Informatics in Medicine Unlocked 24 (2021) 100598

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures

Michael Tschuggnall [a,*], Vincent Grote [b,c,g], Michael Pirchl [b,d], Bernhard Holzner [e,a,f], Gerhard Rumpold [e,a], Michael J. Fischer [b,d,g]

[a] *Evaluation Software Development, Innsbruck, Austria*
[b] *Ludwig Boltzmann Institute for Rehabilitation Research, Vienna, Austria*
[c] *Otto Loewi Research Center, Division of Physiology, Medical University of Graz, Austria*
[d] *Vamed Rehabilitation Center Kitzbuehel, Kitzbuehel, Austria*
[e] *Medical University of Innsbruck, Austria*
[f] *University Hospital of Psychiatry I, Innsbruck, Austria*
[g] *Department of Physical Medicine, Rehabilitation and Occupational Medicine, Medical University of Vienna, Austria*

ARTICLE INFO

Keywords:
Clinical decision making support
Outcome prediction
Machine learning
Rehabilitation

ABSTRACT

A common way to treat hip, knee or foot injuries is by conducting a corresponding physician-guided rehab over several weeks or even months. While health professionals are often able to estimate the treatment success beforehand to a certain extent based on their experience, it is scientifically still not clear to what extent relevant factors and circumstances explain or predict rehab outcomes. To this end, we apply modern machine learning techniques to a real-life dataset consisting of data from more than a thousand rehab patients (N = 1,047) and build models that are able to predict the rehab success for a patient upon treatment start. By utilizing clinical and patient-reported outcome measures (PROMs) from questionnaires, we compute patient-related clinical measurements (CROMs) for different targets like the range of motion of a knee, and subsequently use those indicators to learn prediction models. While we at first apply regression algorithms to estimate the rehab success in terms of percental admission and discharge value differences, we finally also utilize classification models to make predictions based on a three-classed grading scheme. Extensive evaluations for different treatment groups and targets show promising results with F-scores exceeding 65% that are able to substantially outperform baselines (by up to 40%) and thus show that machine learning can indeed be applied for better medical controlling and optimized treatment paths in rehab praxis. Future developments should include further relevant critical success criteria in the rehabilitation routine to further optimize the prognosis models for clinical practice.

1. Introduction

The prevalence of disabling conditions has increased dramatically [1]. Rehabilitation plays a vital role in the mitigation and improvement of functional limitations associated with ageing and chronic conditions. These include in particular degenerative diseases of the musculoskeletal system [2,3]. In those fields, health professionals are often able to estimate rehabilitation success based on their experience regarding clinical measurements (CROMs) at the start of the treatment. Moreover, also patient-reported outcome measures (PROMs) in terms of completed questionnaires may be included in this process. While such subjective estimates are important and valid, it is often not clear what the most influencing and determining factors for a good prognosis are. As outlined in Section 2, to the best of our knowledge there is no existing approach or computer model to properly aid or guide physicians in this estimation process, nor to provide informative feedback on the most influencing factors after a completed rehab treatment. Based on practice and evidence in the literature, we therefore want to establish new technical, validated standards to predict rehabilitation outcomes in post-acute patients, with focus on routine outcome data and machine learning. By utilizing real-life data of more than thousand rehab patients, we aim to create the basis for a more personalized healthcare that benefits from a continuous improvement process using supervised machine learning - a new clinical routine in rehabilitation.

* Corresponding author.
E-mail address: michael.tschuggnall@ches.pro (M. Tschuggnall).

https://doi.org/10.1016/j.imu.2021.100598
Received 22 February 2021; Received in revised form 5 May 2021; Accepted 8 May 2021
Available online 21 May 2021

To sum up, we specifically address the following research questions in this work:

- Can machine learning algorithms be utilized to predict the rehabilitation success of patients suffering from hip, knee or foot injuries, and if yes, to which extent?
- What are the most influencing factors for a good prediction model, which can in further consequence be used by physicians to plan a successful treatment?

The remainder of this paper is structured as follows: After providing a short overview of medical rehabilitation, related work is presented in Section 2. The dataset including all relevant PROMs and CROMs utilized in this study are outlined in Section 3. After describing the main contribution of this study in terms of methodology in Section 4, the corresponding results are presented in detail in Section 5. Finally, Section 6 concludes and discusses possible future work.

1.1. Medical rehabilitation

All over the world, medical rehabilitation is structured in different ways, although a tendency for a standardization of the social and healthcare systems can be observed [4]. In Austria this is the most frequent reason (36%) for an inpatient rehabilitation [5]. An inpatient treatment lasts on average 2–3 hours per day. An individual rehabilitative program consisting of active and passive treatments is provided. Active treatments consist of physical activity including gymnastic and individual physiotherapy sessions and the medical training focuses on underwater, ergometer, nordic walking, strength, balance, relaxation and motion training. Passive treatments contain sessions like massages, thermotherapy, electrotherapy, ultrasound and educational lessons like various lectures, or psychological coaching. Each patient is offered a program of at least 30 hours of therapy across three weeks, split into approximately 50% active and 50% passive treatments. Such multidisciplinary orthopedic rehabilitation in specialized rehabilitation centers improves well-being and physical functioning and reduces risk factors in the majority of patients [6–8]. Especially for rehabilitation after hip and knee endoprosthesis (TEP) a high level of evidence is stated [9,10]. Following an interdisciplinary treatment of medical, physical activity-based and psychological therapy, re-entry into the labor market remains high [11,12].

The structure of the inpatient healthcare unit has a strong influence on the quality of care. Routine collection of standardized outcome measures is recommended to compare different populations, programs and practices [13]. Based on Health Technology Assessment research, a framework of contracts exists with Austrian social security institutions. This framework includes performance agreements that are based on criteria regarding the quality of the processes and treatment outcomes [14]. Data collected in this study could therefore be exported directly from the electronic and verified patient records, based on a common and mandatory routine data collection, in accordance with national legislative guidelines at the points of admission and discharge.

2. Related work

Machine learning and deep learning techniques have been applied in healthcare with increasing frequency in the last decade. Nevertheless, physicians often still rely on traditional methods for decision making or treatment planning. Reasons for that might include that artificial intelligence (AI) has never been applied to the specific field, systems are not mature enough, physicians or patients do not understand machine learning results or simply do not trust them [15]. After laying out the AI impact and improvements on three levels (i.e., clinicians, healthcare and patients), Topol also questions the willingness to apply it, i.e., “whether that will be used to improve the patient-doctor relationship or facilitate its erosion remains to be seen” [16]. To this end, e.g., Norgeot et al. recently made a call to use AI, as “now is the time to create smarter healthcare systems in which the best treatment decisions are computationally learned from electronic health record data” [17].

On the other side, AI is already utilized successfully in many areas of medicine as is broadly outlined by Esteva et al. [18]. Consequently one of the most prominent application areas is probably the processing and categorization of images, that is used in many specific areas including radiology (e.g., [19]), pathology (e.g., [20]), dermatology (e.g., [21]) or cardiology (e.g., [22]). But recently many other fields have also been targeted, including general computer vision, natural language processing, robotic-assisted surgery, genomics, clinical outcome prediction or general decision making [18]. As a result, a wide range of applications show promising results, which often comprise evaluations indicating that AI can match or even exceed clinicians decisions. A few examples include the prediction of dementia [23], automatic extraction of useful information from electronic medical records [24], assessment of mortality risk [25,26], identification of Alzheimer’s disease [27] or even the prediction of potential suicide [28].

Nevertheless, with respect to the specific field of rehabilitation, only a few studies exist to our knowledge, although Zhu et al. [29] already showed the potential of machine learning more than a decade ago. In their comparative study incorporating more than 20,000 home care patients, they found that even quite a simple algorithm like K-nearest neighbor (KNN) can predict the rehabilitation potential better than commonly used clinical assessment protocols. In subsequent studies, Zhu et al. also demonstrated that support vector machines (SVM) and random forests significantly exceed common practices [30,31].

Similar to this study, Lin et al. recently used machine learning to predict the outcome of rehab treatments after strokes [32]. Analyzing the data of approximately 300 patients, logistic regression, SVMs and random forest were used to predict the Barthel Index [33] at discharge. Evaluations showed that regression algorithms are able to estimate the outcome value at a mean absolute error of about 10, and that classifiers can achieve an accuracy of over 70% for categorizing the Barthel Index status in a three-class scheme.

Recently, Huber et al. also conducted a study [34] where machine learning was used to predict patient-reported outcomes after hip and knee replacement surgeries. In contrast to this study, the authors utilized PROMs only and aimed to predict the quality of life, i.e., the estimation of surgery success was not part of the study. Nevertheless, using eight different supervised classifiers promising results have been reported for quality of life.

All the previously mentioned studies focus on specific problems and data sets, and consequently the results are not directly comparable to those of this study. Nevertheless, they show the general potential of machine learning, which we believe could be utilized a lot more frequently in the field of rehabilitation. Along these lines we aim to further fill the gap and add to those studies by applying machine learning to data of patients suffering from hip, knee or foot injuries. By showing the general potential of artificial intelligence to assist health professionals in their decision-making, we hope to additionally motivate other researchers to also apply machine/deep learning on their specific fields and data. Finally, by openly discussing the insights and benefits of this work we hope to counteract the previously mentioned general scepsis for machine learning to be applied on medical data.

3. Dataset

For this study, we utilize an anonymized real-world dataset containing data from 1,047 rehab patients of the Vamed Rehabilitation Center Kitzbühel.[1] More specifically, patients were allocated to five different groups:

[1] https://www.reha-kitz.at.

1. Trauma hip region (HIP_T, N = 148): Proximal femoral fractures, subtrochanteric, peritrochanteric, acetabulum fracture
2. Trauma knee region ($KNEE_T$, N = 109): Distal femoral fractures, proximal tibia and fibula fractures, patella fracture
3. Trauma ankle/heel region (ANKLE, N = 92): Fractures of distal tibia and fibula, calcaneus fractures
4. Hip arthroplasty (HIP_A, N = 292)
5. Knee arthroplasty ($KNEE_A$, N = 406)

Groups 1, 2 and 3 contained patients with fractures or traumatic injuries in the hip, knee or ankle region. Groups 4 and 5 had hip or knee arthroplasty. All patients had inpatient rehabilitation for 21 days.

Depending on the treatment group, the dataset contains a group-specific value representing the state of the patient at the start (T_1) and at the end (T_2) of the rehab. For example, for patients of the *Trauma knee region* group the range of motion is recorded for T_1 and T_2, indicating the mobility of the knee joint before and after rehab. In this case, a higher value at T_2 would indicate a rehab success, as is outlined in more detail in Section 4.

The medical quality outcome measurements established in the performance profile of the Austrian social security institutions served as the basis for this work [14], i.e., serving as input variables. Based on a common and mandatory routine data collection process, patient-reported outcome measurements (PROMs) and clinician-reported measures (CROMs) were extracted from the electronic patient records to obtain data on pre-rehabilitation medical conditions and expected changes related to inpatient rehabilitation. The personal and health-related data were collected as part of routine medical care, as well as the quality assurance and evaluation of doctors and healthcare professionals, in accordance with national legislative guidelines. In addition, the gender and age of patients has been used.

3.1. Clinician reported outcome measures (CROMs)

CROM data contains variables which are assessed by a clinician, e.g., the range of motion (ROM) of the hip joint, the perimeter of the knee or the Timed Up and Go (TUG) test value.[2] The concrete CROMs used for each treatment group are listed in Table 1.

3.2. Patient reported outcome measures (PROMs)

PROM data refer to any completed standardized questionnaire that assesses whether there has been improvement in domains relevant to the outcome of treatment [42,43], with a specific focus on functional status and well-being [44]. PROMs can be divided into two categories: generic measures and specific measures. Generic measures are designed to summarize a spectrum of the concepts of health or quality of life that apply to many different impairments, patients, and populations [45].

Subjective methods of rating intensity and unpleasantness of pain include the visual analogue scale (VAS) [46]. Further generic PROMs include the Health Assessment Questionnaire (HAQ), which is based on 5 patient-centered dimensions: disability, pain, medication effects, costs of care, and mortality [47]. The European Quality of Life-5 Dimensions questionnaire (EQ-5D-5L) is a generic instrument that measures 5 dimensions of health status, each comprised of 5 levels: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression [48]. Physical disability was assessed using the Barthel Index. The Barthel Index is an index for the evaluation of activities of daily life (ADL), the need for care, and independence [33]. The EQ-5D-5L, the Barthel Index, the HAQ disability index, and the HAQ scale for patient global and VAS for pain were used in this study. Specific PROMs refer to a more detailed assessment of outcomes related to a particular injury or disease [49]. They use specific scores, which are not indicative of overall health. For example, the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC) was developed for patients with hip or knee osteoarthritis participating in clinical trials to measure 3 dimensions of pain (5 items), stiffness (2 items), and physical function (17 items) [50].

Both the objective clinician-reported measure (CROM) and subjective PROMs show characteristic changes during rehabilitation. The correlation between these methods is low [6,7].

Similarly to the group target values, the dataset contains T_1 and T_2 values for both CROMs and PROMs. For example, the TUG test is performed at the start and end of the rehab, and questionnaires are entered at each point in time.

Table 1
Utilized CROMs for each Treatment Group.

Treatment Group	CROMs
HIP_T, HIP_A	hip perimeter, ROM hip, TUG value
$KNEE_T$, $KNEE_A$	knee perimeter, ROM knee, TUG value
ANKLE	ankle joint perimeter, ROM ankle joint, TUG value

4. Methodology

The main goal of this study is to predict the success of the rehab of patients based on their health status at the start of the treatment. A general overview is illustrated in Fig. 1. To quantify the success and consequently also the quality of the prediction models, we compare the values of multiple target variables at the start (T_1) and end (T_2) of the rehab and compute a relative change value per patient.[3] In a first step, we utilize regressors to estimate this value, and further define outcome groups (e.g., 'moderate success' or 'significant success') that are used by classifiers to predict the success outcome group. In the following, we describe this procedure in detail, including the utilized features, algorithms, evaluation metrics and the general experimental setup.

4.1. Targets

As we examine the rehab of different treatment groups, we at first define the appropriate target values which should be predicted for each group. Generally, we consider three CROMs (ROM of the knee and hip, TUG value) and two PROMs (the sum scores of the HAQ and WOMAC questionnaires) as targets, i.e., their values at time point T_2. Corresponding possible ranges and optimal values are thereby as follows:

- ROM knee/hip: Standardized outcome measurements of the range of motion of hip and knee joints with a conventional goniometer are based on reference values of the generally accepted normal range of the American Academy of Orthopaedic Surgeons (AAOS active motion score [51–53]). It ranges between $[0\%, 100\%]$, where 100% is optimal (full range of motion).

[2] In orthopedic rehabilitation, there are several techniques for measuring the range of motion (ROM) of hip and knee joints, including estimation by an experienced examiner, using a short arm or a long arm goniometer, a digital goniometer, or a radiographic joint angle measurement [35,36]. The Timed Up and Go test (TUG) documents the time in seconds it takes a person to rise from a standard chair, walk to a line that is 3 meters away, turn 180 degrees, return to the chair, and sit down [37]. It was originally developed to identify elderly people with a fall risk and is used to assess a person's independent mobility based on both static and dynamic balance [38,39]. Prior to this study, the physicians and therapists who collected the outcome measures underwent standardized data collection training in order to obtain valid, reliable, and reproducible data of the ROM and the TUG [40,41].

[3] Note that both T_1 and T_2 values are already present in the dataset, but only T_1 values are needed by the final models to predict rehab success.

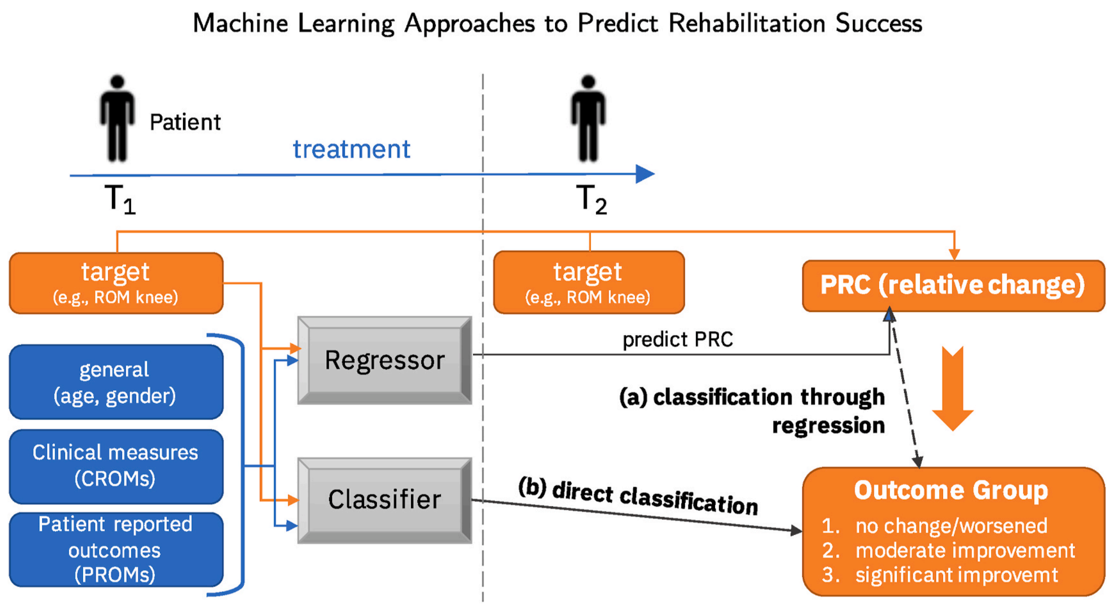

Fig. 1. Methodology Overview. We analyze general, CROM and PROM values at the start of rehab (T_1) and try to predict the outcome of specific targets at the end of rehab (T_2). Concretely, we compute a relative change for the patient (PRC) and categorize it into three outcome groups, which are predicted either implicitly through regressors (a) or directly with classifiers (b).

- TUG value: In theory it ranges between $(0, \infty)$ with 0 being optimal, but which is not achievable due to the test design (the maximum is infinite as patients may not complete the test at all). In real-life scenarios, a cut-off point around 10 seconds or less is suggested as normal for healthy community-dwelling elderly [39]. TUG results below 5 s are very rare in clinical practice and can be considered minimal.
- HAQ sum score (disability index): Per definition [47], it ranges between $[0, 3]$, where 0 is optimal (no disabilities).
- WOMAC sum score: Per definition [50], it ranges between $[0, 240]$, where 0 is optimal (no pain, stiffness or functional limitations).

Obviously, not every target is applicable for each treatment group, but we evaluated each possible combination as is summarized in Table 2.

4.1.1. Patient relative change

Medical outcome quality is defined as the "measurable change in the professionally assessed state of health, the quality of life and the satisfaction of a patient" [4], where the outcomes become visible by examining "the difference between the initial state and the state at treatment end" [41]. A comparison is thus possible between baseline rates before rehabilitation (T_1) and after rehabilitation rates (T_2).

Accordingly, we quantify the rehab success by comparing the value of the specific target at the start of the treatment with the respective value at the end (e.g., ROM hip at T_1 compared to ROM hip at T_2). Thereby the value at T_1 considerably influences the interpretation of "success" of the rehab, as for example the same improvement of the ROM may be moderate for a patient already starting at a high ROM, but may be significant for a patient starting at a low ROM. Consequently, a comparable success value, denoted in a normalized, percental manner is needed. From a medical point of view this value should additionally consider the following factors:

1. The target variable can be improved by at most 100%.
2. There is no limit to the worsening of the target variable, i.e., it can be worsened infinitely in theory.
3. The success of all target variables should be comparable, i.e., a positive value should always indicate an improvement. Concretely, we have to consider that a higher value for the ROM (hip and knee) is better than a low value, whereas on the other side for the TUG value and the sum scores of the questionnaires a lower value is better.

Incorporating these considerations, we finally define the patient relative change (PRC) value for a target X as follows:

$$PRC_X = \begin{cases} \dfrac{100 \cdot (X_{T_2} - X_{T_1})}{max(100 - X_{T_1}, 0.001)} & \text{if a higher value for } X \text{ is an improvement} \\ \dfrac{100 \cdot (X_{T_1} - X_{T_2})}{max(X_{T_1}, 0.001)} & \text{otherwise} \end{cases}$$

where X_{T_1} and X_{T_2} are the values of the target variable at time points T_1 and T_2, respectively. Note that we included the $max(\ldots, 0.001)$ function to cope with edge cases, where the value at the start of the rehab is already optimal (i.e., no further improvement is possible[4]), as otherwise PRC would be undefined due to a division by zero.

To illustrate the PRC with respect to different inputs, we show some examples:

1. **ROM knee** (range: $[0, 100]$, 100 is optimal)
   (a) improvement
   $\text{ROMknee}_{T_1} = 45$, $\text{ROMknee}_{T_2} = 66$
   $PRC_{\text{ROMknee}} = \frac{100 \cdot (66-45)}{100-45} = 38.2\%$
   (b) worsening

Table 2
Evaluated Target Variables with respect to each Treatment Group.

target/group	HIP_T	$KNEE_T$	ANKLE	HIP_A	$KNEE_A$
ROM knee		✓			✓
ROM hip	✓			✓	
TUG value	✓	✓	✓	✓	✓
HAQ sum score	✓	✓	✓	✓	✓
WOMAC sum score	✓	✓	✓	✓	✓

[4] This may be the case, as we evaluate multiple targets for a patient. E.g., a patient who is mainly treated for knee problems may still have an optimal HAQ sum score.

$\text{ROMknee}_{T_1} = 40, \text{ROMknee}_{T_2} = 25$

$PRC_{\text{ROMknee}} = \frac{100 \cdot (25-40)}{100-40} = -25\%$

2. **WOMAC sum score** (range: $[0, 240]$, 0 is optimal)
   (a) improvement

   $\text{WOMAC}_{T_1} = 30, \text{WOMAC}_{T_2} = 2$

   $PRC_{\text{WOMAC}} = \frac{100 \cdot (30-2)}{30} = 93.3\%$

   (b) worsening

   $\text{WOMAC}_{T_1} = 50, \text{WOMAC}_{T_2} = 120$

   $PRC_{\text{WOMAC}} = \frac{100 \cdot (50-120)}{50} = -140\%$

4.1.2. Outcome groups

The PRC value represents an exact and normalized number on how the rehab will affect the patient. Nevertheless, therapists are most often interested in a more coarse-grained view, i.e., if the rehab was successful, and if yes, was it significant or only a slight improvement. Based on the PRC value, we therefore introduce the outcome groups as shown in Table 3. They correspond to standardized effect with additional consideration of the initial value and optimum value, which includes the medical significance of clinically relevant changes.

As described later in this section, these outcome groups will be used by supervised machine learning algorithms to predict the group directly (using classification algorithms), but also indirectly (by estimating a continuous number with regression algorithms and subsequently categorizing it).

4.2. Feature sets and dataset preparation

The main input variables that have been used in this study are CROMs (medical data from different measurements), PROMs (data from completed questionnaires) and the two general variables age and gender. With respect to PROMs, the dataset contains sum scores (subscales) for each questionnaire as well as the answers for each single question. As is outlined later, we aim to find the best feature set combination for predicting rehab success and therefore perform a grid search evaluating multiple configurations. After discussions with the physicians responsible, we finally formed the feature sets as listed in Table 4. In the evaluation in Section 5 we refer to the features containing all PROM data as PROM_{all}, i.e., containing all features from PROM_{sum}, $\text{PROM}_{\text{detail}}$ and $\text{PROM}_{\text{ortho}}$.

4.2.1. Representation and normalization

As a commonly applied technique, we normalized all numerical features between $[0, 1]$. This includes all CROMs, but also several PROMs from questionnaires. For example, for a question like “How far can you walk without aid?” the answers (0 – “I cannot walk alone”), (1 – “50–100 meters”), (2 – “100–300 meters”), (3 – “more than 300 meters”) can be considered ordinal and can thus be normalized, too.

On the other side, for the gender as well as for categorical questions, i.e., where the answers do not imply any order, we applied one-hot encoding. Specifically, the following features are categorical: gender; ORTHO-BASIS items 2, 4-8, 10-13, 15-18, 21-25, 28-30, 33-34, 37-41, 43, 46, 47; Barthel Index items 1-3, 7; HAQ items 10-20, 31-38.

Table 3
Outcome Groups based on the PRC value.

Target	PRC	Outcome Group
ROM knee/hip, TUG value	$\leq 0\%$	no change or worsened (WO)
	$(0\%, 25\%]$	moderate improvement (MI)
	$> 25\%$	significant improvement (SI)
HAQ score, WOMAC score	$\leq 0\%$	no change or worsened (WO)
	$(0\%, 50\%]$	moderate improvement (MI)
	$> 50\%$	significant improvement (SI)

Table 4
Feature sets.

Set	#	Description
General	2	age, gender
CROM	13	all CROM data
PROM_{sum}	15	sum scores from all questionnaires except ORTHO-BASIS
$\text{PROM}_{\text{detail}}$	73	single questions from all questionnaires except ORTHO-BASIS
$\text{PROM}_{\text{ortho}}$	39	single questions from the ORTHO-BASIS questionnaire

Refers to the number of features contained in the respective set.

4.2.2. Cleaning

As we want to investigate predictions for different treatment groups individually, for which not all features are available coincidentally, we additionally clean the data according to the following strategy:

Given a treatment group T (e.g., Trauma hip region) and target X (e.g., ROM hip):

1. Remove all patients not belonging to T.
2. Remove all patients who do not have a value assigned for X.
3. Remove all features for which no value exists for more than 30% of the patients.
4. Remove all patients who do not have a value for more than 30% of the remaining features.
5. Replace missing numerical values with -1, and missing categorical values with a new “missing” class.

By doing so, we can ensure that the dataset finally used for performing the grid search is complete and correct as far as possible. Luckily, the dataset provided by the Rehab Center Kitzbühel turned out to be mostly complete anyway, i.e., for most of the patients all data points were available. In total, Table 5 shows the final remaining data sets with respect to treatment group and target value. Note that we also evaluated combinations of treatment groups with similar body regions, e.g., (HIP1, HIP2) containing patients from both groups.

4.3. Algorithms

In a first step, we use all available data at T_1 and apply regression algorithms to estimate the PRC value, i.e., the relative change in percent

Table 5
Final Dataset Sizes with respect to Targets and Treatment Groups.

Target	Group	Patients
ROM hip	HIP_T	115
	HIP_A	292
	$(\text{HIP}_\text{T}, \text{HIP}_\text{A})$	407
ROM knee	KNEE_T	84
	KNEE_A	406
	$(\text{KNEE}_\text{T}, \text{KNEE}_\text{A})$	490
TUG value	HIP_T	111
	KNEE_T	81
	ANKLE	80
	HIP_A	292
	KNEE_A	406
	all groups	996
HAQ sum score	HIP_T	115
	KNEE_T	89
	ANKLE	83
	HIP_A	287
	KNEE_A	403
	all groups	999
WOMAC sum score	HIP_T	113
	KNEE_T	89
	ANKLE	83
	HIP_A	287
	KNEE_A	404
	all groups	998

of a specific target from the start and end of a treatment. In this way we rely on commonly used algorithms: linear regression, Random Forest regressor [54], Extra Trees regressor [55], Linear Support Vector Regression (SVR) [56] and Kernel Ridge with a polynomial kernel [57].

Further, we project each predicted PRC value to the respective outcome group according to the thresholds listed in Table 3. For example, if the regressor estimates a PRC of 12.7% for ROM knee, the respective group would be 'moderate improvement'. By doing so, we are also able to compute classification metrics as described later on, allowing the results to be compared with those of direct classification.

4.3.1. Direct classification

Alternatively to utilizing regressors for a PRC value prediction, we apply classification algorithms to directly estimate the rehab success in terms of the categories listed in Table 3. In this case, we remove the PRC value from the dataset and replace it with the mapped outcome group as the target class. This means that the classifiers learn the models from the data points at T_1 in combination with the respective outcome group rather than the PRC value. In terms of algorithms, we refer to the following commonly used methods: Random Forest, Extra Trees, Support Vector Classification (SVC) [58] with a linear and nu-kernel [59], Naive Bayes [60] and Linear Discriminant Analysis [61].

In the following, we refer to the two approaches as *classification through regression (reg-cls)* and *direct classification (cls)*.

4.3.2. Baselines

Due to the lack of comparable approaches in the rehabilitation field, we compute the following baselines to estimate the quality of the prediction models:

- For regression, we use a dummy regressor which always predicts the mean PRC value with respect to the dataset in use. Consequently, a constant outcome group is predicted for every dataset, which is projected from the mean value.
- For direct classification, we similarly apply a stratified dummy classifier, which randomly predicts an outcome group by respecting the class distribution of the dataset.

Note that we explicitly refrained from utilizing deep learning techniques (i.e., various types of neural networks), as the dataset sizes are too small compared to the number of parameters [62].

4.4. Experimental setup

To find the best prediction models, we conduct a grid search for each target, treatment group and combination thereof, as listed in Table 5. For each grid search, we perform a 5-fold cross-validation with stratified training and test splits, to find the best hyperparameters for each machine learning algorithm. To measure the performance of the models, we rely on the mean average error (MAE) for regression and the F1 score for classification. According to the PCR value definition (see Section 4.1.1), its range is potentially $[-\infty; 100]$. Nevertheless, to put the MAE in perspective, we list the PRC value ranges of the targets as they appear in the dataset in the following: ROM hip $[-33.3, 70.9]$, ROM knee $[-100, 75]$, TUG value $[-41.2, 68]$, HAQ sum score $[-960, 100]$, WOMAC sum score $[-1010, 100]$. For example, a MAE of 35 for the HAQ sum score is substantially better than for ROM hip.

With respect to the F1 score, we chose to evaluate the macro and weighted variant. In our case, $F1_{macro}$ measures precision and recall for each outcome group and finally computes their unweighted mean, regardless of possible group imbalances. To incorporate the fact that the dataset indeed contains imbalances in outcome groups (the 'no change or worsened' group is especially underrepresented), we additionally compute $F1_{weighted}$, i.e., the weighted mean with respect to outcome group distributions. The weighted variant is also used as an optimization criterion during grid search.

5. Results

In this section, we systematically present the individual evaluation results. We at first compare the performances of the algorithms, and follow this with a comparison of the two evaluation types (i.e., classification through regression and direct classification), and finally present detailed results for each treatment group and target.

5.1. Algorithms in comparison

In general, the performance differences regarding $F1_{weighted}$ of the algorithms are similar over all targets and outcome groups, where it turns out that tree-based methods substantially outperform linear techniques. As a representative example, Table 6a shows the best classification-through-regression results for the target HAQ sum score (for all combined treatment groups). It can be seen that for this target, the Random Forest and Extra Trees regressors achieve the best results. A similar scenario can be seen when inspecting results of direct classification, where the Random Forest and Extra Trees classifiers also substantially outperform all other methods: Table 6b exemplarily shows the results for the TUG value and the treatment group HIP_T. Although not explicitly listed in this work, detailed experiments revealed that the performance differences for both classification through regression as well as direct classification are similar over all targets, i.e., the Random Forest and Extra Trees algorithms always perform best.

With respect to the two types of rehab success prediction, i.e., classification through regression and direct classification, the evaluation results clearly show that direct classification substantially outperforms classification through regression. As a representative example, we show the results of both types for the target ROM hip in Table 7. As stated before, we are more interested in predicting outcome groups in terms of significant, moderate or no success rather than estimating the concrete improvement in percent (PRC). Consequently, we only present the classification results for the remaining targets.

5.2. Target results

In the following, we present the evaluation results for each target and treatment group (combination). Specifically, Tables 7–11 list the best performing algorithms[5] and the corresponding feature sets, including

Table 6
Comparison of the Prediction Performances of the different Machine Learning Algorithms, exemplarily outlined by the HAQ score and TUG value Targets.

(a) Classification through Regression for the HAQ sum score. The respective PRC value range of the HAQ sum score is $[-960; 100]$.

Algorithm	MAE	$F1_{macro}$	$F1_{weighted}$
Random Forest Regressor	35.0	0.462	**0.464**
Extra Trees Regressor	34.0	0.462	**0.462**
Kernel Ridge	34.3	0.421	**0.424**
Linear Regression	38.1	0.412	**0.404**
Linear SVR	46.8	0.347	**0.347**

(b) Direct classification for the TUG value

Algorithm	$F1_{macro}$	$F1_{weighted}$
Random Forest Classifier	0.639	**0.669**
Extra Trees Classifier	0.599	**0.623**
SVC (nu)	0.496	**0.514**
SVC (linear)	0.468	**0.508**
Naive Bayes	0.410	**0.497**
Linear Discriminant Analysis	0.433	**0.484**

[5] Referring to Random Forest as RF and Extra Trees as ET.

Table 7
Comparison of Direct Classification (cls) and Classification through Regression (reg-cls). Exemplarily, the best Results for ROM hip are presented.

Group	Features	Algorithm	$F1_{weighted}$
HIP_T	CROM, general	RF (cls)	0.503
	-	*BASELINE (cls)*	0.254
	CROM, $PROM_{ortho}$	ET (reg-cls)	0.435
	-	*BASELINE (reg-cls)*	0.247
HIP_A	CROM, $PROM_{sum}$	RF (cls)	0.470
	-	*BASELINE (cls)*	0.202
	CROM	ET (reg-cls)	0.434
	-	*BASELINE (reg-cls)*	0.181
(HIP_T, HIP_T)	CROM, $PROM_{ortho}$	RF (cls)	0.448
	-	*BASELINE (cls)*	0.216
	CROM	ET (reg-cls)	0.400
	-	*BASELINE (reg-cls)*	0.213

RF: Random Forest, ET: Extra Trees.

Table 8
Best Direct Classification Results for ROM knee.

Group	Features	Algorithm	$F1_{weighted}$
$KNEE_T$	CROM, general	RF	0.565
	-	*BASELINE*	0.105
$KNEE_A$	CROM, $PROM_{sum}$, general	RF	0.595
	-	*BASELINE*	0.309
($KNEE_T$, $KNEE_A$)	CROM, $PROM_{sum}$	RF	0.586
	-	*BASELINE*	0.290

RF: Random Forest, ET: Extra Trees.

Table 9
Best Direct Classification Results for TUG value.

Group	Features	Algorithm	$F1_{weighted}$
HIP_T	CROM, general	RF	0.669
	-	*BASELINE*	0.306
$KNEE_T$	CROM, $PROM_{sum}$, general	RF	0.594
	-	*BASELINE*	0.286
ANKLE	CROM, $PROM_{sum}$, $PROM_{ortho}$	ET	0.600
	-	*BASELINE*	0.383
HIP_A	CROM, general	RF	0.570
	-	*BASELINE*	0.388
$KNEE_A$	CROM, general	RF	0.585
	-	*BASELINE*	0.405
all	CROM	RF	0.579
	-	*BASELINE*	0.358

RF: Random Forest, ET: Extra Trees.

the respective baselines to estimate the model quality. As stated before and exemplarily shown in Table 7, direct classification substantially outperforms the classification through regression method. Consequently, we only show the direct classification results for Tables 8–11.

In general, the $F1_{weighted}$ ranges from approximately 0.4 up to over 0.65, and in each case the baseline could be exceeded substantially. With respect to the best performing feature groups, CROM variables are almost always included, whereas PROM variables are frequently utilized for the HAQ and WOMAC scores, which reflects the fact that these targets result from questionnaires that were solely completed by patients. Finally, results show that it is of advantage in most cases to build models for each treatment group separately rather than learning one model for combined groups. For a more detailed evaluation of feature importances see Section 5.3.

To further understand the created models, we additionally visualize the normalized confusion matrices in Figs. 2–6. For example, Fig. 3a

Table 10
Best Direct Classification Results for the HAQ sum score.

Group	Features	Algorithm	$F1_{weighted}$
HIP_T	$PROM_{sum}$, $PROM_{detail}$, general	ET	0.625
	-	*BASELINE*	0.185
$KNEE_T$	CROM, $PROM_{all}$, general	ET	0.522
	-	*BASELINE*	0.320
ANKLE	$PROM_{sum}$, $PROM_{detail}$, general	ET	0.625
	-	*BASELINE*	0.359
HIP_A	CROM, $PROM_{sum}$, general	RF	0.583
	-	*BASELINE*	0.243
$KNEE_A$	CROM, $PROM_{all}$, general	ET	0.540
	-	*BASELINE*	0.156
all	CROM, $PROM_{all}$, general	RF	0.553
	-	*BASELINE*	0.171

RF: Random Forest, ET: Extra Trees.

Table 11
Best Direct Classification Results for the WOMAC sum score.

Group	Features	Algorithm	$F1_{weighted}$
HIP_T	CROM, $PROM_{all}$	RF	0.545
	-	*BASELINE*	0.230
$KNEE_T$	$PROM_{sum}$	RF	0.468
	-	*BASELINE*	0.274
ANKLE	$PROM_{all}$, general	ET	0.601
	-	*BASELINE*	0.305
HIP_A	CROM, $PROM_{sum}$, $PROM_{ortho}$, general	RF	0.493
	-	*BASELINE*	0.181
$KNEE_A$	CROM, $PROM_{all}$, general	ET	0.465
	-	*BASELINE*	0.297
all	CROM, $PROM_{all}$, general	RF	0.471
	-	*BASELINE*	0.240

RF: Random Forest, ET: Extra Trees.

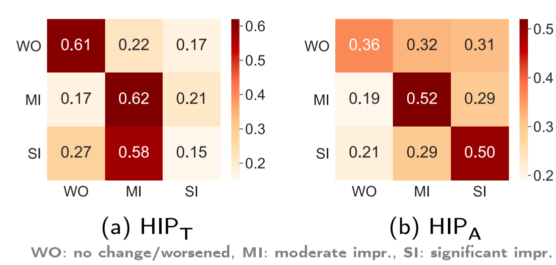

Fig. 2. Normalized Confusion Matrices for ROM hip.

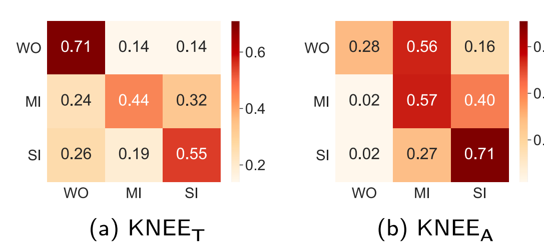

Fig. 3. Normalized Confusion Matrices for ROM knee.

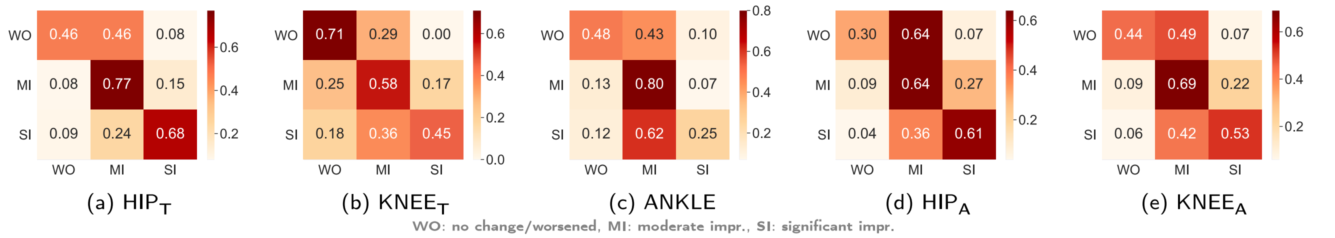

Fig. 4. Normalized Confusion Matrices for the TUG value.

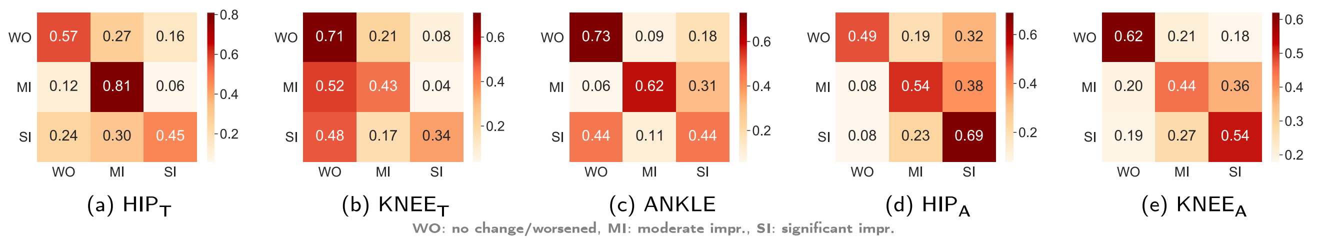

Fig. 5. Normalized Confusion Matrices for the HAQ value.

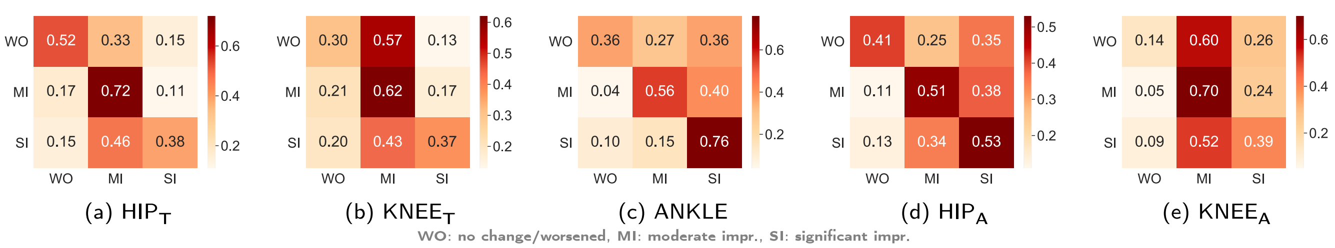

Fig. 6. Normalized Confusion Matrices for the WOMAC value.

shows that 71% of the patients achieving no rehab success or a worsened condition have been correctly assigned, whereas on the other side the model classified only 44% of the patients with moderate improvement correctly, i.e., predicting 24% to worsen and 32% to significantly improve. In general, in the case of individual group predictions the results also vary for each target and treatment group, whereby the correct groups (in the diagonal) show the highest values with an accuracy of up to 81% (see Fig. 5). Note that in this study we did not incorporate the fact that misclassifications may not have the same weight, e.g., classifying a patient who actually improved significantly to achieve only moderate rehab success is clearly better than predicting him/her to worsen (see Section 6 for a more elaborate discussion).

5.3. Feature importances

All the evaluation scenarios achieved their best results using either Random Forest or Extra Trees classifiers (e.g., about 15% better with respect to $F1_{weighted}$ than the support vector machine in the example shown in Table 6b). As both methods are tree-based, we are consequently able to look deeper into the importances of the respective input variables, countering the argument that machine learning in healthcare "comes at the expense of opacity" [63]. To this end, Table 12 depicts the three most important features for each scenario, including their percental importance weight.[6] For PROM variables, *item n* refers to the n^{th} question in the corresponding questionnaire, e.g., HAQ item 25 represents the 25^{th} question of the HAQ 1.0 questionnaire (see Section 3). Interestingly, the age of the patients seem to play a non-negligible role for the rehab success in many cases, whereas the gender does not.

6. Conclusion and future work

In this study, we aimed to predict the potential success of the treatment of knee, hip and foot rehab patients. Therefore, we utilized a real-life, anonymized dataset containing multiple clinical (CROM) and patient-reported (PROM) variables of already completed treatments, stating their progression throughout rehab. After computing a relative success value for each patient and predefined target variable, state-of-the-art regression and classification algorithms were utilized to finally predict the rehab success in terms of a three-classed grading scheme (outcome groups). Individual evaluations for each treatment group and target show that direct classification performs substantially better than using regression algorithms beforehand. In summary, weighted F1 scores from 40% to over 65% could be achieved, whereby simple baselines were exceeded substantially by utilizing the tree-based algorithms Random Forest and Extra Trees. Further investigations on the feature importances indicate that not only are physical parameters like the range of motion of knees important for the prediction, but also the age of the patients together with their self-reported well-being by utilizing questionnaires.

Considering this study as a first pilot work in the area of rehab success prediction, we want to emphasize that the utilized data is collected from patients of a particular geographical region. Although we think that the results of this study are transferable also to other regions (i.e.,

[6] The importances of all input features sum up to 100%. Note that the values of the most important features are usually substantially lower when a high number of features is used.

Table 12
Feature importances. For each scenario, it lists the top three variables utilized by the best prediction model. The corresponding F1 scores are listed in Tables 7–11.

Target PRC	Group	Most Important Features (at T_1)
ROM hip	HIP_T	hip perimeter (26%), age (25%), ROM hip (22%)
	HIP_A	ROM hip (11%), EQ-VAS (8%), hip perimeter (8%)
ROM knee	$KNEE_T$	ROM knee (32%), age (24%), knee perimeter (19%),
	$KNEE_A$	ROM knee (17%), knee perimeter (7%), HAQ sum score (7%)
TUG value	HIP_T	TUG value (31%), age (24%), hip perimeter (19%)
	$KNEE_T$	TUG value (14%), WOMAC pain score (8%), WOMAC sum score (8%)
	ANKLE	ROM ankle joint (7%), WOMAC stiffness score (5%), ankle joint perimeter (5%)
	HIP_A	TUG value (31%), age (23%), hip perimeter (23%)
	$KNEE_A$	TUG value (32%), knee perimeter (23%), age (21%)
HAQ sum score	HIP_T	HAQ item 25 (3%), WOMAC sum score (2%), WOMAC item 13 (1%)
	$KNEE_T$	HAQ sum score (2%), HAQ item 3 (1%), TUG value (1%)
	ANKLE	EQ-5D mobility score (5%), HAQ sum score (3%), HAQ item 30 (2%)
	HIP_A	HAQ sum score (9%), TUG value (7%), age (7%)
	$KNEE_A$	HAQ item 30 (2%), HAQ sum score (2%), HAQ item 29 (1%)
WOMAC sum score	HIP_T	WOMAC sum score (2%), hip perimeter (2%), WOMAC ADL score, (2%)
	$KNEE_T$	WOMAC sum score (11%), EQ-VAS score (11%), EQ-5D general score, (11%)
	ANKLE	WOMAC item 23 (2%), WOMAC stiffness score, (2%), WOMAC item 13 (2%)
	HIP_A	WOMAC ADL score (7%), WOMAC sum score (7%), WOMAC pain score (6%)
	$KNEE_A$	WOMAC sum score (2%), WOMAC ADL score (2%), WOMAC item 7 (1%)

are independent of it), this hypothesis should be evaluated in a proper study. Other than that, multiple future studies are possible. At first, as the performance of machine learning algorithms generally increases with the amount of data available for training, more data coming from newly completed treatments should further increase the quality of the models. Moreover, it should sharpen the view with respect to important rehab success factors, allowing physicians to react correspondingly in an already early treatment phase. For example, a patient may be treated differently if the model predicts that s/he will not respond to common techniques. Finally, the availability of sufficient data could lead to deep learning techniques being utilized and evaluated, as well as the possibility of creating and learning from 'digital twins' [64], i.e., patients with (very) similar conditions.

To further build a more realistic prediction system, the thresholds for outcome groups or even the groups themselves may be refined in close interchange with physicians responsible. Moreover, it should be defined which misclassifications are penalized to which extent. I.e., machine learning algorithms should learn their models based on the fact that, e.g., predicting a medium success is better than predicting a significant success, if the values of the patient actually worsened. By defining corresponding criteria, models could then be built which admittedly may not be able to exactly differentiate between good or very good treatment success, but which have a high precision in estimating whether the treatment can be successful at all.

Furthermore, future work could incorporate the trained models in real-life settings: Once deployed, the models would then be able to give an estimation of the rehab success directly after the patient's first assessments. Finally, such a system could also be built to allow physicians to report misclassifications, information which subsequently could be used to readjust the models, possibly even in real time.

Acknowledgements

Trial registration

This clinical study was entered retrospectively on August 14, 2020 in the German register for clinical studies (registration number: DRKS00022854).

Ethical aspects

The ethics committee of the Medical University of Innsbruck approved the study protocol on August 23, 2019 (Ref: EK Nr:1158/2019). Person-related and health-related data were collected as part of routine medical care and quality management in compliance with all regulations of the Austrian Privacy Act, and in accordance with the Declaration of Helsinki in the currently valid version and the national legislation.

Data availability statement

The datasets analyzed and referred to in this manuscript are not publicly available. The authors can provide descriptive data on individual medical indicators for admission and discharge or the expected change due to inpatient rehabilitation for various groups and diagnoses upon request. Requests to access anonymized datasets should be directed to the corresponding author.

Consent for publication

All authors provided their consent to submit and publish the final version of this manuscript.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the Rehabilitation Center in Kitzbühel under office@reha-kitz.at on reasonable request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

Each author of our work was significantly involved in the conception, design, data acquisition, data analysis and interpretation. All authors contributed to the writing of the manuscript and have released the final version for publication. All authors take responsibility for the accuracy and integrity of all aspects of the research.

The authors would like to thank the team of the Rehabilitation Center in Kitzbühel, Martin Stockinger, the VAMED AG as well as the Ludwig Boltzmann Society for their support in the development and execution of this study.

References

[1] The L. Gbd 2015: from big data to meaningful change. Lancet 2016;388(10053):1447.
[2] Blyth FM, Briggs AM, Schneider CH, Hoy DG, March LM. The global burden of musculoskeletal pain—where to from here? Am J Publ Health 2019;109(1):35–40.

[3] Briggs AM, Woolf AD, Dreinhöfer K, Homb N, Hoy DG, Kopansky-Giles D, Åkesson K, March L. Reducing the global burden of musculoskeletal conditions. Bull World Health Organ 2018;96(5):366.
[4] F. Bachner, J. Bobek, K. Habimana, J. Ladurner, L. Leuschutz, H. Ostermann, L. Rainer, A. Schmidt, M. Zuba, W. Quentin, et al., Austria: health system review.
[5] Austria Statistik. Hospital discharge statistics 2018. 2019. Press release: 12.130-196/19.
[6] Grote V, Unger A, Böttcher E, Muntean M, Puff H, Marktl W, Mur E, Kullich W, Holasek S, Hofmann P, et al. General and disease-specific health indicator changes associated with inpatient rehabilitation. J Am Med Dir Assoc 2020.
[7] Grote V, Unger A, Puff H, Böttcher E. What to expect: medical quality outcomes and achievements of a multidisciplinary inpatient musculoskeletal system rehabilitation. In: Physical therapy effectiveness. London, UK: IntechOpen; 2020.
[8] Bethge M, Müller-Fahrnow W. Wirksamkeit einer intensivierten stationären Rehabilitation bei muskuloskelettalen Erkrankungen: systematischer Review und Meta-Analyse. Rehabil 2008;47:200–9. 04.
[9] Di Monaco M, Castiglioni C. Which type of exercise therapy is effective after hip arthroplasty? a systematic review of randomized controlled trials. Eur J Phys Rehabil Med 2013;49(6):893–907.
[10] Mak JC, Fransen M, Jennings M, March L, Mittal R, Harris IA. Evidence-based review for patients undergoing elective hip and knee replacement. ANZ J Surg 2014;84(1–2):17–24.
[11] Schwarz B, Neuderth S, Gutenbrunner C. Multiprofessional teamwork in work-related medical rehabilitation for patients with chronic musculoskeletal disorders. J Rehabil Med 2015;47(1):58–65.
[12] Momsen A-M, Rasmussen JO, Nielsen CV, Iversen MD, Lund H. Multidisciplinary team care in rehabilitation: an overview of reviews. J Rehabil Med 2012;44(11): 901–12.
[13] Skinner A, Turner-Stokes L. The use of standardized outcome measures in rehabilitation centres in the UK. Clin Rehabil 2006;20(7):609–15.
[14] M. Gyimesi, G. Fülöp, S. Ivansits, E. Pochobradsky, A. Stoppacher, S. Kawalirek, A. Maksimovic, Rehabilitationsplan 2016, Hauptverband der österreichischen Sozialversicherungsträger 273.
[15] Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J Am Med Inf Assoc 2020;27(4):592–600.
[16] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25(1):44–56.
[17] Norgeot B, Glicksberg BS, Butte AJ. A call for deep-learning healthcare. Nat Med 2019;25(1):14–5.
[18] Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, Cui C, Corrado G, Thrun S, Dean J. A guide to deep learning in healthcare. Nat Med 2019; 25(1):24–9.
[19] Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P. Deep learning algorithms for detection of critical findings in head ct scans: a retrospective study. Lancet 2018;392(10162):2388–96.
[20] Bejnordi BE, Veta M, Van Diest PJ, Van Ginneken B, Karssemeijer N, Litjens G, Van Der Laak JA, Hermsen M, Manson QF, Balkenhol M, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama 2017;318(22):2199–210.
[21] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542 (7639):115–8.
[22] Zhang J, Gajjala S, Agrawal P, Tison GH, Hallock LA, Beussink-Nelson L, Lassen MH, Fan E, Aras MA, Jordan C, et al. Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. Circulation 2018;138(16):1623–35.
[23] de Langavant LC, Bayen E, Yaffe K. Unsupervised machine learning to identify high likelihood of dementia in population-based surveys: development and validation study. J Med Internet Res 2018;20(7):e10493.
[24] Yang Z, Huang Y, Jiang Y, Sun Y, Zhang Y-J, Luo P. Clinical assistant diagnosis for electronic medical record based on convolutional neural network. Sci Rep 2018;8 (1):1–9.
[25] Elfiky AA, Pany MJ, Parikh RB, Obermeyer Z. Development and application of a machine learning approach to assess short-term mortality risk among patients with cancer starting chemotherapy. JAMA Netw Open 2018;1(3). e180926–e180926.
[26] Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. BMC Med Inf Decis Making 2018;18(4):122.
[27] Mathotaarachchi S, Pascoal TA, Shin M, Benedet AL, Kang MS, Beaudry T, Fonov VS, Gauthier S, Rosa-Neto P, Initiative ADN, et al. Identifying incipient dementia individuals using machine learning and amyloid imaging. Neurobiol Aging 2017;59:80–90.
[28] Walsh CG, Ribeiro JD, Franklin JC. Predicting risk of suicide attempts over time through machine learning. Clin Psychol Sci 2017;5(3):457–69. https://doi.org/10.1177/2167702617691560.
[29] Zhu M, Chen W, Hirdes JP, Stolee P. The k-nearest neighbor algorithm predicted rehabilitation potential better than current clinical assessment protocol. J Clin Epidemiol 2007;60(10):1015–21.
[30] Zhu M, Zhang Z, Hirdes JP, Stolee P. Using machine learning algorithms to guide rehabilitation planning for home care clients. BMC Med Inf Decis Making 2007;7 (1):41.
[31] Zhu M, Cheng L, Armstrong JJ, Poss JW, Hirdes JP, Stolee P. Using machine learning to plan rehabilitation for home care clients: beyond "black-box" predictions. In: Machine learning in healthcare informatics. Springer; 2014. p. 181–207.
[32] Lin W-Y, Chen C-H, Tseng Y-J, Tsai Y-T, Chang C-Y, Wang H-Y, Chen C-K. Predicting post-stroke activities of daily living through a machine learning-based approach on initiating rehabilitation. Int J Med Inf 2018;111:159–64.
[33] Mahoney FI, Barthel DW. Functional evaluation: the barthel index: a simple index of independence useful in scoring improvement in the rehabilitation of the chronically ill. Md State Med J 1965;6:493–507.
[34] Huber M, Kurz C, Leidl R. Predicting patient-reported outcomes following hip and knee replacement surgery using supervised machine learning. BMC Med Inf Decis Making 2019;19(1):3. https://doi.org/10.1186/s12911-018-0731-6.
[35] Nussbaumer S, Leunig M, Glatthorn JF, Stauffacher S, Gerber H, Maffiuletti NA. Validity and test-retest reliability of manual goniometers for measuring passive hip range of motion in femoroacetabular impingement patients. BMC Muscoskel Disord 2010;11(1):194.
[36] Kim S-G, Kim E-K. Test-retest reliability of an active range of motion test for the shoulder and hip joints by unskilled examiners using a manual goniometer. J Phys Ther Sci 2016;28(3):722–4.
[37] Huber EO, Meichtry A, de Bie RA, Bastiaenen CH. Construct validity of change scores of the chair stand test versus timed up and go test, koos questionnaire and the isometric muscle strength test in patients with severe knee osteoarthritis undergoing total knee replacement. Man Ther 2016;21:262–7.
[38] Wall JC, Bell C, Campbell S, Davis J. The timed get-up-and-go test revisited: measurement of the component tasks. J Rehabil Res Dev 2000;37(1).
[39] Bischoff HA, Stähelin HB, Monsch AU, Iversen MD, Weyh A, Von Dechend M, Akos R, Conzelmann M, Dick W, Theiler R. Identifying a cut-off point for normal mobility: a comparison of the timed 'up and go'test in community-dwelling and institutionalised elderly women. Age Ageing 2003;32(3):315–20.
[40] Bakar Y, Özdemir Ö, Sevim S, Duygu E, Tuğral A, Sürmeli M. Intra-observer and inter-observer reliability of leg circumference measurement among six observers: a single blinded randomized trial. J Med Life 2017;10(3):176.
[41] Holla JF, van der Leeden M, Roorda LD, Bierma-Zeinstra SM, Damen J, Dekker J, Steultjens MP. Diagnostic accuracy of range of motion measurements in early symptomatic hip and/or knee osteoarthritis. Arthritis Care Res 2012;64(1):59–65.
[42] Porter ME, Larsson S, Lee TH. Standardizing patient outcomes measurement. N Engl J Med 2016;374(6):504–6.
[43] Wolpert M. Uses and abuses of patient reported outcome measures (proms): potential iatrogenic impact of proms implementation and how it can be mitigated. Adm Pol Ment Health 2014;41(2):141–5.
[44] Weldring T, Smith S. Patient-reported outcomes (pros) and patient-reported outcome measures (proms) health serv insights 2013;6:61–8. https://doi.org/10.4137/hsi.s11093.
[45] Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. Med Care 1989:S217–32.
[46] Price DD, McGrath PA, Rafii A, Buckingham B. The validation of visual analogue scales as ratio scale measures for chronic and experimental pain. Pain 1983;17(1): 45–56.
[47] Bruce B, Fries JF. The stanford health assessment questionnaire: dimensions and practical applications. Health Qual Life Outcome 2003;1(1):20.
[48] Rabin R, Charro Fd. Eq-sd: a measure of health status from the euroqol group. Ann Med 2001;33(5):337–43.
[49] Brinker MR, O'Connor DP. Stakeholders in outcome measures: review from a clinical perspective. Clin Orthop Relat Res 2013;471(11):3426–36.
[50] McConnell S, Kolopack P, Davis AM. The western ontario and mcmaster universities osteoarthritis index (womac): a review of its utility and measurement properties. Arthritis Care Res: Off J Am Coll Rheumatol 2001;45(5):453–61.
[51] Skinner SB, McVey C. Pocket notes for the physical therapist assistant. Jones & Bartlett Publishers; 2012.
[52] Heck CV, Hendryson IE, Rowe CR. Joint motion: method of measuring and recording. Chicago, IL: American Academy of Orthopedic Surgeons; 1965. p. 30–42.
[53] Roaas A, Andersson GB. Normal range of motion of the hip, knee and ankle joints in male subjects, 30–40 years of age. Acta Orthop Scand 1982;53(2):205–8.
[54] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[55] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63 (1):3–42.
[56] Vapnik V. The nature of statistical learning theory. Springer science & business media; 2013.
[57] Murphy KP. Machine learning: a probabilistic perspective. MIT press; 2012.
[58] Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory; 1992. p. 144–52.
[59] Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. Neural Comput 2000;12(5):1207–45.
[60] John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In: Proceedings of the 11th conference on uncertainty in artificial intelligence. Montreal, Canada: Morgan Kaufmann Publishers Inc.; 1995. p. 338–45.
[61] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media; 2009.
[62] Bartlett PL, Maass W. Vapnik-chervonenkis dimension of neural nets. In: The handbook of brain theory and neural networks; 2003. p. 1188–92.
[63] Grote T, Berens P. On the ethics of algorithmic decision-making in healthcare. J Med Ethics 2020;46(3):205–11.
[64] Tarassenko L, Topol EJ. Monitoring jet engines and the health of people. J Am Med Assoc 2018;320(22):2309–10. https://doi.org/10.1001/jama.2018.16558.