model using routinely collected symptom data. *Lancet Digital Health* 3, e231–e240. doi: 10.1016/S2589-7500(21)00018-2

Brinkman, N., Shah, R., Doornberg, J., Ring, D., Gwilym, S., and Jayakumar, P. (2023). Artificial neural networks outperform linear regression in estimating 9-month patient-reported outcomes after upper extremity fractures with increasing number of variables. *OTA Int.* 7:e284. doi: 10.1097/OI9.0000000000000284

Chen, Y. W., Lin, K. C., Li, Y. C., and Lin, C. J. (2023). Predicting patient-reported outcome of activities of daily living in stroke rehabilitation: a machine learning study. *J. Neuroeng. Rehabil.* 20:25. doi: 10.1186/s12984-023-01151-6

Coley, R. Y., Boggs, J. M., Beck, A., and Simon, G. E. (2021). Predicting outcomes of psychotherapy for depression with electronic health record data. *J. Affect Disord. Rep.* 6:100198. doi: 10.1016/j.jadr.2021.100198

Cruz Rivera, S., Liu, X., Hughes, S. E., Dunster, H., Manna, E., Denniston, A. K., et al. (2023). Embedding patient-reported outcomes at the heart of artificial intelligence health-care technologies. The lancet digital. *Health* 5, e168–e173. doi: 10.1016/S2589-7500(22)00252-7

Detering, K. M., Hancock, A. D., Reade, M. C., and Silvester, W. (2010). The impact of advance care planning on end of life care in elderly patients: randomised controlled trial. *BMJ* 340:c1345. doi: 10.1136/bmj.c1345

Ditto, P. H., and Clark, C. J. (2014). Predicting end-of-life treatment preferences: perils and practicalities. *J. Med. Philos.* 39, 196–204. doi: 10.1093/jmp/jhu007

Dolendo, I. M., Wallace, A. M., Armani, A., Waterman, R. S., Said, E. T., and Gabriel, R. A. (2022). Predictive analytics for inpatient postoperative opioid use in patients undergoing mastectomy. *Cureus* 14:e23079. doi: 10.7759/cureus.23079

Earp, B. D. (2022). Meta-surrogate decision making and artificial intelligence. *J. Med. Ethics* 48, 287–289. doi: 10.1136/medethics-2022-108307

Earp, B. D., Porsdam Mann, S., Allen, J., Salloch, S., Suren, V., Jongsma, K., et al. (2024). A personalized patient preference predictor for substituted judgments in healthcare: technically feasible and ethically desirable. *Am. J. Bioeth.* 24, 13–26. doi: 10.1080/15265161.2023.2296402

Elicit: The AI research Assistant. (2023) Available at: https://elicit.com (Accessed September 09, 2024).

Fagerlin, A., and Schneider, C. E. (2004). Enough: the failure of the living will. *Hast. Cent. Rep.* 34, 30–42. doi: 10.2307/3527683

Farooq, H., Deckard, E. R., Ziemba-Davis, M., Madsen, A., and Meneghini, R. M. (2020). Predictors of patient satisfaction following primary Total knee arthroplasty: results from a traditional statistical model and a machine learning algorithm. *J. Arthroplast.* 35, 3123–3130. doi: 10.1016/j.arth.2020.05.077

Ferrario, A., Gloeckler, S., and Biller-Andorno, N. (2022). Ethics of the algorithmic prediction of goal of care preferences: from theory to practice. *J. Med. Ethics* 49, 165–174. doi: 10.1136/jme-2022-108371

Finkelstein, J. A., Stark, R. B., Lee, J., and Schwartz, C. E. (2021). Patient factors that matter in predicting spine surgery outcomes: a machine learning approach. *J. Neurosurg. Spine* 35, 127–136. doi: 10.3171/2020.10.SPINE201354

Fontana, M. A., Lyman, S., Sarker, G. K., Padgett, D. E., and MacLean, C. H. (2019). Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical Orthopaedics and Related Research®.* 477, 1267–1279.

Ghoreifi, A., Kaneko, M., Peretsman, S., Iwata, A., Brooks, J., Shakir, A., et al. (2023). Patient-reported satisfaction and regret following focal therapy for prostate Cancer: a prospective multicenter evaluation. *Eur. Urol. Open Sci.* 50, 10–16. doi: 10.1016/j.euros.2023.02.003

Halicka, M., Wilby, M., Duarte, R., and Brown, C. (2023). Predicting patient-reported outcomes following lumbar spine surgery: development and external validation of multivariable prediction models. *BMC Musculoskelet. Disord.* 24:333. doi: 10.1186/s12891-023-06446-2

Harris, A. H. S., Kuo, A. C., Bowe, T. R., Manfredi, L., Lalani, N. F., and Giori, N. J. (2021). Can machine learning methods produce accurate and easy-to-use preoperative prediction models of one-year improvements in pain and functioning after knee arthroplasty? *J. Arthroplast.* 36, 112–117.e6. doi: 10.1016/j.arth.2020.07.026

Harrison, C. J., Geoghegan, L., Sidey-Gibbons, C. J., Stirling, P. H. C., McEachan, J. E., and Rodrigues, J. N. (2022). Developing machine learning algorithms to support patient-centered, value-based carpal tunnel decompression surgery. *Plast. Reconstr. Surg. Glob. Open* 10:e4279. doi: 10.1097/gox.0000000000004279

Huber, M., Kurz, C., and Leidl, R. (2019). Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. *BMC Med. Inform. Decis. Mak.* 19:3. doi: 10.1186/s12911-018-0731-6

Hunter, J., Soleymani, F., Viktor, H., Michalowski, W., Poitras, S., and Beaulé, P. E. (2024). Using unsupervised machine learning to predict quality of life after Total knee arthroplasty. *J. Arthroplast.* 39, 677–682. doi: 10.1016/j.arth.2023.09.027

Jardas, E. J., Wasserman, D., and Wendler, D. (2021). Autonomy-based criticisms of the patient preference predictor. *J. Med. Ethics* 48, medethics-2021-107629–medethics-2021-107310. doi: 10.1136/medethics-2021-107629

Karhade, A. V., Fogel, H. A., Cha, T. D., Hershman, S. H., Doorly, T. P., Kang, J. D., et al. (2021). Development of prediction models for clinically meaningful improvement in PROMIS scores after lumbar decompression. *Spine J.* 21, 397–404. doi: 10.1016/j.spinee.2020.10.026

Katakam, A., Karhade, A. V., Collins, A., Shin, D., Bragdon, C., Chen, A. F., et al. (2022). Development of machine learning algorithms to predict achievement of minimal clinically important difference for the KOOS-PS following total knee arthroplasty. *J. Orthop. Res.* 40, 808–815. doi: 10.1002/jor.25125

Kim, S. Y. H. (2014). Improving medical decisions for incapacitated persons: does focusing on "accurate predictions" Lead to an inaccurate picture? *J. Med. Philos.* 39, 187–195. doi: 10.1093/jmp/jhu010

Klemt, C., Uzosike, A. C., Esposito, J. G., Harvey, M. J., Yeo, I., Subih, M., et al. (2023). The utility of machine learning algorithms for the prediction of patient-reported outcome measures following primary hip and knee total joint arthroplasty. *Arch. Orthop. Trauma Surg.* 143, 2235–2245. doi: 10.1007/s00402-022-04526-x

Kober, K. M., Roy, R., Conley, Y., Dhruva, A., Hammer, M. J., Levine, J., et al. (2023). Prediction of morning fatigue severity in outpatients receiving chemotherapy: less may still be more. *Support Care Cancer* 31:253. doi: 10.1007/s00520-023-07723-5

Kober, K. M., Roy, R., Dhruva, A., Conley, Y. P., Chan, R. J., Cooper, B., et al. (2021). Prediction of evening fatigue severity in outpatients receiving chemotherapy: less may be more. *Fatigue* 9, 14–32. doi: 10.1080/21641846.2021.1885119

Kumar, V., Roche, C., Overman, S., Simovitch, R., Flurin, P. H., Wright, T., et al. (2020). What is the accuracy of three different machine learning techniques to predict clinical outcomes after shoulder arthroplasty? *Clin. Orthop. Relat. Res.* 478, 2351–2363. doi: 10.1097/CORR.0000000000001263

Kunze, K. N., Polce, E. M., Nwachukwu, B. U., Chahla, J., and Nho, S. J. (2021). Development and internal validation of supervised machine learning algorithms for predicting clinically significant functional improvement in a mixed population of primary hip arthroscopy. *Arthroscopy J. Arthroscopic Related Surg.* 37, 1488–1497. doi: 10.1016/j.arthro.2021.01.005

Kunze, K. N., Polce, E. M., Rasio, J., and Nho, S. J. (2021). Machine learning algorithms predict clinically significant improvements in satisfaction after hip arthroscopy. *Arthroscopy* 37, 1143–1151. doi: 10.1016/j.arthro.2020.11.027

Kunze, K. N., Polce, E. M., Sadauskas, A. J., and Levine, B. R. (2020). Development of machine learning algorithms to predict patient dissatisfaction after primary Total knee arthroplasty. *J. Arthroplast.* 35, 3117–3122. doi: 10.1016/j.arth.2020.05.061

Lee, S., Deasy, J. O., Oh, J. H., di Meglio, A., Dumas, A., Menvielle, G., et al. (2020a). Prediction of breast Cancer treatment–induced fatigue by machine learning using genome-wide association data. *JNCI Cancer Spectr.* 4:pkaa039. doi: 10.1093/jncics/pkaa039

Lee, S., Deasy, J. O., Oh, J. H., di Meglio, A., Dumas, A., Menvielle, G., et al. (2020b). Prediction of breast cancer treatment–induced fatigue by machine learning using genome-wide association data. *JNCI Cancer Spectrum.* 4:pkaa039. doi: 10.1093/JNCICS/PKAA039

Lian, R., Hsiao, V., Hwang, J., Ou, Y., Robbins, S. E., Connor, N. P., et al. (2023). Predicting health-related quality of life change using natural language processing in thyroid cancer. *Intell. Based Med.* 7:100097. doi: 10.1016/j.ibmed.2023.100097

Loos, N. L., Hoogendam, L., Souer, J. S., Slijper, H. P., Andrinopoulou, E. R., Coppieters, M. W., et al. (2022). Machine learning can be used to predict function but not pain after surgery for thumb carpometacarpal osteoarthritis. *Clin. Orthop. Relat. Res.* 480, 1271–1284. doi: 10.1097/CORR.0000000000002105

Mainz, J. T. (2023). The patient preference predictor and the objection from higher-order preferences. *J. Med. Ethics* 49, 221–222. doi: 10.1136/jme-2022-108427

Martin, R. K., Wastvedt, S., Pareek, A., Persson, A., Visnes, H., Fenstad, A. M., et al. (2022). Predicting subjective failure of ACL reconstruction: a machine learning analysis of the Norwegian knee ligament register and patient reported outcomes. *J ISAKOS* 7, 1–9. doi: 10.1016/j.jisako.2021.12.005

Matsuda, S., Ohtomo, T., Okuyama, M., Miyake, H., and Aoki, K. (2023). Estimating patient satisfaction through a language processing model: model development and evaluation. *JMIR Form. Res.* 7:e48534. doi: 10.2196/48534

McGlothlin, A. E., and Lewis, R. J. (2014). Minimal clinically important difference: defining what really matters to patients. *JAMA* 312, 1342–1343. doi: 10.1001/jama.2014.13128

Munn, J. S., Lanting, B. A., MacDonald, S. J., Somerville, L. E., Marsh, J. D., Bryant, D. M., et al. (2022). Logistic regression and machine learning models cannot discriminate between satisfied and dissatisfied Total knee arthroplasty patients. *J. Arthroplast.* 37, 267–273. doi: 10.1016/j.arth.2021.10.017

Nam, H. S., Ho, J. P. Y., Park, S. Y., Cho, J. H., and Lee, Y. S. (2023). The development of machine learning algorithms that can predict patients satisfaction using baseline characteristics, and preoperative and operative factors of total knee arthroplasty. *Knee* 44, 253–261. doi: 10.1016/j.knee.2023.08.018

Nnamdi, M. C., Shi, W., Tamo, J. B., Iwinski, H. J., Wattenbarger, J. M., Wang, M. D., et al. (2023). Concept Bottleneck Model for Adolescent Idiopathic Scoliosis Patient Reported Outcomes Prediction. *Paper presented at: BHI 2023 - IEEE-EMBS International Conference on Biomedical and Health Informatics, Proceedings 2023.*

Paetkau, O., Weppler, S., Quon, H. C., Tchistiakova, E., and Kirkby, C. (2024). Developing and validating multi-omics prediction models for late patient-reported dysphagia in head and neck radiotherapy. *Biomed. Physics Eng. Exp.* 10:045014. doi: 10.1088/2057-1976/ad4651

Park, C., Mummaneni, P. V., Gottfried, O. N., Shaffrey, C. I., Tang, A. J., Bisson, E. F., et al. (2023). Which supervised machine learning algorithm can best predict achievement of minimum clinically important difference in neck pain after surgery in patients with cervical myelopathy? A QOD study. *Neurosurg. Focus* 54:E5. doi: 10.3171/2023.3.FOCUS2372

Pettit, M. H., Hickman, S. H. M., Malviya, A., and Khanduja, V. (2023). Development of machine learning algorithms to predict attainment of minimal clinically important difference after hip arthroscopy for Femoroacetabular impingement yield fair performance and limited clinical utility. *Arthroscopy* 40, 1153–1163.e2. doi: 10.1016/j.arthro.2023.09.023

Pfob, A., Mehrara, B. J., Nelson, J. A., Wilkins, E. G., Pusic, A. L., and Sidey-Gibbons, C. (2021). Machine learning to predict individual patient-reported outcomes at 2-year follow-up for women undergoing cancer-related mastectomy and breast reconstruction (INSPiRED-001). *Breast* 60, 111–122. doi: 10.1016/j.breast.2021.09.009

Pfob, A., Mehrara, B. J., Nelson, J. A., Wilkins, E. G., Pusic, A. L., and Sidey-Gibbons, C. (2023). Towards patient-centered decision-making in breast Cancer surgery: machine learning to predict individual patient-reported outcomes at 1-year follow-up. *Ann. Surg.* 277, e144–e152. doi: 10.1097/SLA.0000000000004862

Polce, E. M., Kunze, K. N., Fu, M. C., Garrigues, G. E., Forsythe, B., Nicholson, G. P., et al. (2021). Development of supervised machine learning algorithms for prediction of satisfaction at 2 years following total shoulder arthroplasty. *J. Shoulder Elb. Surg.* 30, e290–e299. doi: 10.1016/j.jse.2020.09.007

Rajkomar, A., Oren, E., Chen, K., Dai, A. M., Hajaj, N., Hardt, M., et al. (2018). Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* 1:18. doi: 10.1038/s41746-018-0029-1

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. (2022). AI in health and medicine. *Nat. Med.* 28, 31–38. doi: 10.1038/s41591-021-01614-0

Ramkumar, P. N., Karnuta, J. M., Haeberle, H. S., Owusu-Akyaw, K. A., Warner, T. S., Rodeo, S. A., et al. (2021). Association between preoperative mental health and clinically meaningful outcomes after osteochondral allograft for cartilage defects of the knee: a machine learning analysis. *Am. J. Sports Med.* 49, 948–957. doi: 10.1177/0363546520988021

Ramkumar, P. N., Karnuta, J. M., Haeberle, H. S., Rodeo, S. A., Nwachukwu, B. U., and Williams III, R. J. (2021). 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.* 49, 2177–2186. doi: 10.1177/03635465211015179

Reps, J. M., Wilcox, M., McGee, B., Leonte, M., LaCross, L., and Wildenhaus, K. (2022). Development of multivariable models to predict perinatal depression before and after delivery using patient reported survey responses at weeks 4–10 of pregnancy. *BMC Pregnancy Childbirth* 22. doi: 10.1186/s12884-022-04741-9

Rid, A. (2014). Will a patient preference predictor improve treatment decision making for incapacitated patients? *J. Med. Philos.* 39, 99–103. doi: 10.1093/jmp/jhu005

Rid, A., and Wendler, D. (2010). Can we improve treatment decision-making for incapacitated patients? *Hast. Cent. Rep.* 40, 36–45

Rid, A., and Wendler, D. (2014a). Treatment decision making for incapacitated patients: is development and use of a patient preference predictor feasible? *J. Med. Philos.* 39, 130–152. doi: 10.1093/jmp/jhu006

Rid, A., and Wendler, D. (2014b). Use of a patient preference predictor to help make medical decisions for incapacitated patients. *J. Med. Philos.* 39, 104–129. doi: 10.1093/jmp/jhu001

Roese, N. J., and Vohs, K. D. (2012). Hindsight Bias. *Perspect. Psychol. Sci.* 7, 411–426. doi: 10.1177/1745691612454303

Schwan, B. (2022). Sovereignty, authenticity and the patient preference predictor. *J. Med. Ethics* 48, 311–312. doi: 10.1136/medethics-2022-108292

Shalowitz, D. I., Garrett-Mayer, E., and Wendler, D. (2006). The accuracy of surrogate decision makers: a systematic review. *Arch. Intern. Med.* 166, 493–497. doi: 10.1001/archinte.166.5.493

Sharadin, N. (2018). Patient preference predictors and the problem of naked statistical evidence. *J. Med. Ethics* 44, 857–862. doi: 10.1136/medethics-2017-104509

Siccoli, A., de Wispelaere, M. P., Schröder, M. L., and Staartjes, V. E. (2019). Machine learning–based preoperative predictive analytics for lumbar spinal stenosis. *Neurosurg. Focus.* 46:E5. doi: 10.3171/2019.2.FOCUS18723

Sidey-Gibbons, C., Pfob, A., Asaad, M., Boukovalas, S., Lin, Y. L., Selber, J. C., et al. (2021). Development of machine learning algorithms for the prediction of financial toxicity in localized breast cancer following surgical treatment. *JCO Clinical Cancer Inform.* 5, 338–347. doi: 10.1200/CCI.20.00088

Simmons, C., DeGrasse, J., Polakovic, S., Aibinder, W., Throckmorton, T., Noerdlinger, M., et al. (2024). Initial clinical experience with a predictive clinical decision support tool for anatomic and reverse total shoulder arthroplasty. *Eur. J. Orthop. Surg. Traumatol.* 34, 1307–1318. doi: 10.1007/s00590-023-03796-4

Sniderman, J., Stark, R. B., Schwartz, C. E., Imam, H., Finkelstein, J. A., and Nousiainen, M. T. (2021). Patient factors that matter in predicting hip arthroplasty outcomes: a machine-learning approach. *J. Arthroplast.* 36, 2024–2032. doi: 10.1016/j.arth.2020.12.038

Staartjes, V. E., de Wispelaere, M. P., Vandertop, W. P., and Schröder, M. L. (2019). Deep learning-based preoperative predictive analytics for patient-reported outcomes following lumbar discectomy: feasibility of center-specific modeling. *Spine J.* 19, 853–861. doi: 10.1016/j.spinee.2018.11.009

Suresh, K., Franck, K., Arenberg, J. G., Song, Y., Lee, D. J., and Crowson, M. G. (2023). Development of a predictive model for individualized hearing aid benefit. *Otol. Neurotol.* 44, e1–e7. doi: 10.1097/mao.0000000000003739

Taneja, A., Talavage, T. M., Grawe, B. M. (2024). Influence of preoperative variables on patient satisfaction: a machine learning approach in rotator cuff repair surgeries. *Paper presented at: BHI 2023 - IEEE-EMBS International Conference on Biomedical and Health Informatics, Proceedings 2023.*

Temple, L. K. F., Pusic, A. L., Liu, J. B., Melucci, A. D., Collins, C. E., Kazaure, H. S., et al. (2024). Patient-reported outcome measures within a National Multispecialty Surgical Quality Improvement Program. *JAMA Surg.* 159, 1030–1039. doi: 10.1001/jamasurg.2024.1757

Tschuggnall, M., Grote, V., Pirchl, M., Holzner, B., Rumpold, G., and Fischer, M. J. (2021). Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Inform. Med. Unlocked* 24:100598. doi: 10.1016/j.imu.2021.100598

Twiggs, J., Miles, B., Roe, J., Fritsch, B., Liu, D., Parker, D., et al. (2021). Can TKA outcomes be predicted with computational simulation? Generation of a patient specific planning tool. *Knee* 33, 38–48. doi: 10.1016/j.knee.2021.08.029

Ulivi, M., Orlandini, L., D'Errico, M., Perrotta, R., Perfetti, S., Ferrante, S., et al. (2023). Medium-term patient's satisfaction after primary total knee arthroplasty: enhancing prediction for improved care. *Orthop. Traumatol. Surg. Res.* 110:103734. doi: 10.1016/j.otsr.2023.103734

Van Calster, B., et al. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Med.* 17:230. doi: 10.1186/s12916-019-1466-7

van den Goorbergh, R., van Smeden, M., Timmerman, D., and Van Calster, B. (2022). The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J. Am. Med. Inform. Assoc.* 29, 1525–1534. doi: 10.1093/jamia/ocac093

Verma, D., Bach, K., and Mork, P. J. (2021). Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. *Informatics* 8:56. doi: 10.3390/informatics8030056

Verma, D., Bach, K., and Mork, P. J. (2023). External validation of prediction models for patient-reported outcome measurements collected using the selfBACK mobile app. *Int. J. Med. Inform.* 170:104936. doi: 10.1016/j.ijmedinf.2022.104936

Verma, D., Jansen, D., Bach, K., Poel, M., Mork, P. J., and d'Hollosy, W. O. N. (2022). Exploratory application of machine learning methods on patient reported data in the development of supervised models for predicting outcomes. *BMC Med. Inform. Decis. Mak.* 22:227. doi: 10.1186/s12911-022-01973-9

Wang, Y., Yu, Y., Liu, Y., Ma, Y., and Pang, P. C. I. (2023). Predicting Patients' satisfaction with mental health drug treatment using their reviews: unified interchangeable model fusion approach. *JMIR Mental Health* 10:e49894. doi: 10.2196/49894

Weinfurt, K. P., and Reeve, B. B. (2022). Patient-reported outcome measures in clinical research. *JAMA* 328, 472–473. doi: 10.1001/jama.2022.11238

Wendler, D., Wesley, B., Pavlick, M., and Rid, A. (2016). A new method for making treatment decisions for incapacitated patients: what do patients think about the use of a patient preference predictor? *J. Med. Ethics* 42, 235–241. doi: 10.1136/medethics-2015-103001

Werneburg, G. T., Werneburg, E. A., Goldman, H. B., Mullhaupt, A. P., and Vasavada, S. P. (2023). Neural networks outperform expert humans in predicting patient impressions of symptomatic improvement following overactive bladder treatment. *Int. Urogynecol. J.* 34, 1009–1016. doi: 10.1007/s00192-022-05291-6

Xu, C., Pfob, A., Mehrara, B. J., Yin, P., Nelson, J. A., Pusic, A. L., et al. (2023). 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.* 30, 7046–7059. doi: 10.1245/s10434-023-13971-w

Ye, Z., Zhang, T., Wu, C., Qiao, Y., Su, W., Chen, J., et al. (2022). Predicting the objective and subjective clinical outcomes of anterior cruciate ligament reconstruction: a machine learning analysis of 432 patients. *Am. J. Sports Med.* 50, 3786–3795. doi: 10.1177/03635465221129870

Zhang, S., Chen, J. Y., Pang, H. N., Lo, N. N., Yeo, S. J., and Liow, M. H. L. (2021). Development and internal validation of machine learning algorithms to predict patient satisfaction after total hip arthroplasty. *Arthroplasty* 3:33. doi: 10.1186/s42836-021-00087-3

Zhang, S., Lau, B. P. H., Ng, Y. H., Wang, X., and Chua, W. (2022). Machine learning algorithms do not outperform preoperative thresholds in predicting clinically meaningful improvements after total knee arthroplasty. *Knee Surg. Sports Traumatol. Arthrosc.* 30, 2624–2630. doi: 10.1007/s00167-021-06642-4

Zhou, Y., Dowsey, M., Spelman, T., Choong, P., and Schilling, C. (2023). SMART choice (knee) tool: a patient-focused predictive model to predict improvement in health-related quality of life after total knee arthroplasty. *ANZ J. Surg.* 93, 316–327. doi: 10.1111/ans.18250

Zhou, Y., Kantarcioglu, M., and Clifton, C. (2023). On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). Society for Industrial and Applied Mathematics, 874–882.

Ziobrowski, H. N., Kennedy, C. J., Ustun, B., House, S. L., Beaudoin, F. L., An, X., et al. (2021). Development and validation of a model to predict posttraumatic stress disorder and major depression after a motor vehicle collision. *JAMA Psychiatry* 78, 1228–1237. doi: 10.1001/jamapsychiatry.2021.2427

Zrubka, Z., Csabai, I., Hermann, Z., Golicki, D., Prevolnik-Rupel, V., Ogorevc, M., et al. (2022). Predicting patient-level 3-level version of EQ-5D index scores from a large international database using machine learning and regression methods. *Value Health* 25, 1590–1601. doi: 10.1016/j.jval.2022.01.024