User Modeling and User-Adapted Interaction (2022) 32:1019–1062
https://doi.org/10.1007/s11257-021-09298-4

A pharmaceutical therapy recommender system enabling shared decision-making

Felix Gräßer[1] · Falko Tesch[2] · Jochen Schmitt[2] · Susanne Abraham[3] · Hagen Malberg[1] · Sebastian Zaunseder[4]

Received: 12 June 2020 / Accepted in revised form: 17 July 2021 / Published online: 5 August 2021

Abstract
Data-based clinical decision support systems (CDSSs) can provide personalized support in medical applications. Such systems are expected to play an increasingly important role in the future of healthcare. Within this work, we demonstrate an exemplary CDSS which provides individualized pharmaceutical drug recommendations to physicians and patients. The core of the proposed system is a neighborhood-based collaborative filter (CF) that yields data-based recommendations. CFs are capable of integrating data at different scale levels and a multivariate outcome measure. This publication provides a detailed literature review, a holistic comparison of various implementations of CF algorithms, and a prototypical graphical user interface (GUI). We show that similarity measures, which automatically adapt to attribute weights and data distribution perform best. The illustrated user-friendly prototype is intended to graphically facilitate explainable recommendations and provide additional evidence-based information tailored to a target patient. The proposed solution or elements of it, respectively, may serve as a template for future CDSSs that support physicians to identify the most appropriate therapy and enable a shared decision-making process between physicians and patients.

Keywords Health recommender systems · Therapy recommender systems · Drug recommender systems · Clinical decision support systems · Real world evidence · Shared decision-making · Collaborative filtering · Patient similarity

✉ Felix Gräßer
felix.graesser@tu-dresden.de

Extended author information available on the last page of the article

1 Introduction

1.1 Clinical decision-making

The ability to make accurate and timely treatment decisions is a core skill and critical aspect of physician performance in medical practice (Croskerry 2009; Groves 2012). Based on diagnosis and additional patient risk factors, such as demographic data, comorbidities, and life situation, the attending physician is tasked to make an estimation on the natural history of a disease and to predict the response to possible treatment options for a patient and time (Del Mar et al. 2007). Outcome, however, is typically multifactorial (Calero Valdez et al. 2016), meaning that multiple aspects, such as benefits and harms, are to be considered. At the same time, additional factors such as costs and the way of application determine the treatment decision. A precise definition of the targeted outcome (Kaplan and Frosch 2005) and an accurate prognosis are the foundation of optimal treatment decisions.

Depending on condition and indication, a great variety of pharmaceutical drugs and drug combinations may be available. Consequently, selecting the potentially most appropriate therapy option for an individual patient poses a challenging task to prescribers. As a result, treatment choices are often subjective and cognitive biases (Avorn 2018; Croskerry 2009; Trimble and Hamilton 2016), a high inter-rater and intra-rater variability (Kaplan and Frosch 2005), conflicts of interests (Larkin et al. 2017) and errors ("To Err is Human" (IOM 1999)) are daily fare. Assuming that one optimal treatment for a patient and time exists, the aforementioned factors suggest that many patients are not treated optimally.

Moreover, patients self-perception shifts toward a more active role and the desire to be engaged in a participative decision-making process (*patient empowerment*) (Sim 2001; Kaplan and Frosch 2005; Barratt 2008). Trade-offs need to be found between the medical requirements and the patients' preferences and expectations to support patients' satisfaction and adherence to treatment. Explainability of treatment decisions becomes an increasingly important factor. Physicians not only need to decide on one treatment but will be increasingly requested to clarify decisions and to provide detailed prognoses for the full range of options.

1.2 Evidence-based medicine

To reduce medication errors and remedy the stated inconsistency of treatment choices, evidence-based medicine (EbM) and evidence-based guidelines are supposed to supplement a physician's opinion with the best available external evidence from the scientific literature. EbM and guidelines, however, are susceptible to various issues. Clinical studies, which evidence is based on, often lack generalizability. In particular, the presence of multimorbidity and polypharmacy can lead to differing therapy outcomes and increases the risk of drug interactions, adverse or unforeseen effects, or contraindications (Fortin et al. 2006; Campbell-Scherer 2010; Frankovich et al. 2011; Faries et al. 2013; Longhurst et al. 2014; Sönnichsen et al. 2016).

Potential differences between clinical study collectives and real patient collectives, but also long-term effects, are often insufficiently evaluated before market introduction which makes pharmacovigilance an important process for drug safety. Moreover, clinical study endpoints frequently differ from the patients' actual objectives such as Patient Reported Outcomes (PROs).

1.3 Data-driven decision support

To seamlessly integrate the most recent evidence from literature and guidelines into the clinical work process, appropriate technical tools are not yet available. Beyond that as stated above, the selection of patient-specific therapy options often cannot be provided on the basis of evidence from the literature and guidelines alone. An obvious way to address these challenges is to complement this *external evidence* by clinical experience from past patient encounters and routine care, which is stored in local or global data bases such as electronic health records (EHRs) and clinical registries. Exploiting such *practice-based* (Sim 2001) or *real-world evidence* (Sherman et al. 2016) facilitates the attending physician with empirical experience and supplements external evidence where evidence from the literature is missing, inappropriate, or inaccessible (Frankovich et al. 2011; Gallego et al. 2015; Celi et al. 2014; Longhurst et al. 2014). Data-driven approaches, which assist by incorporating such information into treatment decision-making, can be expected to play a significant role in future healthcare.

To date, data-driven clinical decision support systems (CDSSs) are rare in clinical practice. This is certainly due to challenges regarding integration of such systems into the clinical workflow and due to challenges related to access the relevant clinical data in a processable format. Lacking interpretability and explainability of recommendations further hamper acceptance of such systems. Particularly, collaborative filter (CF) algorithms, which are widely and successfully applied in other applications, such as e-commerce or music and movie streaming services, have many obvious analogies with the therapy recommendation setting as outlined above. A large number of optional items are ranked according to personalized preference predictions. Here, treatment options can be regarded as items and any of the multifactorial outcome indicators as preference. The potentially most successful treatment with respect to an addressed outcome objective can be recommended to a physician and patient. By providing outcome predictions for each option and aspect, the final treatment decision can be made together with the patient and with special focus on his or her values and objectives. Moreover, especially neighborhood-based CF methods have the additional feature of being very intuitive. Predictions and recommendations are transparent and explainable in terms of the included neighboring consultations. On the one hand, this neighborhood can be inspected directly if kept at a moderate size. On the other hand, the computation of local summary statistics or a "Prototype Patient" can be supplementary or alternative means of providing insight into the outcome prediction and recommendation process.

1.4 Scope of this paper

Overall, we envision a CDSS as schematized in Fig. 1 that supports with clinical decision-making by integrating multiple sources of information such as (collective) clinical experience stored in health records or clinical registries and clinical evidence from scientific literature, expert information and advisory platforms, respectively. However, also patient reviews captured by online pharmacies or drug rating portals can be included as valuable source of patient experience, e.g., by means of sentiment analysis methods (Gräßer et al. 2018). This vision of a CDSS implements a closed loop in order to feedback treatment decisions and outcome. Consequently, this interactive machine learning (iML) approach (Holzinger 2016), encompassing a doctor-in-the-loop (DiL), facilitates a learning therapy recommender system which continuously improves by extending the clinical databases and adapts to applied research and pharmacovigilance findings.

Within this work, we present a comparative study of various implementations of a data-driven therapy recommender system. The applied methodologies exploit (phenotypic) patient characteristics and information on outcome of previously applied treatments. This data is considered to capture (collective) clinical experience concerning therapy options. On the basis of this previous experience, the goal is to provide *patient-specific* therapy recommendations which are optimized for a given patient and time considering his or her individual characteristics. Therefore, we transfer and adapt methodologies from CF research to the therapy recommendation application. In order to illustrate the recommender system's intended use, we present a prototype and Graphical User Interface (GUI) concept which addresses—based on the CF recommendation algorithms—explainable recommendations and features shared decision-making as introduced above. Beyond that the demonstrated prototype provides evidence-based information regarding treatment options, which is tailored to a target patient and consultation. The exemplary application focuses on

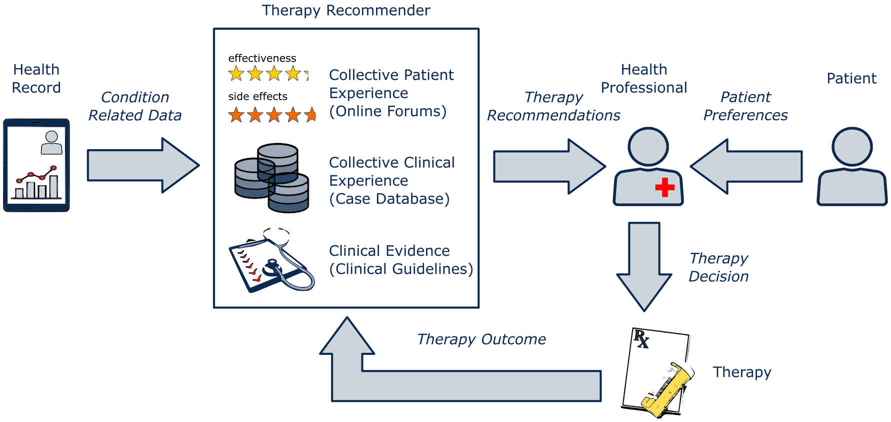


Fig. 1 Therapy recommender system framework integrating multiple sources of information and encompassing a DiL

the chronic inflammatory skin disease *Psoriasis*, however, is intended to be transferable to other conditions as well.

Basic ideas of this work were initially demonstrated in our previous publications (Gräßer et al. 2017, 2019). Within this paper, we demonstrate the first time a comprehensive evaluation and comparison of various implementations and the integration into an overall prototypical therapy recommender system.

1.5 Organization of this paper

The remainder of this article is organized as follows. Related works are summarized in Sect. 2. The exemplary application and available data are detailed in Sect. 3. Sections 4 and 5 provide an overview on the characteristics of the implemented CF algorithms and its results, respectively. Algorithmic details can be found in Appendix A. In Sect. 6, the fundamental requirements and ideas behind the GUI concept are presented. A detailed description and screenshots are included in Appendix B. Finally, Sects. 7 and 8 discuss benefits and shortcomings and provide some conclusions, respectively.

2 Related work

CDSS in general are broadly defined as computer systems which are designed to aid clinical decision-making by providing patient-specific assessments or recommendations at the point in time that these decisions are made (Berner and La Lande 2016).

Several essential characteristics, which determine acceptance and successful application of CDSSs, are described in the literature. According to the analyses by Kawamoto (2005), decision support must be actionable recommendations, rather than just assessments and must be provided automatically as part of the clinician's workflow at the time and location of decision-making. Moreover, the decision support should be provided timely and be accurate, interpretable, and tailored to the current needs (Beeler et al. 2014). Finally, the growing engagement of patients in clinical decision-making should be considered (Sim 2001; Kaplan and Frosch 2005).

Research on CDSSs in general has emerged from earlier Artificial Intelligence research, which aimed to design computer programs to simulate human decision-making (INTERNIST-I (Miller et al. 1982), MYCIN (Shortliffe 2012), DXplain (Barnett et al. 1987)). Today, the literature describes a mass of CDSSs varying greatly in design, function and use (Shortliffe 1987; Garg et al. 2005; Wright et al. 2011; Berner and La Lande 2016; Pandey and Mishra 2009; Sutton et al. 2020). Berner and La Lande (2016) and Pandey and Mishra (2009) distinguish CDSS approaches according to their implementation properties into *knowledge-based*, which typically consist of compiled rules or probabilistic associations, and *non-knowledge-based* approaches, which apply machine learning (ML) or other statistical pattern recognition methods to automatically learn from past experiences stored in the clinical data. The first approach features explainability but relies on complete and accurate knowledge bases which are difficult to obtain and keep to up-to-date. The latter approach, on the other hand, has the potential to automatically reveal knowledge from data and adapt to changes. However, ML methods typically require large amounts of data to build reliable models and decisions are often difficult to interpret. Within the context of *non-knowledge-based*, data-driven approaches, especially the information captured in EHRs is expected to play an important role in the future of healthcare CDSSs (Haas 2005; Gallego et al. 2015).

Several works propose to mimic personalized observational studies by dynamically identifying a subgroup of patients in the database of past patient encounters (Frankovich et al. 2011; Leeper et al. 2013; Gallego et al. 2015). Such virtual cohorts of similar patients can be assumed to be more likely to represent a realistic population with similar characteristics than those assembled for clinical trials (Gallego et al. 2015). Clinicians using an EHR ideally generate such *practice-based evidence* as a by-product of routine health care. Longhurst et al. (2014) propose a *Green Button* which is intended to provide real-time and personalized practice-based evidence and treatment recommendations by systematic analysis, appraisal, and presentation of such observational experience.

Deriving recommendations based on the collective experience and preferences of users is a widely and successfully used techniques to support users with the decision-making task in multiple, especially online applications (Ricci et al. 2011; Su and Khoshgoftaar 2009). Recommender System (RS) algorithms which rely on a subset of similar users, namely *neighborhood-based* CF, are especially popular and successfully applied (Ricci et al. 2011; Su and Khoshgoftaar 2009). The underlying concept is very similar to the identification of similar patients to obtain practice-based evidence from, as described above.

Various works employ RS-related techniques for health applications, denoted as Health Recommender Systems (HRSs), which are summarized in Table 1. The publications can be broadly categorized into reviews or frameworks on the one hand, and approaches addressing the objectives adverse drug event and side effect prediction and prevention, outcome prediction and therapy recommendation, and disease risk stratification, on the other hand. As can be seen, only ten works can be categorized into the group of treatment recommendations including those recommending clinical orders in general. In this group, in turn, no work deals with the recommendation of pharmaceutical treatments exclusively. Four works in this group use treatment or clinical order history, and six works use patient data as basis for recommendations. In summing up, it can be said that the application of RS methods for treatment recommendations, especially in the domain of pharmaceutical treatments, are not widely represented in the scientific literature.

3 Data and application

Clinical data today is an expensive asset and benchmark datasets, suitable for development and evaluation of a therapy recommendation CDSS, are unfortunately hardly accessible due to data protection issues. Moreover, feedback on interventions from longitudinal observations is difficult to obtain and often associated with long time constants. This shortage can be considered as one major reason for the small number of comparable works in the literature. Moreover, clinical data is rarely recorded in a structured and processable format but requires extensive preprocessing and transformation which is subject to uncertainties and noise.

In our previous work (Gräßer et al. 2019), we presented such a dataset which will also be used within this work. The data represents the routine health care of patients suffering from different types of the chronic inflammatory skin disease *Psoriasis* and is provided by an university outpatient clinic in Germany. Psoriasis is incurable

Table 1 Related works regarding HRSs ordered by year of publication, analyzed in terms of underlying algorithms, category of application, and data source

Reference	Year	Algorithm	Category	Data
Miyo et al. (2007)	2007	Memory-based	ADE prediction	Treatment history
Duan et al. (2008)	2008	Association analysis	Treatment	Treatment history
Hassan and Syed (2010)	2010	Memory-based	Disease prediction	Patient data
Folino and Pizzuti (2010)	2010	Association analysis	Disease prediction	Clinical history
Davis et al. (2010)	2010	Memory-based	Disease prediction	Clinical history
Duan et al. (2011)	2011	Association analysis	Treatment	Treatment history
Lu et al. (2012)	2012	Memory-based	Treatment	Patient data
Komkhao et al. (2012)	2012	Memory-based	Treatment	Patient data
Sezgin and Özkan (2013)	2013	–	Review	-
Sodsee and Komkhao (2013)	2013	Memory-based, content-based, knowledge-based	Treatment	Patient data
Chawla and Davis (2013)	2013	Memory-based	Disease prediction	Clinical history
Chen and Altman (2013)	2013	Association analysis	Treatment	Order history
Wiesner and Pfeifer (2014)	2014	Content-based	Information	Patient data
Folino and Pizzuti (2015)	2015	Association analysis, HMM	Disease prediction	Clinical history
Chen and Altman (2014)	2016	Association analysis	Treatment	Order history
Zhang et al. (2016b)	2016	Memory-based, ANN	Treatment	Patient data
Zhang et al. (2016a)	2016	Memory-based	ADE prediction	ADE data
Valdez et al. (2016)	2016	–	Framework	-
Hao and Blair (2016)	2016	Memory-based	Disease prediction	Patient data
Schäfer et al. (2017)	2017	–	Framework	-
Gräßer et al. (2017)	2017	Memory-based	Treatment	Patient data
Chiang et al. (2018)	2018	model-based, LR	ADE prediction	ADE data
Hors-Fraile et al. (2018)	2018	–	Review	-
Lattar et al. (2020)	2019	–	Review	–
Mustaqeem et al. (2020)	2020	Memory-based	Treatment	Patient data

Whereas *patient data* includes (condition related) attributes such as demographic information, diagnoses, or laboratory results, *treatment history* comprises previous treatments, *order history* previous clinical orders in general, and *clinical history* previous diagnoses only. *ADE data* summarizes datasets containing experience with drug–drug interactions. The 25 identified works stem from 20 research groups. Only publications which focus on clinical applications and utilize data related to health records do derive recommendations are included. Works dealing with recommending hospitals, doctors or social networks, but also nutrition or lifestyle change and behavior recommendations, are neglected and requires lifelong treatment and rehabilitation and hence patients often have a long treatment history including various pharmaceutical treatment options. The given data comprises $N = 1242$ consultations from $P = 239$ patients. Eight hundred and fifty-two consultations from 209 patients are associated with one of the $M = 22$ systemic treatment options and outcome information is given. These 852 consultations can be utilized for evaluation.

Each consultation is represented by patient data (i.e., demographic data, diagnosis, clinical findings, comorbidities, and life situation), treatment history (i.e., outcome of previously applied treatments), and outcome of the treatment applied in the current consultation. As it is prevalent with clinical data, the patient data is characterized by missing values. Since several algorithms used in the following are dependent on complete data, imputation strategies, depending on the mechanism underlying the missingness, were applied to complete this data. Outcome for each applied drug is measured by *effectiveness*, ΔPASI and occurrence of adverse drug events (ADEs). Whereas *effectiveness* is the physician's assessment on a scale *good*, *moderate*, and *bad*, ΔPASI quantifies the effect of an applied treatment in terms of the psoriasis area and severity index (PASI), which combines both the skin area affected and the severity of lesions (Fredriksson and Pettersson 1978). All three outcome indicators are summarized in the *affinity* score as demonstrated in (Gräßer et al. 2017) which is defined on the interval between 0 (bad outcome) and 1 (good outcome). To compute the *affinity* score, *effectiveness* and relative change of PASI, both transferred to the interval [0, 1], are averaged and penalized if ADEs have been occurred after application of the respective treatment.

Patient data of the overall $N = 1242$ available consultations is stored in the $D = 159$ dimensional *consultation representation matrix* $\tilde{\mathbf{X}}$. Moreover, the sparse $N \times M$ *historic consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{hist}$ represents the *affinity* scores of all treatments applied to a patient p previously to his or her nth consultation. The sparse $N \times M$ *complete consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{all}$ holds *affinity* scores for all treatments applied up to and including the current consultation n. Thus, a vector $\tilde{\mathbf{a}}^{all}$ associated with the nth consultation of patient p corresponds to the vector $\tilde{\mathbf{a}}^{hist}$ associated with the consultation succeeding consultation n (consultation $n + 1$) of this patient p. Finally, the sparse $N \times M$ *outcome matrix* $\tilde{\mathbf{Y}}$ holds the *affinity* scores of treatments applied in consultation n.

4 Collaborative filtering for therapy recommendation

4.1 Background

Particularly in e-commerce applications, CF methods have gained increasing impact and are an active topic of research. Personalized product recommendations are typically based on estimating a user's preference in order to derive a ranked list of items. As already introduced in Sect. 1 and demonstrated in previous works (Gräßer et al. 2017), personalized therapy recommendations can be regarded as a comparable task considering patients as target users and the therapy options as items. However, with essential differences as specified in Table 2.

In this work and our previous publication (Gräßer et al. 2017), we employ the concept of user-based CF in the therapy recommendation setting. The proposed therapy recommendation approach focuses on recommending treatment options which optimize outcome for a given patient and time. To meet the multifactorial outcome aspects described in Sect. 3, the proposed algorithms are optimized and evaluated with respect to the summarizing *affinity* score. Consultations are regarded as users and therapy options m as items. The intention is to exploit consultation similarity, i.e., similarity between patients at a point in time. We compare variations of two basic neighborhood-based, i.e., memory-based, methods differing in the data used to represent a consultation. Thus, patterns in response to previous treatments alone or supplemented by patient characteristics are supposed to be revealed.

All approaches have in common to (1) predict outcome of therapy options and (2) rank the treatments according to this prediction. The intention is not to recommend treatments based on general popularity or average efficiency, but rather to make a selection that is tailored to a target patient and consultation at hand. Furthermore, to leverage trust and reduce risk of the automatically generated therapy recommendations, (3) exclusion rules, such as contraindications and recommendations regarding the sequence of treatments can be applied in a post-filtering layer to highlight or exclude treatments from the recommendations list. The evaluation of such heuristics, however, is not part of this work. Figure 2 shows the processing and evaluation chain for a recommendation query together with all inputs and associated outputs.

In the reporting of our methodological specifications and results, we are guided by the guidelines for Machine Learning in Biomedical Research of Luo et al. (2016). However, not all suggested reporting items are applicable to the recommendation setting. Following the guideline's categorizations, the present outcome prediction problem can be considered as a prognostic regression task as we predict a temporal event. The study itself is retrospective, as we use data collected previously to our experiments.

4.2 Evaluation strategy

Within this work, two evaluation criteria are utilized. Accuracy of the predicted outcome is evaluated by computing the root-mean-square error (RMSE) between predicted and actually observed outcome (*affinity* score). The quality of the ranked list of recommendations is assessed by computing the agreement between recommendations derived from outcome predictions and actually applied therapies, i.e., the recommendations from the attending physician. The top-3 recommendations are assessed in the following using mean average precision (MAP) at position 3. However, as the objective is rather to recommend potentially successful therapy options than imitating the attending physician, the reference standard for the MAP@3 are recommendations only, which have actually been applied and for which good outcome was observed, i.e., *affinity* scores exceed a predefined threshold $thr_{good} = 0.5$.

Table 2 Key aspects that distinguish conventional RS applications from therapy recommendation applications

Aspect	Description
Multifactorial outcome	CFs typically aim to optimize overall user satisfaction by recommending items with the highest predicted purchase probability or the highest predicted rating. In contrast to such uni-dimensional criteria, in case of therapy recommendations, the aspects for an optimal item span multiple dimensions such as effectiveness and the risk of ADEs of a drug, but also way and practicability of application and cost–benefit balance. The priority of the individual aspects differs among patients.
Objective and subjective outcomes	CFs leverage implicit or explicit information to derive personal suggestions. Whereas implicit feedback, such as previous purchases, is generated automatically, explicit feedback, such as subjective ratings on a predefined scale, is provided actively by the user. Also in the therapy recommendation setting, the application of a treatment can be regarded as implicit feedback and aspects such as practicability can be measured by explicit and subjective patient feedback. Aspects such as effectiveness or ADEs, however, are not subjective ratings but rather objective measures which quantify treatment response. Prerequisite for a meaningful explicit feedback is in any case the patient's adherence to the recommended drug.
Shared decision-making	In contrast to traditional CF applications, in the therapy recommender system setting at least two stakeholders should be involved into the final choice of the item. The treatment decision is made by the attending physician. However, to increase the patient's satisfaction and adherence to the recommendation, the priorities are ideally defined in agreement with the patient's values and preferences. Moreover, additional "indirect" stakeholders are involved such as health insurance companies, which strive for reducing the overall treatment costs.
Patient information	Conventional CF applications base recommendations on purchase history only. Especially in health applications, patient characteristics, such as comorbidities and state of health, can be assumed to contain essential additional information for accurate outcome prediction and meaningful therapy recommendation.

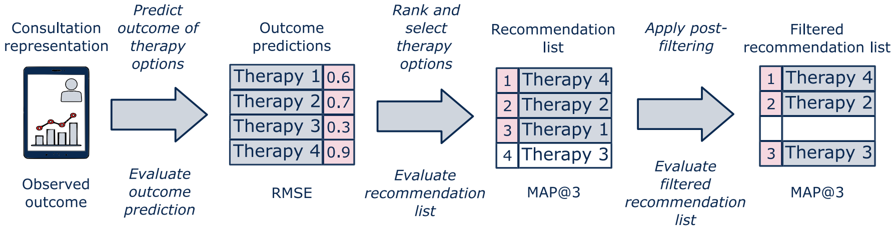


Fig. 2 Therapy recommendation processing and evaluation pipeline

As an accurate outcome prediction is the foundation for appropriate therapy recommendation, primary focus is put on this criterion in the following.

As the temporal consultations of the individual patients cannot be regarded to be independent and identically distributed, we apply a patient-wise evaluation scheme in this work. To make most of the available data and ideally provide an unbiased estimate of the true generalization error, a $P \times 5$ *nested cross-validation* approach is applied for model selection and generalization performance estimation, which was found to provide almost unbiased performance estimates (Raschka 2018). The realized approach is a nesting of two patient-wise cross-validation loops as pictured in Fig. 3 exemplarily for the *consultation representation matrix* $\tilde{\mathbf{X}}$.

The outer loop implements a leave-one-patient-out cross-validation, which in each iteration $p \in P$ holds out the consultations of the test patient p for evaluation. For this test patient p, an individual model on the basis of all patients apart from p is evaluated. For each consultation of the hold out test patient, accuracy of the predicted outcome (RMSE) and quality of the ranked list of recommendations (MAP@3) are assessed. The average RMSE and MAP@3 scores reflect the overall performance of this model applied to the test patient p's consultations. Finally, average and variance of RMSE and MAP@3 is computed over all iterations p to estimate the overall generalization performance.

The inner loop applies shuffled fivefold cross-validation for model selection on the basis of all consultations apart from test patient p. To avoid bias due to potential sample dependencies as described above, also the inner loop is implemented such that in no iteration i the same patient enters different folds in the same iteration. The data partitioning is carried out in such a way that each fold approximately contains the equal number of consultations. Within this inner loop, the fivefold cross-validation performance is calculated for all considered model variants (grid search) and the best performing model is selected.

4.3 Implementation of outcome prediction

In the user-based CF approach, as it is applied in this work, outcome prediction can be regarded as a neighborhood-based regression problem. Outcome estimates $\hat{y}_m^n$ for individual therapy options m are computed as a linear combination of observed *affinity* scores in the neighborhood of a test consultation n. This neighborhood is derived from the training subset which is defined by the cross-validation iterations p and i. Each training consultation k is represented by a respective vector $\tilde{\mathbf{a}}^k$ from $\tilde{\mathbf{A}}^{all}$ and holds the outcome of previous and current therapies as described in Sect. 3. The $\tilde{\mathbf{a}}^k$ for a cross-validation iteration are aggregated in a matrix $\tilde{\mathbf{A}}_{train}$. Figure 4 shows the outcome prediction for a treatment option m_1 and an exemplary test consultation n.

The neighborhood of size K is determined using heuristic similarity measures $s^{n,k}$ for each test consultation n. The similarity measures $s^{n,k}$ are further employed as the $k \in K$ regression coefficients to estimate $\hat{y}_m^n$ by computing the (weighted) average of all observed outcomes for each m according to

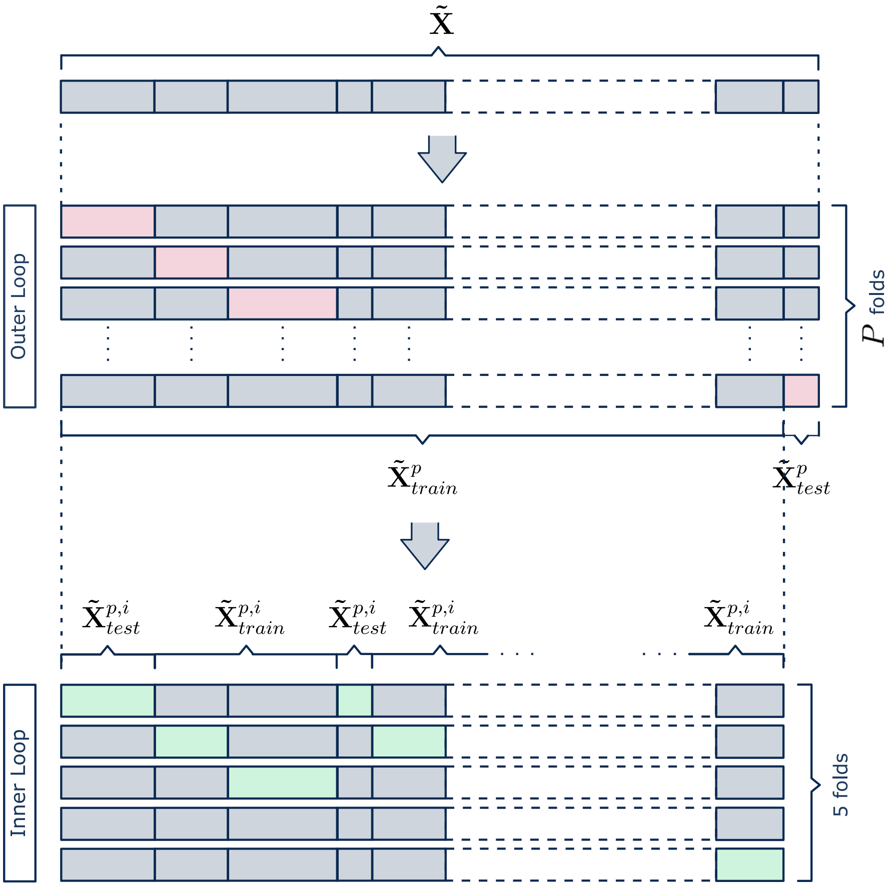

$\tilde{\mathbf{X}}$
Outer Loop
P folds
$\tilde{\mathbf{X}}^{p}_{train}$
$\tilde{\mathbf{X}}^{p}_{test}$
$\tilde{\mathbf{X}}^{p,i}_{test}$
$\tilde{\mathbf{X}}^{p,i}_{train}$
$\tilde{\mathbf{X}}^{p,i}_{test}$
$\tilde{\mathbf{X}}^{p,i}_{train}$
$\tilde{\mathbf{X}}^{p,i}_{train}$
Inner Loop
5 folds

Fig. 3 *Nested cross-validation* approach for model selection and evaluation. The outer loop implements a patient-wise cross-validation over all $p \in P$ patients, the inner loop implements a fivefold cross-validation, however, without mixing consultations of a patient p into test and training partition in any iteration i. Here, the example for the *consultation representation matrix* $\tilde{\mathbf{X}}$ is shown. $\tilde{\mathbf{A}}^{hist}$, $\tilde{\mathbf{A}}^{all}$, and $\tilde{\mathbf{Y}}$ are partitioned the identical way

$$\hat{y}^{n}_{m} = \frac{\sum_{k=1}^{K} \tilde{a}^{k}_{m} \cdot s^{n,k}}{\sum_{k=1}^{K} |s^{n,k}|} \tag{1}$$

Here, it must be kept in mind that outcome estimates can be computed for therapies only which appear at least once in the neighborhood of n. That means, besides predicting outcome the algorithm already selects a subset of therapies from all available options.

In a subsequent recommendation step, all treatment options for which an *affinity* prediction is available are ranked according to that prediction. The top-N ranked entries are evaluated to assess recommendation quality. If ties occur, i.e., the *affinity*

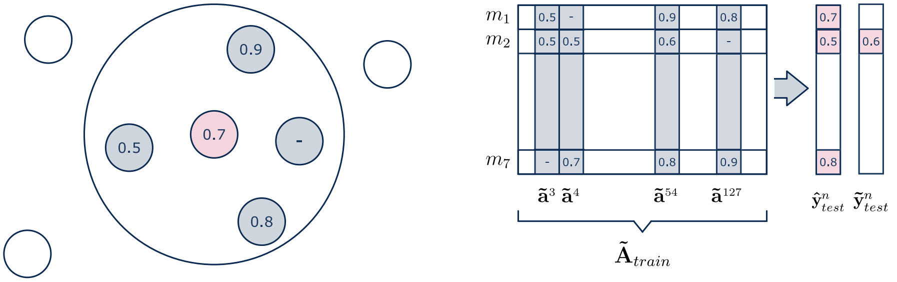

0.9
0.5
0.7
-
0.8
m_1
m_2
m_7
0.5 - 0.9 0.8
0.5 0.5 0.6 -
- 0.7 0.8 0.9
$\tilde{\mathbf{a}}^3$ $\tilde{\mathbf{a}}^4$ $\tilde{\mathbf{a}}^{54}$ $\tilde{\mathbf{a}}^{127}$
$\tilde{\mathbf{A}}_{train}$
0.7
0.5
0.8
0.6
$\hat{\mathbf{y}}^n_{test}$ $\tilde{\mathbf{y}}^n_{test}$

Fig. 4 Outcome $\hat{y}^n_m$ of treatment options m_1 is estimated for a test consultation n by aggregating all outcomes observed for m_1 in the treatment history of the K most similar training data consultations. Therefore, the (weighted) average of all outcomes for m_1 observed in that neighborhood is computed

score prediction of two therapy options equal, they are broken by recommending the more effective treatment according to the training partition.

To evaluate the accuracy of the predicted outcome, RMSE between predicted and actually observed outcome is computed as described in Sect. 4.2. For each test consultation n, only one ground truth value, i.e., applied therapy and known outcome is available in $\tilde{\mathbf{y}}^n_{test}$. Furthermore, prerequisite to compute a RMSE is that an *affinity* score estimate can be provided for this actually applied therapy. This in turn depends on whether the therapy is available in the neighborhood under consideration. Missing overlap of prediction and ground truth does not affect the RMSE calculation as the average score is only calculated using the existing values. However, reliability of RMSE suffers if computed from little overlapping observations. Beyond that this overlap directly affects the MAP@3, which quantifies the quality of the ranked list of recommendations. On the one hand, one can assume that a neighborhood of similar consultations is not only characterized by similar outcome but is also characterized by commonly applied therapies yielding good MAP@3 scores even when recommending only few options. On the other hand, for small neighborhood sizes K, the *coverage* of available treatment options can become very low, which reduces the possibility of recommendations overlapping with the actually applied treatment. Therefore, ratio of neighbors from which RMSE can be computed (*overlap*) and ratio of overall recommended treatment options (*coverage*) are also monitored. When defining the neighborhood sizes K, a trade-off needs to be found, as large K increase *overlap* at the expense of deteriorating prediction accuracy and recommendation quality due to inclusion of inappropriate consultations. Based on those considerations and with respect to the overall objective to optimize outcome prediction accuracy, two criteria are defined to be met for a model to be selected in the inner cross-validation loop: (1) the average number of recommendations overlapping with the actually applied treatment is *overlap* $\geq 75\,\%$ and (2) prediction accuracy (RMSE) is minimal.

The data to represent consultations and the applied similarity measure $s^{n,k}$ to compare consultation representations, have crucial impact on the prediction results. In the following, six CF variations are compared which differ in the data and similarity measure utilized. Additionally, prediction accuracy and therapy ranking performance are benchmarked against two baseline approaches. All implemented algorithms are introduced briefly below and their characteristic features summarized in Table 3. Detailed descriptions are given in Appendix A.

CF (Cosine), CF (Euclidean): Firstly, a *conventional* user-based CF approach, described in Appendix A.1, is implemented (Gräßer et al. 2017). Consultations are compared based on the outcome of commonly applied therapies. Consultations are regarded as similar if outcome on commonly applied therapies is similar according to the applied similarity measure. The experience with therapies observed in the neighborhood of a target consultation can then be transferred to this consultation. Two metrics to measure similarity are contrasted, *Cosine similarity* (CF (Cosine)) and *Euclidean distance* (CF (Euclidean)).

DR (Gower), (DR (Euclidean): The proposed conventional CF approach requires the associated test patient to have experience with at least one therapy in its therapy history (*cold start* problem). Moreover, reliability of the computed similarity can depend on the number of co-occurring therapies in consultation representation vector which can affect the accuracy of recommendations. To overcome such limitations and to make use of the additional, presumably meaningful information in the patient data, the described conventional CF is extended to a hybrid approach which additionally incorporates the available patient data into the similarity computation (Gräßer et al. 2017). Firstly, the *Gower similarity coefficient* is applied to compare consultation representations (DR (Gower)). It is inherently capable of measuring similarity at the presence of mixed data types and can cope with missing values. Secondly, *Euclidean distance*, converted to a similarity measure by means of a Gaussian radial basis function (RBF), is applied (DR (Euclidean)). In contrast to the *Gower similarity* approach, this similarity measure requires data transformation and normalization.

DR-RBA: Individual attributes typically are of varying importance concerning the similarity coefficient $s^{n,k}$. The *curse of dimensionality* further requires the dimension of the data to be as low as possible to facilitate a meaningful concept of similarity. As a consequence, both, the unweighted inclusion of attributes and the inclusion of irrelevant or redundant attributes, can affect the performance of neighborhood-based CF algorithms substantially. Accordingly, it is an obvious strategy to modify the above-proposed patient-data CF approach in order to weight the individual attributes according to their relevance (attribute weighting) and to remove irrelevant ones (attribute selection) before computing similarity. Here, a relief-based algorithm (RBA) is adapted to the given problem as detailed in Appendix A.3. The proposed algorithm weights and selects attributes on the basis of *a priori* assumptions regarding similarity an dissimilarity of instances. Finally, *Gower similarity coefficient*, which allows to assign weights w_d or discard attributes, is applied to compare consultation representations in the scaled and reduced attribute space.

DR-LMNN: Especially linear transformation, which takes correlations among attributes and the data's distribution in the attributes space into account, is a widely and successfully used preprocessing strategy in the context of classification and data analysis. In contrast to the RBA approach, such transformations not only scale the dimensions of the attribute space but rotate the basis of the coordinate system in order to adapt to the data at hand. This bears the potential to yield more meaningful

Table 3 Summary and qualitative comparison of the CF algorithms and variants applied in this work

Method	Advantages	Disadvantages
CF (Cosine)	–Only treatment history required	–New patient *Cold start* problem
CF (Euclidean)		–Dependent on co-occurring therapies
DR (Gower)	–No new patient *cold start* problem	–High-dimensional attribute space
DR (Euclidean)	–Additional patient information included	–Importance of attributes disregarded
	–Attribute data-type considered in *DR (Gower)*	
	–Capable of handling missing values	
DR-RBA	–see *DR (Gower)*	–Correlations and redundancies not respected
	–Attributes weighted according to importance	–Dependent on sufficient and informative data
	–Dimensionality reduction	
	–Physical meaning of attributes is maintained	
DR-LMNN	–see *DR (Euclidean)*	–Attribute data-type not considered
	–Multivariate distribution of the data is accounted for	–Dependent on sufficient and informative data
		–No dimensionality reduction
		–Physical meaning of attributes get lost

neighborhoods. Also in the context of the proposed patient-data CF, it is assumed that the multivariate distribution of the data has crucial impact on the similarity computation and hence the outcome estimation of the regression algorithm. Furthermore, it is assumed that certain attributes are redundant or correlate strongly. Hence, in order to improve outcome prediction accuracy, linear transformation of the data before computing similarity may be a beneficial preprocessing approach. Here, a generalized *Mahalanobis metrics* is learned from the data based on *a priori* information which can be regarded as a linear transformation of the attribute space before applying *Euclidean distance*. As detailed in Appendix A.4, the large Margin nearest neighbor (LMNN) algorithm proposed by Weinberger et al. (2005) and tailored to the problem at hand, is employed for *Mahalanobis metrics* learning.

Baseline approaches: Additionally, two baseline approaches are compared with the proposed CF algorithms. Firstly, *average efficiency*, i.e., the average *affinity* scores for each treatment, is computed as outcome prediction baseline (Base-EFF). Ranking those predictions according to outcome provides one recommendation baseline. Secondly, *overall popularity*, i.e., the individual therapies' frequency of application in the training partitions, are employed as second recommendation baseline (Base-POP). As no outcome prediction is provided, no RMSE can be computed for the *overall popularity* baseline.

5 Results

In the following, the performance of the introduced CF algorithms and variations are compared. In subsection 5.1, results from model selection, i.e., the inner cross-validation loop, are contrasted, and the best model for each approach is selected. In subsection 5.2, generalization performance estimates, yielded in the outer cross-validation loop, are compared and discussed.

5.1 Model selection

Depending on algorithm, various free parameters need to be optimized. The most crucial parameter which all approaches have in common is the neighborhood size K. As specified in Sect. 4.3, the primary evaluation criterion for model selection is the accuracy of outcome predictions. However, as additional criterion, the ratio of neighbors overlapping the actually applied therapy is defined to exceed *overlap* ≥ 0.75 to base the selection on reliable values. Table 4 summarizes mean values and standard deviations of the inner cross validation results (i.e., average over all 5 folds) for each of the discussed scores and the selected K. Further parameter settings are discussed in the following.

In case of the conventional CF (CF (Cosine) and CF (Euclidean)), optimal K of the *Cosine similarity* approach is considerably smaller than K of the *Euclidean distance*. Nevertheless, outcome prediction performance (RMSE), which in both cases deteriorates with increasing neighborhood size, shows superior results when using the *Euclidean distance* compared with *Cosine similarity*. Regarding the ability to rank the actually applied and successful therapy among the top options, *Cosine similarity* outperforms the *Euclidean distance*. *Cosine similarity* is capable of retrieving already at very small neighborhood sizes a large ratio of neighbors overlapping the actually applied therapy (*overlap*). Simultaneously, *coverage* is comparably low, meaning that the retrieved neighboring consultations are very accurate with respect to the applied treatments and hence introduce only little noise into the recommendation. Both results in high MAP@3 values. Yet, the neighboring consultations only allow for comparably bad outcome prediction. The *Euclidean distance*, on the other hand, facilitate much better RMSE values which, however, is based on smaller *overlap*. The comparably large *coverage* yields inaccurate recommendations and low-quality therapy ranking.

In case of the patient-data CF (DR (Gower) and DR (Euclidean)), outcome prediction performance (RMSE) is comparable for both measures compute similarity between consultation representations. However, K is distinctly smaller for the *Gower similarity*. With this K, the *Gower similarity* approach is capable of yielding larger agreement with the physician's successful recommendations (MAP@3) with smaller *coverage*. When considering the course of MAP@3 over K, MAP@3 is even larger for smaller K than the point where the *overlap* criterion is met. These observations indicate that considering scale of measurement, i.e., data type, is obviously beneficial when comparing attributes. Yet, for both similarity measures the patient-data CF only allows for bad outcome predictions compared with the conventional CF.

As introduced in Sect. 4.3, the proposed RBA approach scales each attribute d according to assigned importance weights w_d before computing *Gower similarity*, whereas only those attributes assigned with positive weights are taken into account. The free parameters, number of nearest hits and nearest misses K_{RBA} and neighborhood size K, are determined by means of a grid search within the inner cross-validation loop. Concerning K_{RBA}, the best RMSE could be constantly found for $K_{RBA} = 15$. As given in Table 4, by applying this attribute weighting approach, the prediction error is reduced compared to the unweighted *Gower similarity* approach

Table 4 Inner cross-validation loop results

Method	K	RMSE	MAP@3	Coverage	Overlap
CF (Cosine)	13.13 (5.01)	0.18 (0.00)	0.77 (0.01)	0.31 (0.03)	0.90 (0.01)
CF (Euclidean)	62.32 (6.23)	0.14 (0.00)	0.40 (0.01)	0.64 (0.02)	0.77 (0.02)
DR (Gower)	14.11 (18.20)	0.21 (0.00)	0.46 (0.09)	0.32 (0.10)	0.80 (0.07)
DR-RBA	12.22 (7.85)	0.17 (0.00)	0.55 (0.05)	0.36 (0.08)	0.85 (0.05)
DR (Euclidean)	39.78 (12.02)	0.22 (0.00)	0.34 (0.06)	0.46 (0.05)	0.94 (0.03)
DR-LMNN	9.68 (3.37)	0.14 (0.00)	0.66 (0.02)	0.29 (0.03)	0.85 (0.03)
Base-EFF	–	0.29 (0.00)	0.13 (0.02)	0.97 (0.00)	0.99 (0.00)
Base-POP	–	–	0.11 (0.01)	0.97 (0.00)	0.99 (0.00)

Mean and standard deviation of neighborhood size K, outcome prediction accuracy (RMSE), recommendation list agreement (MAP@3), average *overlap* with applied treatment and *coverage* of treatment options and also MAP@3 is overall increased. DR-RBA outperforms the unweighted version for an even smaller neighborhood size. Nevertheless, *coverage* is generally larger and the recommender hence tends to be less selective.

Two free parameters, additional to the CF neighborhood size K, must be defined for the metric learning method: the LMNN neighborhood size K_{LMNN}, which determines the included *target neighbors* and *impostors*, μ, which controls the impact of the competing objectives ϵ_{pull} and ϵ_{push}, and learning rate ν. Best results could be found for $K_{LMNN} = 10$, $\nu = 0.5$, and $\nu = 0.001$ for the entire range of evaluated K. Furthermore, as the yielded distances are distributed over a wider range after data transformation, the RBF spread parameter is adjusted to $\sigma = 0.5$. The DR-LMNN *Overlap* is rather large already for small K which results in a very small neighborhood size K. This large ratio of overlapping treatments which coincides with very small RMSEs values is a clear indicator for a meaningful neighborhood. Moreover, also MAP@3 is increased and *coverage* reduced compared to the basic *Euclidean distance* patient-data CF especially for rising K.

According to the inner cross-validation loop, all neighborhood-based CF approaches are clearly capable of outperforming the two baselines. In terms of RMSE, average efficiency as outcome predictions (Base-EFF) is still inferior to all other methods. Nevertheless, ranking treatment according to this predictions is still superior to only ranking treatments according to overall popularity (Base-POP). Not all treatment options are present in all inner cross-validation folds, resulting in *coverage* and *overlap* below 100%.

5.2 Generalization performance evaluation

When considering the outer cross-validation results summarized in Table 5 and visualized in Fig. 6a and b, the large variance of the results becomes apparent. Within each outer cross-validation loop, all consultations except the test patient p are available. Hence, the applied leave-one-patient-out cross-validation approach is assumed to be almost unbiased. The major downside of many small folds is the large variance of the individual estimates as it is observed. In each iteration p, the performance estimate is based on the consultations of patient p only, which is highly variable. Especially variance of MAP@3 scores, pictured in Fig. 6b, is remarkably large and partly spread over the entire value range.

Statistical hypothesis tests are applied to evaluate the proposed algorithms performance differences with respect to their statistical significance. Both, central tendency of outcome prediction (RMSE) and of recommendation quality (MAP@3) are examined. Due to multiple algorithms to be compared, firstly an omnibus test under the null hypothesis is conducted and, in case of rejection of the null hypothesis, pairwise *post hoc* tests are performed. The null hypotheses are that the RMSE and MAP@3 results from each algorithm, including the baselines *average efficiency* and *overall popularity*, stem from the equal distribution. The pre-defined level of significance is $\alpha = 0.05$.

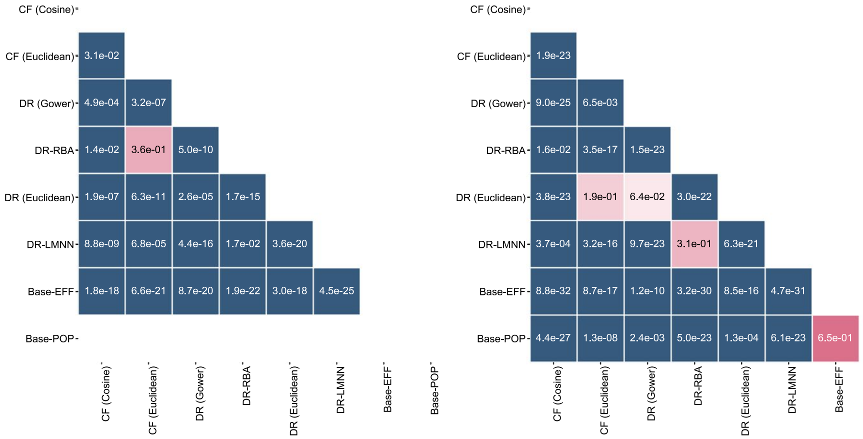
As the leave-one-patient-out cross-validation uses the identical patients and consultations for evaluation, the individual algorithms' results are considered to be paired. Both, RMSE and MAP@3 results are numerical values but cannot be considered to be normally distributed. As the majority of errors are small and the frequency decreases as the error value increases, the RMSE distribution is right-skewed. In case of the MAP@3 score, the MAP@3 distribution is left-skewed as the majority of observed scores are large or is bimodal. Consequently, a nonparametric, namely the *Friedman test* (Friedman 1937) is used in both cases although having less statistical power than parametric tests. The probability distribution of the Friedman test statistic is approximated by the Chi-squared distribution. Only the intersection of patients with available RMSE or MAP@3 score are used for the hypothesis testing in the following, encompassing $n = 193$ and $n = 201$ observations, respectively. As both, the number of algorithms to be compared ($k = 8$) and the number of included partitions ($n = 193$ and $n = 201$) are sufficiently large, this distribution assumption can be regarded to be valid and provide reliable *p*-values. Concerning both, RMSE and MAP@3 score, the Friedman test shows significant differences between the algorithms included.

For pairwise *post hoc* testing, we applied *Wilcoxon signed-rank tests* (Wilcoxon 1945) on all pairs of algorithms and both evaluation metrics. To account for multiple testing, the *Bonferoni–Holm*-correction is applied (Holm 1979). The individual test samples in each outer cross-validation iteration can be regarded identically distributed but cannot be considered independent due to overlapping data. As a consequence, the test results should be interpreted with caution.

Overall, the estimated generalization performance reproduces the inner cross-validation results in terms of both aspects, RMSE and MAP@3. Whereas central tendency—in most cases—are comparable, variance of the outer loop results is, as initially discussed, remarkably large, especially for MAP@3.

Looking at the results summarized in Table 5 and *p*-values in Fig. 5a and b, it becomes obvious that all examined algorithms perform significantly better than the two baseline methods in terms of both, outcome prediction and therapy ranking. Hence, it can be concluded that estimating outcome based on local data alone is highly beneficial.

In the case of the conventional CFs, the prediction performance of the *Euclidean distance* is significantly superior to the *Cosine similarity* and even outperforms all other approaches apart from the metric learning (DR-LMNN) and attribute scaling (DR-RBA) patient-data CFs. In terms of agreement of the recommendation list with the attending physician's successful choices, however, a statistically significant superiority of the *Cosine similarity* conventional CF algorithm over all other evaluated approaches is evident. As was already observed for the inner loop, prediction accuracy is improved at the expense of MAP@3 and *vice versa*. In terms of prediction accuracy (RMSE), both patient-data CFs are clearly inferior to the conventional CF and also MAP@3 is at the lower end of the results overall being achieved. Regarding MAP@3, the *Gower similarity* approach performs even worse in the outer than in the inner loop. Whereas outcome prediction obviously benefits from the data type sensitive similarity measure, no statistically significant performance difference can be shown between *Gower similarity* and the *Euclidean distance* in terms of therapy ranking quality.



	CF (Cosine)	CF (Euclidean)	DR (Gower)	DR-RBA	DR (Euclidean)	DR-LMNN	Base-EFF	Base-POP
CF (Cosine)								
CF (Euclidean)	3.1e-02							
DR (Gower)	4.9e-04	3.2e-07						
DR-RBA	1.4e-02	3.6e-01	5.0e-10					
DR (Euclidean)	1.9e-07	6.3e-11	2.6e-05	1.7e-15				
DR-LMNN	8.8e-09	6.8e-05	4.4e-16	1.7e-02	3.6e-20			
Base-EFF	1.8e-18	6.6e-21	8.7e-20	1.9e-22	3.0e-18	4.5e-25		
Base-POP								

	CF (Cosine)	CF (Euclidean)	DR (Gower)	DR-RBA	DR (Euclidean)	DR-LMNN	Base-EFF	Base-POP
CF (Cosine)								
CF (Euclidean)	1.9e-23							
DR (Gower)	9.0e-25	6.5e-03						
DR-RBA	1.6e-02	3.5e-17	1.5e-23					
DR (Euclidean)	3.8e-23	1.9e-01	6.4e-02	3.0e-22				
DR-LMNN	3.7e-04	3.2e-16	9.7e-23	3.1e-01	6.3e-21			
Base-EFF	8.8e-32	8.7e-17	1.2e-10	3.2e-30	8.5e-16	4.7e-31		
Base-POP	4.4e-27	1.3e-08	2.4e-03	5.0e-23	1.3e-04	6.1e-23	6.5e-01	

Fig. 5 Results regarding outcome prediction and list agreement of almost all algorithms are significantly different. p-values of pairwise *post hoc* tests (*Wilcoxon signed-rank tests*), comparing all presented algorithms concerning (**a**) outcome prediction (RMSE) and (**b**) recommendation list agreement (MAP@3). Statistical significant performance differences ($p > \alpha$) are colored blue

Table 5 Outer cross-validation loop results

Method	K	RMSE	MAP@3	Coverage	Overlap
CF (Cosine)	13.13 (5.01)	0.16 (0.11)	0.77 (0.30)	0.29 (0.16)	0.91 (0.19)
CF (Euclidean)	62.32 (6.23)	0.14 (0.11)	0.39 (0.33)	0.63 (0.11)	0.82 (0.28)
DR (Gower)	14.11 (18.20)	0.20 (0.11)	0.30 (0.33)	0.52 (0.14)	1.00 (0.00)
DR-RBA	12.22 (7.85)	0.14 (0.12)	0.70 (0.34)	0.30 (0.14)	0.99 (0.09)
DR (Euclidean)	39.78 (12.02)	0.22 (0.11)	0.34 (0.35)	0.51 (0.14)	1.00 (0.00)
DR-LMNN	9.68 (3.37)	0.11 (0.11)	0.67 (0.31)	0.34 (0.14)	0.98 (0.10)
Base-EFF	–	0.29 (0.13)	0.12 (0.16)	1.00 (0.00)	1.00 (0.00)
Base-POP	–	–	0.15 (0.32)	1.00 (0.00)	1.00 (0.00)

Mean and standard deviation of outcome prediction accuracy (RMSE), recommendation list agreement (MAP@3), average *overlap* with applied treatment and *coverage* of treatment options

Applying the proposed RBA algorithm significantly improves the *Gower similarity* baseline regarding both aspects, prediction accuracy and recommendation quality. This finding indicates that linear attribute scaling and the inherent attribute selection is a suitable approach. The RMSE improvement yielded by the LMNN approach is even larger compared to the *Euclidean distance* baseline. Rotation of the attribute space beyond attribute scaling is obviously additionally beneficial. The distance metric optimized to the given task achieves the best prediction accuracy compared to all demonstrated methods.

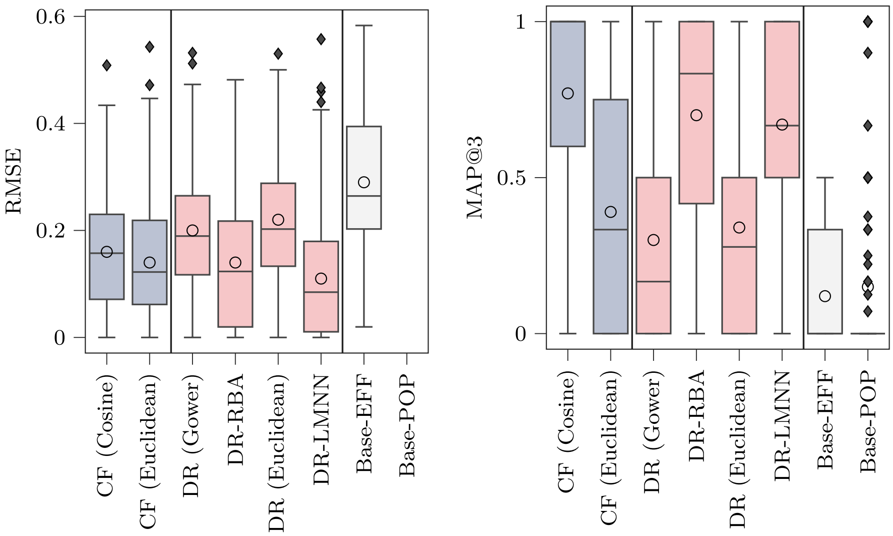


Fig. 6 Outer cross-validation loop results. (**a**) Outcome prediction accuracy (RMSE—lower is better) and (**b**) recommendation list agreement (MAP@3—larger is better) evaluated for all proposed methods

6 Therapy recommender system prototype

In order to clarify the presented algorithms' intended use and application, a prototypical recommender system including a GUI is developed and presented within this work. We have opted for a web application approach (client-server-model) and a browser-based GUI as it ensures independence and easy portability among platforms and devices. It further facilitates maintenance and problem resolution. As already introduced in Sect. 2, several CDSSs requirements, which determine acceptance and successful application are described in the literature (Kawamoto 2005; Beeler et al. 2014; Sim 2001; Kaplan and Frosch 2005). Moreover, according to a survey of explanations in recommender systems (Tintarev and Masthoff 2007), seven criteria to assess interpretability and explainability of recommendations can be defined. From those, especially *transparency* how recommendations are generated, *effectiveness* and *efficiency* of recommendations, and user's *trust* into the system and recommendations are of particular interest for the given therapy recommendation task. Additionally, communication of recommendation reliability and uncertainties, as mentioned in (Calero Valdez et al. 2016), can be regarded an important requirement. To summarize, during prototype and GUI design, we put focus on the following characteristics in order to meet these requirements:

- Provide actionable recommendations, rather than just assessments.
- Provide recommendations instantaneously at the time and location of decision-making.
- Provide recommendations tailored to the current patient characteristics and needs.

- Provide interpretable and explainable recommendations and communicate uncertainties.
- Facilitate the integration of patient preferences and values and shared decision-making.

The developed prototype implements two basic functionalities: on the one hand patient data input and presentation, and on the other hand treatment recommendation visualization (recommendation dashboard). The implementation of both functionalities, under consideration of the above characteristics, is described in detail in Appendix B along with screenshots.

7 Discussion

7.1 Insights and findings

Overall, it can be concluded that the neighborhood-based CF methods, which estimate outcome and rank treatment options based on local data, i.e., a virtual cohort of similar patients only, show great potential. Patient-specific recommendations can be facilitated to supplement a physician's experience and external evidence with practice-based evidence as proposed by Longhurst et al. (2014). Assuming an EHR with relevant condition related data to be a by-product of routine care and data to be provided in a structured format, such recommendations can be provided automatically as part of the clinician's workflow at the time and location of decision-making as demanded in (Kawamoto 2005). The proposed recommendation dashboard exploits the intuitive characteristic of the neighborhood-based CF. The visualization of outcome prognoses and statistics on all outcome aspects from the included local data can help to find and optimal treatment which is in accordance with the patient's preferences and needs. The patient can be incorporated into decision-making, which, in turn, can be expected to increase his or her satisfaction and adherence to the applied treatment. The presentation format can be regarded as actionable recommendations tailored to the current needs as requested in (Beeler et al. 2014) and also the growing demand for patient-engagement (Sim 2001; Kaplan and Frosch 2005) is accounted for. Moreover, the proposed visualization of the included data can provide additional insight into recommendations and recommendation reliability to remedy acceptance issues. Both are important features to push acceptance of CDSSs (Berner and La Lande 2016; Sim 2001) but are hardly addressed in the related works identified in Sect. 2. Particularly, the evidence-based post-filtering to highlight or exclude treatments from the recommendations list, but also making patient-specific information regarding external evidence available, leverages trust into such a CDSS and reduces the risk of automatically generated therapy recommendations.

The essential strength of the neighborhood-based CF methods is twofold. On the one hand, the modeling based on local data clearly increases accuracy when predicting outcome of the actually applied therapy. On the other hand, CF additionally features the selection of a subset of therapy options which benefits the recommendation quality, i.e., MAP@3. Only treatment options are included into the recommendation list which are observed in that neighborhood of the target patient. As a consequence, the proposed neighborhood-based CF methods are capable of predicting outcome of therapy options more accurately than average outcome (*average efficiency* baseline). Based on outcome predictions and selection of a subset of options, the potentially most successful therapy recommendations are derived which are independent from treatment popularity. As was shown, these recommendations clearly outperform the *overall popularity* baseline concerning agreement with the attending physicians successful choice.

In case of the conventional CF, the similarity measure must be chosen dependent on the main objective whether to improve outcome prediction accuracy or the agreement (RMSE) between recommendations and actually and successfully applied treatments (MAP@3). On the one hand, a large overlap of commonly applied treatments increases similarity in case of *Cosine similarity*. Therefore, this similar measure is more selective concerning treatments observed in the neighborhood which yields larger MAP@3 and lower *coverage*. The *Euclidean distance*, on the other hand, especially focuses on similar outcome when computing similarity which results in small RMSE scores. This metric is, however, not sensitive to the number of co-occurring treatments in two vectors to be compared. As introduced above, algorithm selection here is based on outcome prediction accuracy rather than the ranking of treatment options. Hence, the conventional CF using *Euclidean distance* can be considered as the overall preferable conventional CF algorithm.

Considering either of the evaluation criteria, the patient-data CF approaches are clearly inferior to the conventional approaches. There are two data properties that basically contribute to the observed performance difference. Firstly, the significantly larger attribute space (22 vs. 159) increases the *curse of dimensionality* effects. The computed similarity or distance measures, which are fundamental for selecting a patient's neighborhood, become imprecise and meaningless with increasing attribute space. Secondly, lacking relevance but also redundancy of attributes introduces significant noise into the similarity or distance computation and worsen its informative value. Attributes which are not relevant for the outcome prediction problem degrade accuracy. Hence, attribute selection and weighting is a crucial factor of the patient data approach. Therefore, results could be significantly improved by the proposed supervised attribute selection and scaling method (DR-RBA). As a by-product this approach additionally features insight into relevancy of attributes and lowers computational complexity and required storage. The underlying distance metric optimized by the LMNN algorithm is not reducing the attribute space and also the physical meaning of attributes get lost in the transformed space. However, in terms of outcome prediction accuracy, the DR-LMNN approach even yields larger improvements and is the overall best-performing algorithm. On the one hand, the shown RBA and LMNN results indicate that the applied optimization algorithms are suitable for the given task. On the other hand, also the assumptions regarding similarity, which provide the ground truth for this demonstrated supervised learning methods, prove to be valid. Overall, the hypothesis that the additional patient data contributes important information is proven. To conclude, the optimized DR approaches are the preferable algorithms. Firstly, they are capable of yielding superior prediction accuracy. Secondly, *cold start* problems can be overcome and reliability issues caused by sparse consultation representation can be reduced.

The demonstrated system targets the treatment of Psoriasis and optimizes therapy recommendations regarding a summarizing outcome objective (*affinity* score). However, this algorithm framework allows to optimize recommendations in terms of each treatment response individually or regarding other criteria and is transferable to other conditions as well. In general, optimizing and evaluating treatment decision support regarding outcome rather than agreement with expert recommendations or guidelines can be considered more reliable. Using outcome as ground truth can be regarded more objective in contrast to subjective and sometimes ambiguous physician decisions or inadequate guideline recommendations.

7.2 Limitations and future works

The general challenge with applying neighborhood-based CF methods to the therapy recommendation domain is the dependence on representative and reliable data. Such data must consist of structured patient representations as complete and as error-free as possible. In particular, the outcome criterion to be optimized must be reported objectively and free of gaps. Such data, however, is hardly generated in routine care today which limits the integration of such recommender systems into existing infrastructures. Data protection and usage regulations make implementation additionally difficult in practice. Another limitation is the integration of new treatment options which are underrepresented in the data (*cold start* problem) as they are less likely to appear in the neighborhood and hence recommendation list of a target patient.

The major challenge and limitation of the presented comparative study is the small data foundation on which it is based. Two factors determine the demand for a large data basis. On the one hand, a large variety of patients must be included in order to find a sufficiently homogeneous neighborhood for each target patient. On the other hand, sufficient representations of each relevant treatment option must be available within this homogeneous neighborhood to provide reliable outcome statistics. As was stated, benchmark datasets with suitable longitudinal data are not available, which emphasizes the uniqueness of this work. Based on larger datasets, also state of the art model-based CF algorithm, such as matrix factorization (MF) (Koren et al. 2009) or sparse linear method (SLIM) (Ning and Karypis 2011), can become alternative approaches, however, making visualization and explainability of recommendations difficult. Another critical issue is the aspect of only partially observed (*hidden*) ground truth (Mei et al. 2015), meaning that only one outcome per recommended and applied treatment option for each consultation is available. On the background of low inter-rater agreement, it is obvious that the given ground truth derived from the physicians' recommendations and consequently the MAP@3 scores lack reliability. But also RMSE ground truth derived from the observed outcome relies heavily on the patients' adherence to the recommended treatment. Both limitations can be countered by a larger dataset that covers a wide variety of patients and treatment options.

A key drawback of all proposed algorithms is their reduced capability to consider the temporal dependencies of consultations. The sequence of a patient's consultations can be considered as observations over a defined period of time resulting in time sequences of varying length. On the one hand, treatment recommendation considering these time dependencies can be formulated as a sequence classification or regression task. Here, one model for each treatment option is trained to predict an outcome which characterizes the entire (multivariate) input sequence. Exemplary algorithms capable of performing such tasks while considering time dependencies are, e.g., hidden Markov models (HMMs) (Rabiner 1989) but also recurrent neuronal networks (RNNs) such as long short-term memories (LSTMs) (Hochreiter and Schmidhuber 1997) or gated recurrent units (GRUs) (Cho et al. 2014). On the other hand, this consultation sequence can be considered as session-aware recommendation scenario (Symeonidis et al. 2020) which exploits past sessions of registered users. Various approaches have been presented in the literature which also apply neuronal networks such as GRUs (Hidasi and Karatzoglou 2018) or combinations of convolutional neuronal networks (CNNs) and LSTMs (Moreira et al. 2019). However, besides interpretability issues, the required data volume hampers the usage of such methods within the scope of this work.

Particularly notable are the very small included neighborhoods in case of the best performing approach DR-RBA and DR-LMNN. In this context, future works will focus on further visualization concepts such as means to inspect the neighborhood of a target patient directly. Moreover, the integration of further information sources, such as information from advisory platforms or patient reviews from online pharmacies or drug rating portals will be considered in future works. Finally, the proposed system's benefits, applicability, acceptance and usability—with special focus on explanation—must be evaluated in a clinical study.

Future work will also address the generalizability and transfer of the proposed methods to other applications. As with the presented application psoriasis, the challenge will be the definition of patient representations, i.e., relevant attributes, and the identification of appropriate condition specific outcome criteria. Both are highly application dependent but crucial and must be done in close alignment with the various stakeholders.

8 Conclusion

Within this work, the application and adaption of neighborhood-based CF methods for therapy recommendation was demonstrated. Moreover, a GUI concept that, based on the CF algorithm, intends to present recommendations in an intuitive and interpretable format was introduced. Beyond visualizing recommendations, prognosis for several outcome aspects and information from external evidence tailored to a target patient are given. Regarding the underlying CF algorithms, in particular the incorporation of patient-data yields small outcome prediction errors and recommendation lists which overlap with the actually applied and successful therapy to a large extend. A prerequisite is attribute weighting or transformation of the attribute space before computing similarity among patients and consultations. Two supervised methods are proposed and successfully applied, namely a RBA and LMNN metric learning. Even though evaluated on a small data basis, we consider this work to be an important contribution to the HRS domain and motivation to further research.

Appendix: Algorithmic details

Conventional collaborative filter (CF)

This approach is comparable to recommending items based on users' rating behavior on previously purchased products. The underlying assumption of this approach is that therapies applied to a given patient within his or her treatment history and the associated outcomes reincorporate meaningful information about that respective patient and consultation.

Here, the similarity measure $s^{n,k}$ is defined by a function $s(\tilde{\mathbf{a}}^{n}_{test}, \tilde{\mathbf{a}}^{k})$ which calculates a pairwise similarity between the test consultation representation $\tilde{\mathbf{a}}^{n}_{test}$, derived from the *historic consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{hist}$, and all training consultation representations $\tilde{\mathbf{a}}^{k}$, derived from the training partition of the *complete consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{all}$ and which are stored in matrices $\tilde{\mathbf{A}}_{train}$ and $\tilde{\mathbf{A}}_{test}$, respectively. Figure 7 visualizes the neighborhood of an exemplary test consultation representation $\tilde{\mathbf{a}}^{n}_{test}$. All attributes in $\tilde{\mathbf{A}}_{train}$ and $\tilde{\mathbf{A}}_{test}$ have equal quantitative data type and are measured with equal scale. Hence, no normalization of the data is required to equal the impact of the individual dimensions. There are numerous functions for computing pairwise similarity $s^{n,k}$ between two consultation representations n and k with such properties. In the context of CF, especially *Cosine similarity* is widely used, however, also *Minkowski metrics*, such as the *Euclidean distance*, are appropriate for similarity computation. As the proposed CF algorithm is based on similarity measures $s^{n,k}$, *Euclidean distance* metric need to be converted to similarity measures. Here, a RBF with spread parameter $\sigma = 0.25$ is employed for that purpose.

The proposed conventional CF approach requires the associated test patient to have experience with at least one therapy in its therapy history (*cold start* problem). Moreover, reliability of the computed similarity can depend on the number of co-occurring therapies in consultation representation vector which can affect the accuracy of recommendations.

Patient-data collaborative filter (DR)

The described conventional CF is extended to a hybrid approach which additionally incorporates the available patient data into the similarity computation (Gräßer et al. 2017). The advantage is twofold. Firstly, the *cold start* problem,

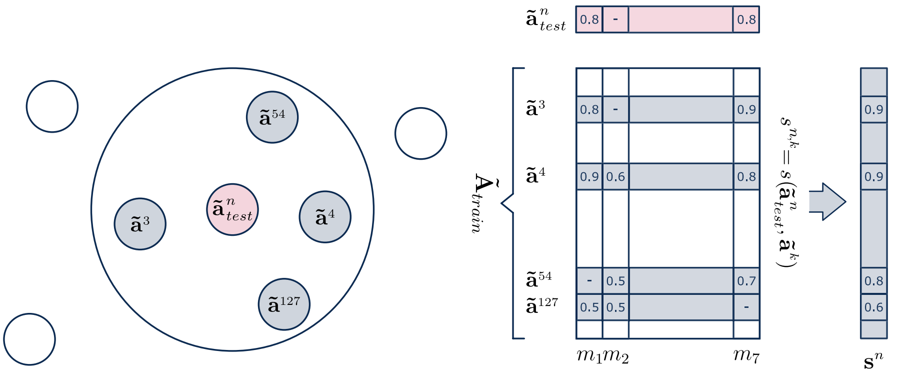

$\tilde{\mathbf{a}}^{n}_{test}$
0.8
-
0.8
$\tilde{\mathbf{a}}^{54}$
$\tilde{\mathbf{a}}^{3}$
$\tilde{\mathbf{a}}^{n}_{test}$
$\tilde{\mathbf{a}}^{4}$
$\tilde{\mathbf{a}}^{127}$
$\tilde{\mathbf{A}}_{train}$
$\tilde{\mathbf{a}}^{3}$ 0.8 - 0.9
$\tilde{\mathbf{a}}^{4}$ 0.9 0.6 0.8
$\tilde{\mathbf{a}}^{54}$ - 0.5 0.7
$\tilde{\mathbf{a}}^{127}$ 0.5 0.5 -
$m_1 m_2$
m_7
$s^{n,k} = s(\tilde{\mathbf{a}}^{n}_{test}, \tilde{\mathbf{a}}^{k})$
0.9
0.9
0.8
0.6
$\mathbf{s}^{n}$

Fig. 7 In the conventional CF approach, consultations are compared regarding treatment history stored in $\tilde{\mathbf{a}}^k$ and $\tilde{\mathbf{a}}^n_{test}$, respectively

which encounters for patients having no experience with previous treatments, and reliability issues of to view co-occurring items to be compared can be met. Secondly, the additional, presumably meaningful information in the patient data is made use of.

In the hybrid *patient-data* CF approach described in the following, consultations n and k are represented by vectors $\tilde{\mathbf{x}}^n_{test}$ and $\tilde{\mathbf{x}}^k$ which are derived from the *consultations data matrix* $\tilde{\mathbf{X}}$ and stored in matrices $\tilde{\mathbf{X}}_{train}$ and $\tilde{\mathbf{X}}_{test}$, respectively. $\tilde{\mathbf{X}}$ combines both, patient data and outcome of previously applied therapies. Hence, the heuristic similarity measure $s^{n,k}$ which determines the included neighborhood and the regression coefficients is defined by the function $s(\tilde{\mathbf{x}}^n_{test}, \tilde{\mathbf{x}}^k)$. Figure 8 visualizes the neighborhood of an exemplary test consultation representation $\tilde{\mathbf{x}}^n_{test}$.

In contrast to the *consultation-therapy outcome matrices* $\tilde{\mathbf{A}}_{train}$ and $\tilde{\mathbf{A}}_{test}$, the attributes in the *consultations data matrix* $\tilde{\mathbf{X}}$ are highly heterogeneous, i.e., they are of various level of measurement. Thus, the similarity function applied in Sect. 1 to determine $s^{n,k}$ are not appropriate for the concatenated data. Two similarity measures are contrasted

The *Gower similarity coefficient* measures similarity at the presence of mixed data types and can even cope with missing values. The similarity function $s_{GSC}(\tilde{\mathbf{x}}^n_{test}, \tilde{\mathbf{x}}^k)$ defines an overall coefficient $s^{n,k}$ which is computed out of the individual attribute similarities $\rho^{n,k}_d$, depending on their presence $\delta^{n,k}_d$ and assigned weights w_d.

$$s^{n,k} = \frac{\sum_{d=1}^{D} \delta^{n,k}_d \cdot w_d \cdot \rho^{n,k}_d}{\sum_{d=1}^{D} \delta^{n,k}_d \cdot w_d} \tag{2}$$

$\rho^{n,k}_d$ quantifies the similarity between two instances according to the dth attribute, depending on the data type. The coefficient $\delta^{n,k}_d$ controls whether to include $\rho^{n,k}_d$ into the similarity computation or not. $\delta^{n,k}_d$ is set to 1 if the respective attribute is known for both instances and set to 0 otherwise.

Furthermore, also the *Euclidean distance* can be employed do derive a similarity function $s_{Euc}(\tilde{\mathbf{x}}^n_{test}, \tilde{\mathbf{x}}^k)$ using a RBF as introduced above. Prerequisite for

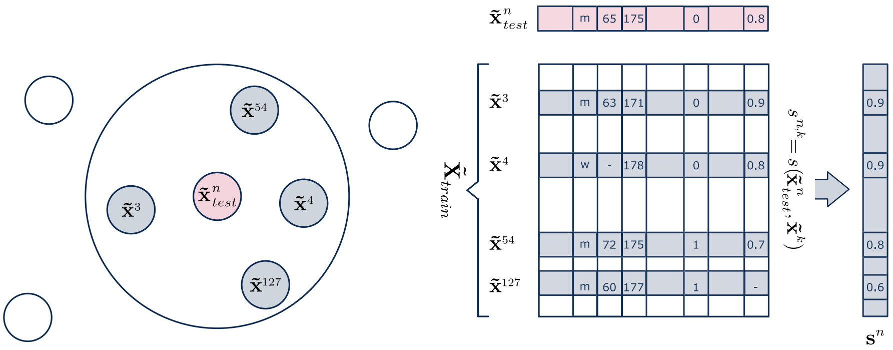

$\tilde{\mathbf{x}}^{54}$
$\tilde{\mathbf{x}}^{3}$
$\tilde{\mathbf{x}}^{n}_{test}$
$\tilde{\mathbf{x}}^{4}$
$\tilde{\mathbf{x}}^{127}$
$\tilde{\mathbf{x}}^{n}_{test}$
m 65 175 0 0.8
$\tilde{\mathbf{X}}_{train}$
$\tilde{\mathbf{x}}^{3}$
m 63 171 0 0.9
$\tilde{\mathbf{x}}^{4}$
w - 178 0 0.8
$\tilde{\mathbf{x}}^{54}$
m 72 175 1 0.7
$\tilde{\mathbf{x}}^{127}$
m 60 177 1 -
$s^{n,k} = s(\tilde{\mathbf{x}}^{n}_{test}, \tilde{\mathbf{x}}^{k})$
0.9
0.9
0.8
0.6
$\mathbf{s}^{n}$

Fig. 8 In this patient-data CF approach, consultations are compared regarding patient data and treatment history stored in $\tilde{\mathbf{x}}^k$ and $\tilde{\mathbf{x}}^n_{test}$, respectively

computing the *Minkowski metric* are all attributes in the attribute space having a numeric (quantitative) data type which allows for pairwise attribute subtraction. Hence, categorical (qualitative) attributes must be converted to a numeric scale, namely at least the interval scale. Subtraction of dichotomous attributes can be regarded to yield valid distance measures in the value range [0, 1]. Nominal attributes are converted by applying *one-hot-encoding* which creates one dichotomous dummy feature for each of the available categories of a specific attributes. Those, in turn, allow for subtraction as stated before. Ordinal attributes, finally, require transformation to an interval scale under appropriate assumptions regarding the distance between adjacent ordinal categories. Here, for the sake of simplicity, all ordinal variables are assumed to have equidistant categories. As a result of this attribute preprocessing strategy, the dimension of the consultation representation $\tilde{\mathbf{X}}$ is further expanded to $D = 159$ attributes. Additionally, in comparison with the patient-data CF utilizing the *Gower similarity coefficient*, which already incorporates data normalization, utilizing *Euclidean distance* requires normalization as an essential preprocessing step. All attributes are rescaled to the closed unit interval [0, 1] by subtracting minimum values and dividing each attribute $\tilde{\mathbf{x}}$ by its range (min-max normalization). Comparably to the *Gower similarity coefficient*, also *Euclidean distance* is only computed on mutually available attributes when comparing consultation representations. Hence, $s_{Euc}(\tilde{\mathbf{x}}^n_{test}, \tilde{\mathbf{x}}^k)$ defines the similarity coefficient $s^{n,k}$ as

$$s^{n,k} = K_\sigma\left(\sqrt{\frac{\sum_{d=1}^{D} \delta_d^{n,k} \cdot (\tilde{\mathbf{x}}_d^k - \tilde{\mathbf{x}}_{test,d}^n)^2}{\sum_{d=1}^{D} \delta_d^{n,k}}}\right) \tag{3}$$

with the coefficient $\delta_d^{n,k}$ controlling whether to include the dth attribute into the similarity computation or not and the RBF kernel $K_\sigma(\cdot)$.

Attribute weighting (DR-RBA)

Selection or adjustment of attribute weights can either be based on *a priori* knowledge or is extracted automatically from the given data. Various filter methods from the feature selection domain (Kira and Rendell 1992; Peng et al. 2005; Yu and Liu 2003; Pudil et al. 1994) are suitable to assign numeric values to individual attributes which reflect their importance regarding a given classification or regression task. Based on supervised information, such as known class labels or dependent variables, univariate or multivariate criteria are typically defined which measure importance. In the proposed neighborhood-based CF approach, it is an obvious choice to incorporate *a priori* assumptions concerning similarity and dissimilarity between consultation representations and determine attribute weights such that a similarity criterion is optimized.

A widely and successfully used class of attribute weighting and selection algorithms, which exploit the concept of similarity, are RBAs, as initially proposed in Kira and Rendell (1992) and, among others, extended by Kononenko et al. (1997). Within this work, a generalization of the mentioned algorithms is adapted to the given patient-data CF approach as was demonstrated in Gräßer et al. (2019). The attribute weights are determined for each outer cross-validation loop using assumptions regarding similarity and dissimilarity of training instances $\tilde{\mathbf{X}}^{p}_{train}$. Within an iterative process, the attribute weights w_d are updated such that large distances for similar instances are penalized and *vice versa*.

$$w_d = w_d + (\bar{\rho}_d^{hits} - \bar{\rho}_d^{misses})/J \tag{4}$$

The adaption of an attribute weight w_d is determined by the K_{RBA} nearest neighbors of the target with the same class, i.e., the *nearest hits* and the K_{RBA} nearest neighbors with different class, i.e., *nearest misses*. The average of observed value differences $\bar{\rho}_d^{hits}$ and $\bar{\rho}_d^{misses}$ computed for an attribute d between target $\tilde{\mathbf{x}}^j$ and the respective neighboring instances determine the update of the attribute weight w_d in each iteration. The values $\bar{\rho}_d^{hits}$ and $\bar{\rho}_d^{misses}$ are normalized by the number of iterations J. In accordance with the applied *Gower similarity coefficient*, similarity between two samples is quantified with a ρ_d depending on the data type of the dth attribute. As proposed by Kira and Rendell (1992), all attribute weights dropping below a predefined *relevance threshold* thr_w, are discarded. In total, three additional hyperparameter need to be determined within the inner cross-validation loop. The number of nearest hits and nearest misses K_{RBA}, the initial feature weight vector $\mathbf{w}^{init}$, and the weight threshold thr_w for feature selection.

The RBA assumes a supervised classification problem where each sample is associated with a distinct class. In the context of the present problem, each sample, i.e., consultation, is characterized by a numeric outcome indicator for the applied therapy option and unknown outcome for all other options which have not been applied (*hidden ground truth*). Consequently, *a priori* assumptions concerning the relationship, i.e., similarity or dissimilarity between a pair of consultations can only be derived from those samples which applied therapies in common and for which in both cases outcome is known. Regarding this relationship, three groups can be distinguished. (1) Two consultations are similar to each other, if the respective patients respond similarly to the given treatment option. Both consultations are labeled with the same therapy and outcome is similar. (2) Two consultations are dissimilar to each other, if the respective patients respond differently to the given treatment. Both consultations are labeled with the same treatment but outcome differs. (3) No information on similarity is available for a pair of consultations which are labeled with different therapies. The response of the respective neighboring patient on the treatment given to the target patient is unobserved. As stated, in the context of the RBA algorithm, *nearest hits* are the closest observations to the target observation which are considered to be similar, whereas *nearest misses* are the closest observations which are considered to be dissimilar. Therefore, applying the groups described above, *nearest hits* to a target consultation $\tilde{\mathbf{x}}^j$ are the K_{RBA} closest consultations associated with equal therapy and similar patient response, whereas *nearest misses* are the K_{RBA} closest observations to $\tilde{\mathbf{x}}^j$ associated to equal therapy but differing outcome. Here, similar response means that both outcome indicators, i.e., *affinity* scores have the same polarity regarding a predefined threshold $thr_{good} = 0.5$ which divides treatment responses into *good* and *bad* outcome classes. The neighboring consultations to a target consultation $\tilde{\mathbf{x}}^j$ associated with different therapy options are, independent of their outcome, not included into the K_{RBA} neighbors as they hold no information regarding the relationship between $\tilde{\mathbf{x}}^j$ and those consultations.

Figure 9 illustrates an exemplary neighborhood of the representation $\tilde{\mathbf{x}}^j$ of a target consultation j where the applied treatment, here m_1, showed good response (*affinity* > 0.5). All K_{RBA} neighboring consultations $\tilde{\mathbf{x}}^k$ are labeled as similar to $\tilde{\mathbf{x}}^j$ if the same treatment is present in $\tilde{\mathbf{a}}^k$, which is derived from the *complete consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{all}$, and if the respective treatment has also shown good response, i.e., equal polarity (green). Conversely, all K_{RBA} neighboring consultations are labeled as dissimilar to consultations $\tilde{\mathbf{x}}^j$ if the same treatment is present in $\tilde{\mathbf{a}}^k$ but this treatment has shown bad response, i.e., has different polarity (red). Neighboring consultation representations with equal treatment applied and equal polarity are considered as *nearest hits* and representations with equal treatment applied but differing polarity are considered as *nearest misses*. Regarding neighboring consultation representations $\tilde{\mathbf{x}}^j$ for which is true that the in consultation j applied therapy was never applied, no information regarding the similarity label is available. Training consultation $k = 54$ (white) is not associated with therapy m_1 but with different therapy options and hence is discarded.

Metric learning (DR-LMNN)

Mahalanobis distance (Lowsky et al. 2013; Panahiazar et al. 2015) incorporates linear transformation before computing *Euclidean distance* between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ in the transformed attribute space according to

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) \quad (5)$$

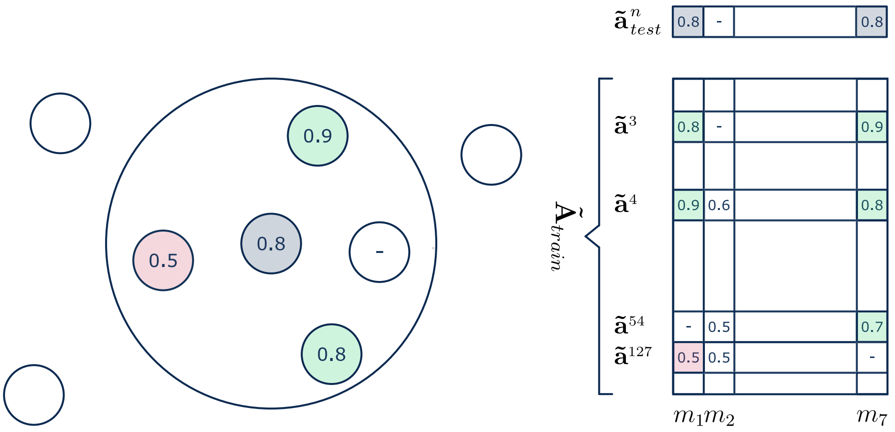


Fig. 9 RBA algorithm assumes a supervised classification problem. Consultations in the neighborhood of a target consultation j are labeled as similar or dissimilar if the same treatment was applied, here treatment option m_1, and the outcome polarity is equal (green) or different (red), respectively. No information regarding the similarity label is available for consultations were the treatment applied in consultation j was never applied (white)

Any positive semidefinite matrix is applicable as squared transformation matrix $\mathbf{M}$ in order to yield a valid (pseudo-) metric. Employing the inverse covariance matrix as $\mathbf{M}$, the data is decorrelated by rotating the basis, and scaled to unit variance. Accordingly, the classical *Mahalanobis distance* considers the distribution of the data by measuring distance in standard deviations along the principal components of the present data. Generalized *Mahalanobis metrics* can exploit additional *a priori* information. The objective of such supervised approaches is to learn a *Mahalanobis metric* based on a transformation matrix $\mathbf{M}$ that takes into account both, the distribution of the data and known similarity and dissimilarity constraints. The LMNN algorithm proposed by Weinberger et al. (2005) and adapted to therapy recommendation in our previous work (Gräßer et al. 2019) learns such a generalized *Mahalanobis metric* and is especially intended for neighborhood-based classification algorithms. Prerequisite for applying linear transformation to the data is that all attributes in the attribute space must have equal quantitative data type and are normalization to the closed unit interval [0, 1] as described in "Patient-data Collaborative Filter (DR)" section.

Comparable to the proposed RBA algorithm, a squared transformation matrix $\mathbf{M}$ is learned for each outer cross-validation loop using the entire training sets $\tilde{\mathbf{X}}^p_{train}$. The overall intention of the LMNN algorithm is to learn a global transformation such that it causes a target consultation representation $\tilde{\mathbf{x}}^j$ to be surrounded by consultations of the same class while being separated from consultations of different classes. To do so, the loss function to be minimized is composed of two competing objectives ϵ_{pull} and ϵ_{push}. Its relative impact is controlled using a meta parameter $\mu \in [0, 1]$ which is to be tuned in the inner cross-validation loops. Firstly, for each target consultation representation $\tilde{\mathbf{x}}^j$, the K_{LMNN} nearest neighbors with the same class, denoted as *target neighbors* should be close. Therefore, large average distances between $\tilde{\mathbf{x}}^j$ and the K_{LMNN} closest consultation representations $\tilde{\mathbf{x}}^k$ labeled as similar are penalized. Here, the binary matrix $\boldsymbol{\eta}$ indicates whether $\tilde{\mathbf{x}}^k$ is a target neighbor of $\tilde{\mathbf{x}}^j$ and the binary matrix $\mathbf{y}^0$ indicates whether labels in $\tilde{\mathbf{y}}^j$ and $\tilde{\mathbf{y}}^k$ match, respectively.

$$\epsilon_{pull}(\mathbf{L}) = \sum_{j,k} \eta_{jk} ||\mathbf{L}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^k)||^2 \tag{6}$$

Secondly, small distances between $\tilde{\mathbf{x}}^j$ and consultations labeled as dissimilar and which invade the perimeter established by the *target neighbors*, denoted as *impostors*, are penalized. To increase the robustness of the underlying K-nearest-neighbor classifier (KNN) classification and to cope with noise in the training data, an additional unit margin is added around the KNN decision boundaries, i.e., the perimeters established by the *target neighbors*. The *hinge loss* $[z]_+ = max(z, 0)$ ensures not all samples with different label but only *impostors* to contribute to the loss function.

$$\epsilon_{push}(\mathbf{L}) = \sum_{j,k,l} \eta_{jk}(1 - y^0_{jl})[1 + ||\mathbf{L}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^k)||^2 - ||\mathbf{L}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^l)||^2]_+ \tag{7}$$

By minimizing the combined loss function

$$\epsilon(\mathbf{L}) = (1 - \mu)\epsilon_{pull}(\mathbf{L}) + \mu\epsilon_{push}(\mathbf{L}) \tag{8}$$

a transformation is learned which pulls the K_{LMNN} *target neighbors* toward $\tilde{\mathbf{x}}^j$ and pushes *impostors* outside the KNN decision boundaries plus unit margin.

As the RBA, the LMNN algorithm assumes a supervised classification problem where each sample is associated with one class label which corresponds to a distinct ground truth. *Target neighbors* are the K_{LMNN} closest observations to a target observation, which are considered to be similar, whereas *impostors* are too close observations, which are considered to be dissimilar. Applying the three groups described in "Appendix Attribute Weighting (DR-RBA)", *target neighbors* are the K_{LMNN} closest consultations associated with equal therapy and similar patient response, whereas *impostors* are consultation representations invading the neighborhood defined by the *target neighbors* which are labeled with equal therapy but differing outcome. Equally to the RBA definition, treatment responses are divided into *good* and *bad* outcome classes by applying the predefined affinity threshold $thr_{good} = 0.5$. All consultations which are labeled with different therapy options compared to the target consultation $\tilde{\mathbf{x}}^j$ are not included into the respective cost definition as they hold no information regarding the relationship between $\tilde{\mathbf{x}}^j$ and these consultations.

Figure 10 illustrates an exemplary neighborhood of a target patient $\tilde{\mathbf{x}}^j$ with good outcome where all $K_{LMNN} = 3$ *target neighbors* (green), i.e., neighboring consultations with equal polarity, are supposed to be pulled toward $\tilde{\mathbf{x}}^j$. Consultation representations with differing polarity, i.e., bad outcome, which invade the neighborhood defined by the *target neighbors* are considered as *impostors* (red) and are supposed to be pushed outside the KNN decision boundaries plus unit margin. Consultation $k = 54$ (white) is associated with a different therapy options and hence is discarded.

Therapy recommender system user interface

Patient data input and presentation

For each patient and new visit, the physician is prompted to update or confirm condition related patient data, such as demographic data, diagnosis, disease severity, comorbidities, and life situation. Severity and impairment caused by the disease is measured using clinical scores such the PASI score. The development of such numeric scores over the course of treatment is visualized in charts. Moreover, the treating physician is requested to assess outcome of the currently applied treatment (efficiency and ADEs) and to add the newly prescribed or administered treatment option. PASI development and the physician's assessments are utilized to derive the summarizing *affinity* score associated with an applied drug and hence is the foundation of the above proposed recommendation algorithms. The individual data categories are grouped into cards. Figure 11 displays the data of an exemplary patient and visit.

Recommendation dashboard

The proposed recommendation dashboard is shown in Fig. 12. Outcome estimates (i.e., *affinity* score) for each therapy option are visualized in a bar chart and are intended to give an overall prognosis and actionable decision support. By hovering

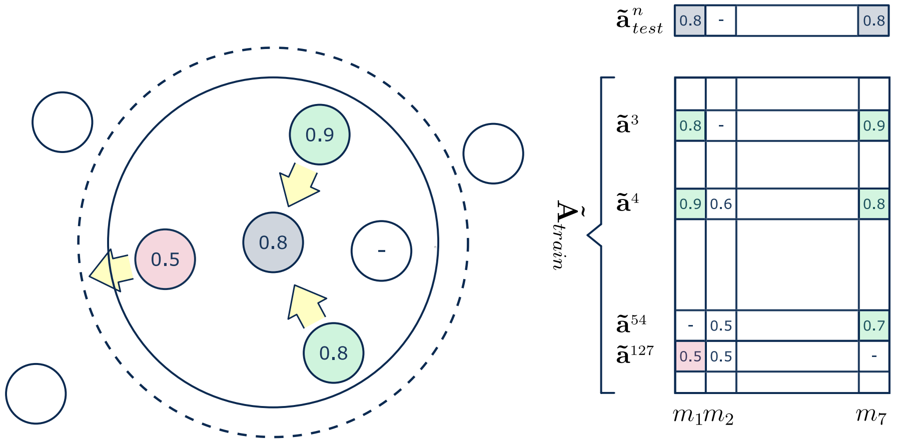
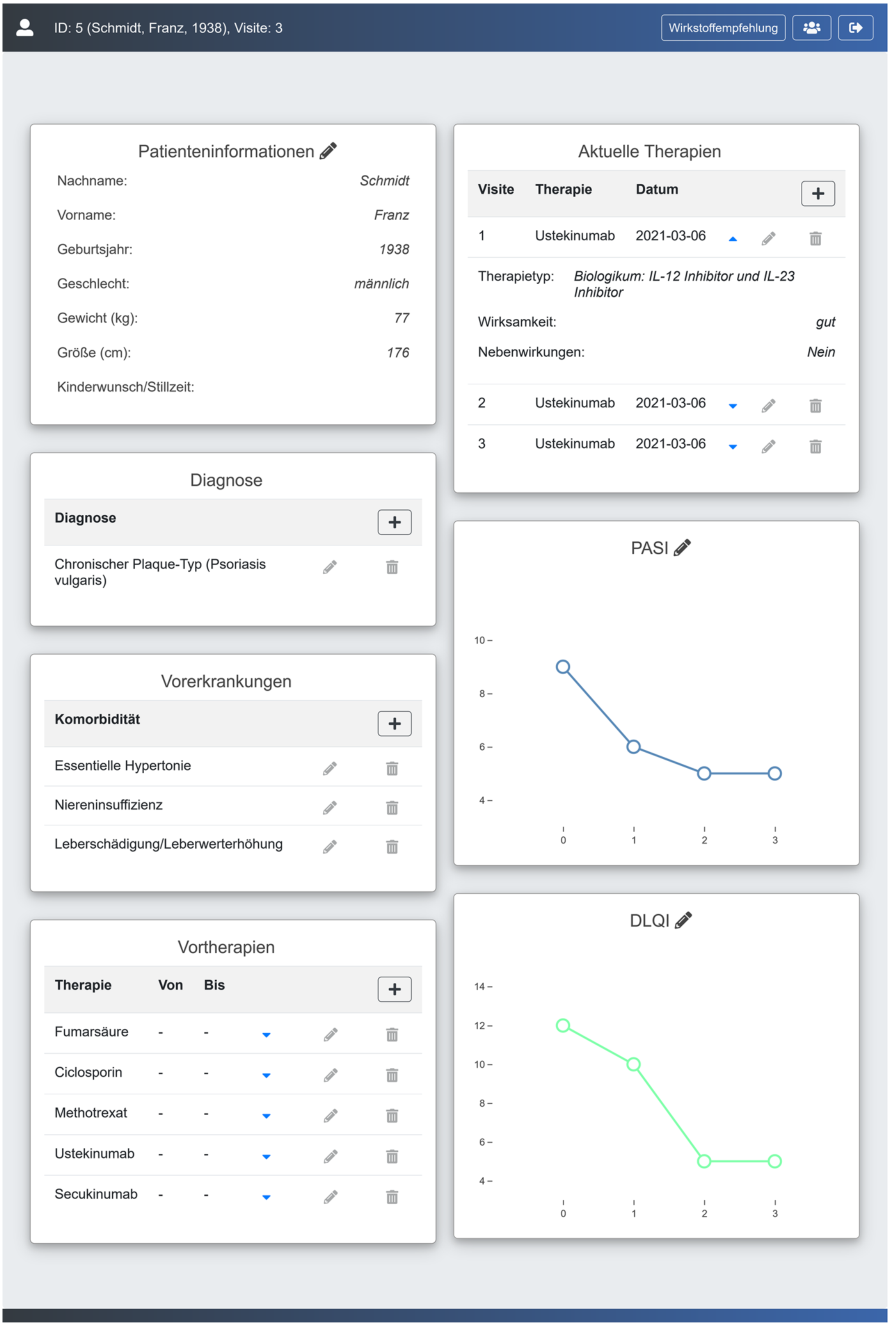


Fig. 10 LMNN algorithm assumes a supervised classification problem. Consultations are labeled with respect to a target consultation j and *a priori* similarity and dissimilarity assumptions as introduced in Sect. 3. The LMNN algorithm intents to cause the target consultation representation $\tilde{\mathbf{x}}^{j}$ to be surrounded by samples of the same class while being separated from samples of different classes



ID: 5 (Schmidt, Franz, 1938), Visite: 3

Patienteninformationen

Nachname:	*Schmidt*
Vorname:	*Franz*
Geburtsjahr:	*1938*
Geschlecht:	*männlich*
Gewicht (kg):	*77*
Größe (cm):	*176*
Kinderwunsch/Stillzeit:	

Diagnose

Diagnose
Chronischer Plaque-Typ (Psoriasis vulgaris)

Vorerkrankungen

Komorbidität
Essentielle Hypertonie
Niereninsuffizienz
Leberschädigung/Leberwerterhöhung

Vortherapien

Therapie	Von	Bis
Fumarsäure	-	-
Ciclosporin	-	-
Methotrexat	-	-
Ustekinumab	-	-
Secukinumab	-	-

Aktuelle Therapien

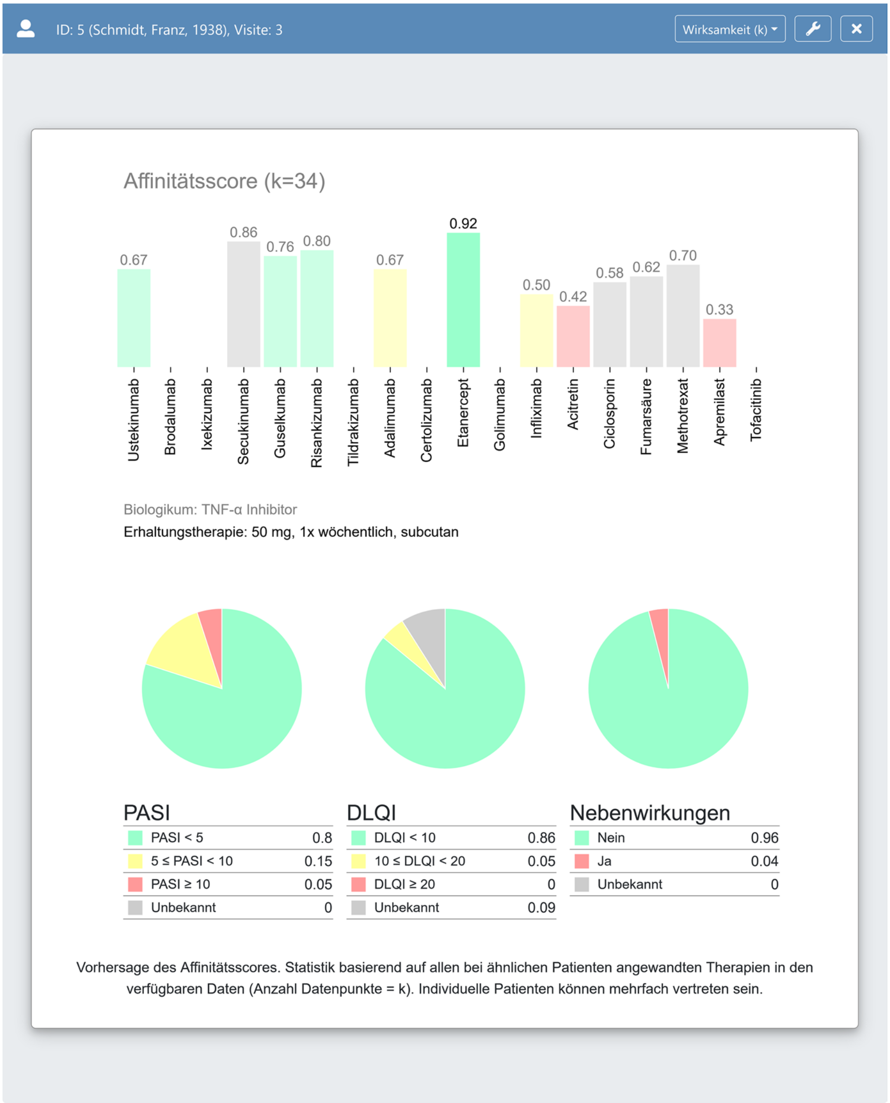
Visite	Therapie	Datum
1	Ustekinumab	2021-03-06

Therapietyp:	*Biologikum: IL-12 Inhibitor und IL-23 Inhibitor*
Wirksamkeit:	*gut*
Nebenwirkungen:	*Nein*

Visite	Therapie	Datum
2	Ustekinumab	2021-03-06
3	Ustekinumab	2021-03-06

Fig. 11 Psoriasis therapy recommender system GUI: Patient and previous therapy data presentation. Patient data, such as demographic data, diagnosis, comorbidities, and clinical scores, as well as information on previous therapies and outcomes are presented for the selected patient and consultation and are editable

over a bar, additional information as brand name and biochemical or physiologic effects, i.e., pharmacodynamics of the selected drug, are displayed below the chart. Optionally, therapy options can be sorted alphabetically, by pharmacodynamics, or by predicted *affinity* score. Clicking on a treatment option opens a pop-up window



Biologikum: TNF-α Inhibitor

Erhaltungstherapie: 50 mg, 1x wöchentlich, subcutan

PASI	
🟩 PASI < 5	0.8
🟨 5 ≤ PASI < 10	0.15
🟥 PASI ≥ 10	0.05
▨ Unbekannt	0

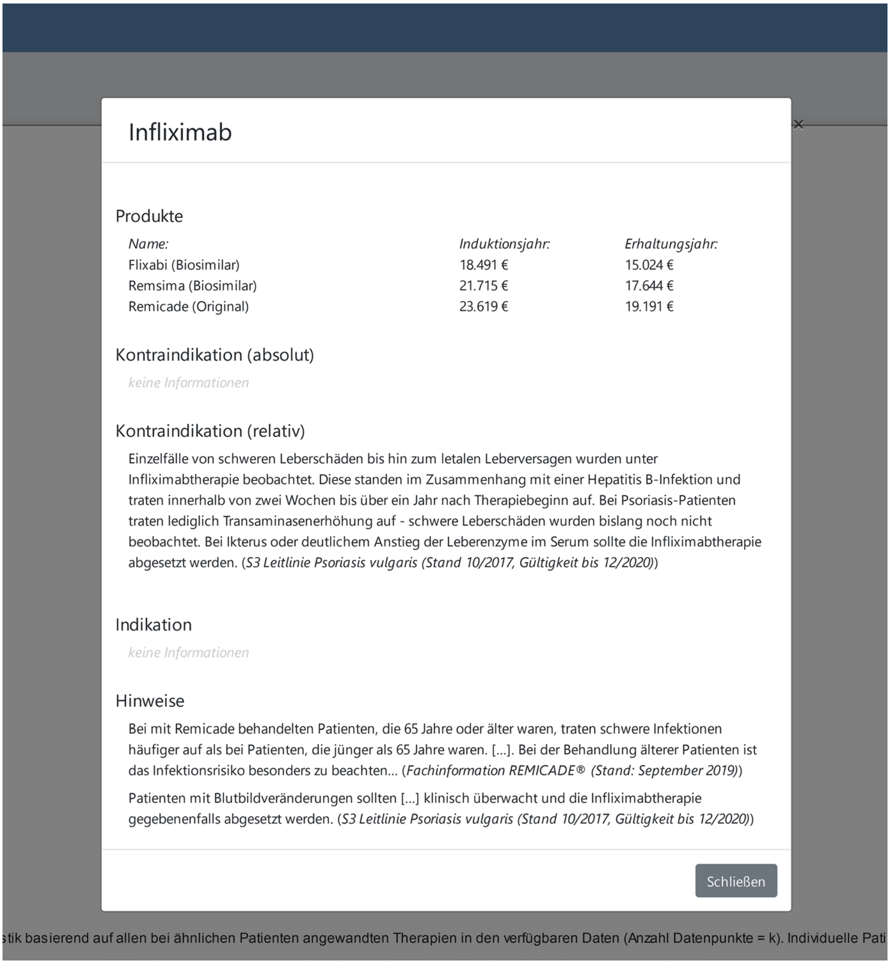
DLQI	
🟩 DLQI < 10	0.86
🟨 10 ≤ DLQI < 20	0.05
🟥 DLQI ≥ 20	0
▨ Unbekannt	0.09

Nebenwirkungen	
🟩 Nein	0.96
🟥 Ja	0.04
▨ Unbekannt	0

Vorhersage des Affinitätsscores. Statistik basierend auf allen bei ähnlichen Patienten angewandten Therapien in den verfügbaren Daten (Anzahl Datenpunkte = k). Individuelle Patienten können mehrfach vertreten sein.

Fig. 12 Psoriasis therapy recommender system GUI: Recommendation dashboard. The predicted *affinity* scores for each therapy option after post filtering are visualized in a bar chart. The bar colors indicate already applied and discontinued treatments (grey), and relative (yellow) and absolutely (red) contraindicated options. By hovering over an option, summary statistics derived from the local neighborhood of the target consultation are shown for each of the outcome indicators. Moreover, the pharmaceutical's mechanism of action, mode of administration and dosage information are shown

with supplementary information, namely treatment costs and information on indication and absolute and relative contraindications as exemplarily shown in Fig. 13. This information is derived from the relevant clinical guideline (Nast et al. 2017) and pharmaceutical directories and is tailored to the selected patient's characteristics. Absolute and relative contraindications are further grouped into the four categories (i) drug not approved for the present diagnosis, (ii) contraindication due to comorbidities or life situation (e.g., pregnancy), (iii) contraindication due to drug–drug interaction, and (iv) violation of the sequence of drug application defined by the guideline. Therapy options can optionally be color-coded or filtered, if absolute or



Infliximab

Produkte

Name:	*Induktionsjahr:*	*Erhaltungsjahr:*
Flixabi (Biosimilar)	18.491 €	15.024 €
Remsima (Biosimilar)	21.715 €	17.644 €
Remicade (Original)	23.619 €	19.191 €

Kontraindikation (absolut)

keine Informationen

Kontraindikation (relativ)

Einzelfälle von schweren Leberschäden bis hin zum letalen Leberversagen wurden unter Infliximabtherapie beobachtet. Diese standen im Zusammenhang mit einer Hepatitis B-Infektion und traten innerhalb von zwei Wochen bis über ein Jahr nach Therapiebeginn auf. Bei Psoriasis-Patienten traten lediglich Transaminasenerhöhung auf - schwere Leberschäden wurden bislang noch nicht beobachtet. Bei Ikterus oder deutlichem Anstieg der Leberenzyme im Serum sollte die Infliximabtherapie abgesetzt werden. (*S3 Leitlinie Psoriasis vulgaris (Stand 10/2017, Gültigkeit bis 12/2020)*)

Indikation

keine Informationen

Hinweise

Bei mit Remicade behandelten Patienten, die 65 Jahre oder älter waren, traten schwere Infektionen häufiger auf als bei Patienten, die jünger als 65 Jahre waren. [...]. Bei der Behandlung älterer Patienten ist das Infektionsrisiko besonders zu beachten... (*Fachinformation REMICADE® (Stand: September 2019)*)

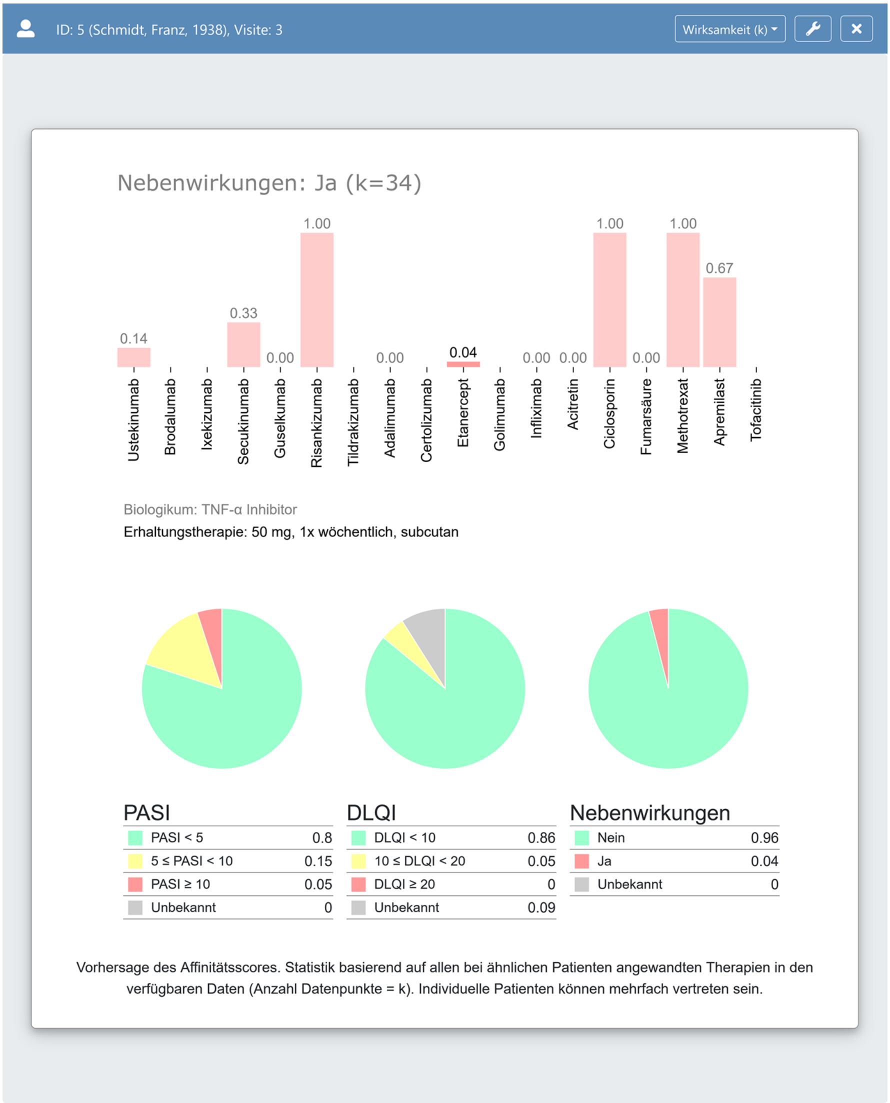
Patienten mit Blutbildveränderungen sollten [...] klinisch überwacht und die Infliximabtherapie gegebenenfalls abgesetzt werden. (*S3 Leitlinie Psoriasis vulgaris (Stand 10/2017, Gültigkeit bis 12/2020)*)

stik basierend auf allen bei ähnlichen Patienten angewandten Therapien in den verfügbaren Daten (Anzahl Datenpunkte = k). Individuelle Pati

Fig. 13 Psoriasis therapy recommender system GUI: Recommendation dashboard. Clicking on a treatment option opens a pop-up window with supplementary information. For the selected systemic pharmaceutical, available products along with costs are shown. Additionally, detailed information on indication, absolute and relative contraindications, which is derived from clinical guidelines and pharmaceutical directories, are displayed

relative contraindications are applicable for that particular therapy option and patient characteristics. Also the included information source—clinical guideline and pharmaceutical directory—and the included contraindication category mentioned above can be customized by the user. The proposed presentation of patient-specific information and the evidence-based post-filtering of recommendations allows an integration of external evidence and a reduced risk of recommending contraindicated or even health-endangering drugs. The customization of added information and filters intends to meet the requirement for tailoring recommendations to patient characteristics and current needs.

Beyond the overall recommendation, the dashboard provides supplementary details regarding individual outcome aspects for a selected therapy option. For each outcome indicator, summary statistics from the local neighborhood, on which the recommendation is based, are shown for a selected therapy option. Therefore, within this neighborhood the fraction of manifestations—in case of nominal or ordinal indicators—or observed values falling into a respective bin—in case of nominal indicators—, are visualized in pie charts and summarized in tables. When hovering over an outcome aspect and category, either on pie chart or table, the respective fraction of the selected outcome aspect and category is shown for each therapy option instead of the *affinity* score bars. This function is exemplarily shown in Fig. 14 for the aspect “ADEs” and category “yes”. The benefit of this visualization approach is twofold. Firstly, presenting the observed local distribution of the individual *affinity* score components increases transparency and interpretability of recommendations and facilitates assessment of recommendation reliability. Secondly, presenting the fraction of an outcome aspect and category for each therapy option can be interpreted as the probability for the individual treatments to yield a similar response. This, in turn, facilitates to choose treatment options not based on the summarizing *affinity* score prediction but with respect to a specific outcome aspect depending on the target patient’s preferences. If, for example, the target patient’s main objective is a treatment with low ADE risk, the fraction of neighboring patients with no ADEs



Nebenwirkungen: Ja (k=34)

Biologikum: TNF-α Inhibitor

Erhaltungstherapie: 50 mg, 1x wöchentlich, subcutan

PASI	
🟩 PASI < 5	0.8
🟨 5 ≤ PASI < 10	0.15
🟥 PASI ≥ 10	0.05
⬜ Unbekannt	0

DLQI	
🟩 DLQI < 10	0.86
🟨 10 ≤ DLQI < 20	0.05
🟥 DLQI ≥ 20	0
⬜ Unbekannt	0.09

Nebenwirkungen	
🟩 Nein	0.96
🟥 Ja	0.04
⬜ Unbekannt	0

Vorhersage des Affinitätsscores. Statistik basierend auf allen bei ähnlichen Patienten angewandten Therapien in den verfügbaren Daten (Anzahl Datenpunkte = k). Individuelle Patienten können mehrfach vertreten sein.

Fig. 14 Psoriasis therapy recommender system GUI: Recommendation dashboard. By hovering over an outcome aspect and category, either on pie chart or table, the respective fraction of the selected outcome aspect and category is shown for each therapy option instead of the *affinity* score bars. This function is exemplarily shown for the aspect "ADEs" and category "yes"

is shown for each treatment options and the treatment with the largest value can be chosen. This allows patient preferences and values to be taken into account and treatment decisions to be made together with the patient.

Acknowledgements This work is part of the project *DARE* which is funded by *Sächsisches Staatsministerium für Wissenschaft, Kultur und Tourismus (SMWK).*

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Avorn, J.: The psychology of clinical decision making - Implications for medication use. (2018). https://doi.org/10.1056/NEJMp1714987

Barnett, G.O., Cimino, J.J., Hupp, J.A., Hoffer, E.P.: DXplain: an evolving diagnostic decision-support system. JAMA J. Am. Med. Assoc. **258**(1), 67–74 (1987)

Barratt, A.: Evidence based medicine and shared decision making: the challenge of getting both evidence and preferences into health care. Patient Educ. Counsel. (2008). https://doi.org/10.1016/j.pec.2008.07.054

Beeler, P.E., Bates, D.W., Hug, B.L. (2014). Clinical decision support systems, Swiss Med. Weekly, https://doi.org/10.4414/smw.2014.14073

Berner, E.S., La Lande, T.J.: Overview of Clinical Decision Support Systems. pp 1–17, (2016) https://doi.org/10.1007/978-3-319-31913-1_1

Calero Valdez, A., Ziefle, M., Verbert, K., Felfernig, A., Holzinger, A.: Recommender Systems for Health Informatics: State-of-the-Art and Future Perspectives, pp. 391–414. Springer International Publishing, Cham (2016)

Campbell-Scherer, D.: Multimorbidity: A challenge for evidence-based medicine. (2010) https://doi.org/10.1136/ebm1154

Celi, L.A., Zimolzak, A.J., Stone, D.J.: Dynamic clinical data mining: Search engine-based decision. (2014). https://doi.org/10.2196/medinform.3110

Chawla, N.V., Da, Davis: Bringing big data to personalized healthcare: a patient-centered framework. J. General Internal Med. **28**, 660–665 (2013). https://doi.org/10.1007/s11606-013-2455-8

Chen, J.H., Altman, R.B.: Mining for clinical expertise in (undocumented) order sets to power an order suggestion system. AMIA Joint Summits Translat. Sci. proceed. **2013**, 34–8 (2013)

Chen, J.H., Altman, R.B.: Automated Physician Order Recommendations and Outcome Predictions by Data-Mining Electronic Medical Records. AMIA Summits on Translational Science proceedings pp 206–210 (2014)

Chiang, W.H., Shen, L., Li, L., Ning, X.: Drug Recommendation toward Safe Polypharmacy. KDD'18 1803.03185 (2018)

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, (2014) https://doi.org/10.3115/v1/d14-1179, 1406.1078

Croskerry, P.: Clinical cognition and diagnostic error: applications of a dual process model of reasoning. Adv. Health Sci. Educ. (2009). https://doi.org/10.1007/s10459-009-9182-2

Davis, D.A., Chawla, N.V., Christakis, N.A., Barabási, A.L.: Time to CARE: a collaborative engine for practical disease prediction. Data Mining Knowl. Discov. (2010). https://doi.org/10.1007/s10618-009-0156-z

Del Mar, C., Doust, J., Glasziou, P.: Clinical thinking: evidence. Commun. Decis. Mak. (2007). https://doi.org/10.1002/9780470750568

Duan, L., Street, W.N., Lu, D.F.: A Nursing Care Plan Recommender System Using A Data Mining Approach. Proceedings of the 3rd INFORMS Workshop on Data Mining and Health Informatics pp 1–6 (2008)

Duan, L., Street, W.N., Xu, E.: Healthcare information systems: data mining methods in the creation of a clinical recommender system. Enterprise Info. Syst. **5**(2), 169–181 (2011). https://doi.org/10.1080/17517575.2010.541287

Faries, D.E., Chen, Y., Lipkovich, I., Zagar, A., Liu, X., Obenchain, R.L.: Local control for identifying subgroups of interest in observational research: persistence of treatment for major depressive disorder. Int. J. Methods Psych. Res. (2013). https://doi.org/10.1002/mpr.1390

Folino, F., Pizzuti, C.: A comorbidity-based recommendation engine for disease prediction. Proceedings - IEEE Symposium on Computer-Based Medical Systems (2010). https://doi.org/10.1109/CBMS.2010.6042664

Folino, F., Pizzuti, C.: A recommendation engine for disease prediction. Info. Syst. e-Business Manage. **13**(4), 609–628 (2015). https://doi.org/10.1007/s10257-014-0242-7

Fortin, M., Dionne, J., Pinho, G., Gignac, J., Almirall, J., Lapointe, L.: Randomized controlled trials: do they have external validity for patients with multiple comorbidities? Annals Family Med. (2006). https://doi.org/10.1370/afm.516

Frankovich, J., Longhurst, C.A., Sutherland, S.M.: Evidence-Based Medicine in the EMR Era. New England Journal of Medicine **365**(19):1758–1759, (2011) https://doi.org/10.1056/NEJMp1108726, arXiv:1011.1669v3

Fredriksson, T., Pettersson, U.: Severe psoriasis - oral therapy with a new retinoid. Dermatology **157**(4), 238–244 (1978). https://doi.org/10.1159/000250839

Friedman, M.: The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Statist. Assoc. (1937). https://doi.org/10.1080/01621459.1937.10503522

Gallego, B., Walter, S.R., Day, R.O., Dunn, A.G., Sivaraman, V., Shah, N., Longhurst, C.A., Coiera, E.: Bringing cohort studies to the bedside: Framework for a ”green button” to support clinical decision-making. (2015). https://doi.org/10.2217/cer.15.12

Garg, A.X., Adhikari, N.K.J., McDonald, H., Rosas-Arellano, M.P., Devereaux, P.J., Beyene, J., Sam, J., Haynes, R.B.: Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. JAMA (2005). https://doi.org/10.1001/jama.293.10.1223

Gräßer, F., Beckert, S., Küster, D., Schmitt, J., Abraham, S., Malberg, H., Zaunseder, S.: Therapy decision support based on recommender system methods. J. Healthcare Eng. (2017). https://doi.org/10.1155/2017/8659460

Gräßer, F., Malberg, H., Kallumadi, S., Zaunseder, S.: Aspect-Based sentiment analysis of drug reviews applying cross-Domain and cross-Data learning. ACM International Conference Proceeding Series, doi **10**(1145/3194658), 3194677 (2018)

Gräßer, F., Malberg, H., Zaunseder, S.: Neighborhood optimization for therapy decision support. Curr. Direct. Biomed. Eng. **5**(1), 1–4 (2019). https://doi.org/10.1515/cdbme-2019-0001

Groves, M.: Understanding clinical reasoning: The next step in working out how it really works. (2012). https://doi.org/10.1111/j.1365-2923.2012.04244.x

Haas, P.J.: Medizinische Informationssysteme und Elektronische Krankenakten. Springer-Verlag (2005). https://doi.org/10.1007/b138207

Hao, F., Blair, R.H.: A comparative study: classification vs user-based collaborative filtering for clinical prediction. BMC Med. Res. Methodol. (2016). https://doi.org/10.1186/s12874-016-0261-9

Hassan, S., Syed, Z.: From netflix to heart attacks: Collaborative filtering in medical datasets. IHI'10 - Proceedings of the 1st ACM International Health Informatics Symposium pp 128–134, (2010) https://doi.org/10.1145/1882992.1883012

Hidasi, B., Karatzoglou, A.: Recurrent neural networks with Top-k gains for session-based recommendations. In: International Conference on Information and Knowledge Management, Proceedings, Association for Computing Machinery, pp 843–852, (2018) https://doi.org/10.1145/3269206.3271761, 1706.03847

Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computat. **9**(8), 1735–1780 (1997)

Holm, S.: A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics (1979)

Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform. **3**(2), 119–131 (2016)

Hors-Fraile, S., Rivera-Romero, O., Schneider, F., Fernandez-Luque, L., Luna-Perejon, F., Civit-Balcells, A., de Vries, H.: Analyzing recommender systems for health promotion using a multidisciplinary taxonomy: a scoping review. Int. J. Med. Inform. **114**, 143–155 (2018). https://doi.org/10.1016/j.ijmedinf.2017.12.018

IOM: Institute of Medicine. To Err Is Human: Building a Safer Health System. In: To Err Is Human: Building a Safer Health System., p 8, (1999) https://doi.org/10.1017/S095026880100509X

Kaplan, R.M., Frosch, D.L.: Decision making in medicine and health care. Ann. Rev. Clin. Psychol. **1**(1), 525–556 (2005). https://doi.org/10.1146/annurev.clinpsy.1.102803.144118

Kawamoto, K.: Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ (2005). https://doi.org/10.1016/S0022-0728(02)01149-X

Kira, K., Rendell, L.A.: A Practical Approach to Feature Selection. In: International Conference on Machine Learning, pp 249–256 (1992)

Komkhao, M., Lu, J., Zhang, L.: Determining Pattern Similarity in a Medical Recommender System. pp 103–114 (2012)

Kononenko, I., Šimec, E., Robnik-Šikonja, M.: Overcoming the myopia of inductive learning algorithms with RELIEFF. Appl. Intell. **7**(1), 39–55 (1997). https://doi.org/10.1023/A:1008280620621

Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **42**, 42–49 (2009)

Larkin, I., Ang, D., Steinhart, J., Chao, M., Patterson, M., Sah, S., Wu, T., Schoenbaum, M., Hutchins, D., Brennan, T., Loewenstein, G.: Association between academic medical center pharmaceutical detailing policies and physician prescribing. JAMA J. Am. Med. Assoc. (2017). https://doi.org/10.1001/jama.2017.4039

Lattar, H., Ben Salem, A., Hajjami Ben Ghézala, H., Boufares, F.: Health Recommender Systems: A Survey. (2020) https://doi.org/10.1007/978-3-030-21005-2_18

Leeper, N.J., Bauer-Mehren, A., Iyer, S.V., LePendu, P., Olson, C., Shah, N.H.: Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS ONE (2013). https://doi.org/10.1371/journal.pone.0063499

Longhurst, C.A., Harrington, R.A., Shah, N.H.: A "green button" for using aggregate patient data at the point of care. Health Affairs (2014). https://doi.org/10.1377/hlthaff.2014.0099

Lowsky, D.J., Ding, Y., Lee, D.K.K., McCulloch, C.E., Ross, L.F., Thistlethwaite, J.R., Sa, Zenios: A K-nearest neighbors survival probability prediction method. Statist. Med. **32**(12), 2062–2069 (2013). https://doi.org/10.1002/sim.5673

Lu, X., Huang, Z., Duan, H.: Supporting adaptive clinical treatment processes through recommendations. Comp. Methods Prog. Biomed. (2012). https://doi.org/10.1016/j.cmpb.2010.12.005

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T.B., Venkatesh, S., Berk, M.: Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. J. Med. Internet Res. **18**(12), e323 (2016)

Mei, J., Liu, H., Li, X., Xie, G., Yu, Y.: A decision fusion framework for treatment recommendation systems. Stud. Health Technol. Inform. **216**, 300–304 (2015). https://doi.org/10.3233/978-1-61499-564-7-300

Miller, R.A., Pople, H.E., Myers, J.D.: Internist-I, an experimental computer-based diagnostic consultant for general internal medicine. New England J. Med. (1982). https://doi.org/10.1056/NEJM198208193070803

Miyo, K., Nittami, Y.S., Kitagawa, Y., Ohe, K.: Development of case-based medication alerting and recommender system: a new approach to prevention for medication error. Stud. Health Technol. Inform. **129**(Pt 2), 871–874 (2007)

Moreira, G.D.S.P., Jannach, D., Cunha, A.M.D.: Contextual hybrid session-based news recommendation with recurrent neural networks. IEEE Access (2019). https://doi.org/10.1109/ACCESS.2019.2954957

Mustaqeem, A., Anwar, S.M., Majid, M.: A modular cluster based collaborative recommender system for cardiac patients. Artif. Intell. Med. (2020). https://doi.org/10.1016/j.artmed.2019.101761

Nast, A., Amelunxen, L., Augustin, M., Boehncke, W.H., Dressler, C., Gaskins, M., Härle, P., Hoffstadt, B., Klaus, J., Koza, J., Mrowietz, U., Ockenfels, H.M., Philipp, S., Reich, K., Rosenbach, T., Rzany, B., Schlaeger, M., Schmid-Ott, G., Sebastian, M., von Kiedrowski, R., Weberschock, T.: S3 - Leitlinie zur Therapie der Psoriasis vulgaris Update 2017 (2017)

Ning, X., Karypis, G.: SLIM: Sparse LInear Methods for top-N recommender systems. In: Proceedings - IEEE International Conference on Data Mining, ICDM, pp 497–506, (2011) https://doi.org/10.1109/ICDM.2011.134, 0803.0476

Panahiazar, M., Taslimitehrani, V., Pereira, N.L., Pathak, J.: Using EHRs for heart failure therapy recommendation using multidimensional patient similarity analytics. Stud. Health Technol. Inform. **210**, 369–373 (2015). https://doi.org/10.3233/978-1-61499-512-8-369

Pandey, B., Mishra, R.B.: Knowledge and intelligent computing system in medicine. Comp. Biol. Med. **39**(3), 215–30 (2009). https://doi.org/10.1016/j.compbiomed.2008.12.008

Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transact. Pattern Anal. Mach. Intell. **27**(8), 1226–1238 (2005)

Pudil, P.: Pattern recognition letters floating search methods in feature selection NORTH-HOLLAND AMSTERDAM Floating search methods in feature selection. Pattern Recognit. Lett. **15**(1), 1–119 (1994)

Rabiner, L.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition (1989)

Raschka, S.: Model evaluation, model selection, and algorithm selection in machine learning Performance Estimation: Generalization Performance Vs. Model Selection. arXiv **1811**, 12808 (2018)

Ricci, F., Rokach, L., Shapira, B., Kantor, P., Ricci, Francesco, Rokach, Lior, Shapira, Bracha, Kantor, Paul B.: Recommender Systems Handbook. Springer, US, (2011). https://doi.org/10.1007/978-0-387-85820-3

Schäfer, H., Hors-Fraile, S., Karumur, R.P., Calero Valdez, A., Said, A., Torkamaan, H., Ulmer, T., Trattner, C.: (2017) Towards health (aware) recommender systems. In: Proceedings of the 2017 International Conference on Digital Health, Association for Computing Machinery, New York, NY, USA, DH '17, p 157–161, https://doi.org/10.1145/3079452.3079499

Sezgin, E., Özkan, S., (2013) A systematic literature review on Health Recommender Systems. In, : E-Health and Bioengineering Conference. EHB **2013**,(2013). https://doi.org/10.1109/EHB.2013.6707249

Sherman, R.E., Anderson, S.A., Dal Pan, G.J., Gray, G.W., Gross, T., Hunter, N.L., LaVange, L., Marinac-Dabic, D., Marks, P.W., Robb, M.A., Shuren, J., Temple, R., Woodcock, J., Yue, L.Q., Califf, R.M.: Real-world evidence - what is it and what can it tell us? The New England J. Med. **23**(375), 2293–2297 (2016). https://doi.org/10.1056/NEJMsb1609216

Shortliffe, E.: Computer-based medical consultations: MYCIN, vol. 2. Elsevier (2012)

Shortliffe, E.H. (1987), Computer Programs to Support Clinical Decision Making. JAMA: The Journal of the American Medical Association. https://doi.org/10.1001/jama.1987.03400010065029

Sim, I.E.A.: Clinical decision support systems for the practice of evidence-based medicine. JAMIA J. Am. Med. Inform. Assoc. **8**(6), 527 (2001)

Sodsee, S., Komkhao, M.: Evidence-based Medical Recommender Systems: A. Review. **4**, 114–120 (2013)

Sönnichsen, A., Trampisch, U.S., Rieckert, A., Piccoliori, G., Vögele, A., Flamm, M., Johansson, T., Esmail, A., Reeves, D., Löffler, C., Höck, J., Klaassen-Mielke, R., Trampisch, H.J., Kunnamo, I.: Polypharmacy in chronic diseases-reduction of inappropriate medication and adverse drug events in older populations by electronic decision support (PRIMA-eDS): study protocol for a randomized controlled trial. Trials (2016). https://doi.org/10.1186/s13063-016-1177-8

Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. Adv. Artif. Intell. (2009). https://doi.org/10.1155/2009/421425

Sutton, R.T., Pincock, D., Baumgart, D.C., Sadowski, D.C., Fedorak, R.N., Kroeker, K.I. (2020), An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Medicine, https://doi.org/10.1038/s41746-020-0221-y

Symeonidis, P., Kirjackaja, L., Zanker, M.: Session-aware news recommendations using random walks on time-evolving heterogeneous information networks. User Model. User Adapt. Interact. (2020). https://doi.org/10.1007/s11257-020-09261-9

Tintarev, N., Masthoff, J.: A survey of explanations in recommender systems. Proceedings - International Conference on Data Engineering pp 801–810, (2007) https://doi.org/10.1109/ICDEW.2007.4401070

Trimble, M., Hamilton, P.: The thinking doctor: Clinical Decision making in contemporary medicine. (2016). https://doi.org/10.7861/clinmedicine.16-4-343

Valdez, C., Ziefle, M., Verbert, K.: Machine Learning for Health Informatics **9605**,(2016). https://doi.org/10.1007/978-3-319-50478-0

Weinberger, K., Blitzer, J., Kl, S.: Distance metric learning for large margin nearest neighbor classification. Nips **10**, 207–244 (2005)

Wiesner, M., Pfeifer, D.: Health recommender systems: concepts, requirements, technical basics and challenges. Int. J. Environ. Res. Public Health **11**(3), 2580–2607 (2014). https://doi.org/10.3390/ijerph110302580

Wilcoxon, F.: Individual comparisons by ranking methods. Biometr. Bullet. (1945). https://doi.org/10.2307/3001968

Wright, A., Sittig, D.F., Ash, J.S., Feblowitz, J., Meltzer, S., McMullen, C., Guappone, K., Carpenter, J., Richardson, J., Simonaitis, L., Evans, R.S., Nichol, W.P., Middleton, B.: Development and evaluation of a comprehensive clinical decision support taxonomy: comparison of front-end tools in commercial and internally developed electronic health record systems. J. Am. Med. Inform. Assoc. (2011). https://doi.org/10.1136/amiajnl-2011-000113

Yu, L., Liu, H.: Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. International Conference on Machine Learning (ICML) pp 1–8, citeulike-article-id:3398512 (2003)

Zhang, Q., Zhang, G., Lu, J., Wu, D.: A framework of hybrid recommender system for personalized clinical prescription. In: Proceedings - The 2015 10th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2015, (2016a) https://doi.org/10.1109/ISKE.2015.98

Zhang, W., Zou, H., Luo, L., Liu, Q., Wu, W., Xiao, W.: Predicting potential side effects of drugs by recommender methods and ensemble learning. Neurocomputing (2016b). https://doi.org/10.1016/j.neucom.2015.08.054

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Felix Gräßer received a diploma degree from the faculty of electrical engineering at TU Dresden (TUD), Germany in 2012. Currently, he is research associate and Ph.D. candidate at the Institute of Biomedical Engineering at TUD. His research interests cover health recommender systems, machine learning for health applications and clinical decision support systems.

Falko Tesch received a B.A. in Social Science from the University of Rostock in 2009 and a M.Sc. in Survey Statistics from Otto Friedrich University Bamberg in 2014. He was working at the Max Planck Institute for Demographic Research in Rostock and the professorship of Demography at the Otto Friedrich University of Bamberg. Since 2015, he is working as a research associate at the Center for Evidence-Based Healthcare of the University Hospital and Faculty of Medicine Carl Gustav Carus at TU Dresden. His research interests cover epidemiology and medical record data.

Jochen Schmitt MD, MPH, is a professor for health services research and social medicine at the Medical Faculty Carl Gustav Carus, TU Dresden, Germany. Prof. Schmitt graduated in Medicine in Leipzig, Germany in 2002 and received an MPH degree from the Johns Hopkins Bloomberg School of Public Health in Baltime, USA, in 2004. He then worked as a dermatologist and clinical researcher before he was appointed as the director of the Center for Evidence-based Healthcare at the TU Dresden in 2012. The Center currently employs about 50 researchers and focusses on outcomes research, implementation research, effectiveness studies, as well as evidence synthesis and evidence communication. Prof. Schmitt led and coordinated several national and international multi-center studies and published multiple papers in peer review journals.

Susanne Abraham MD, is a senior physician at the department of Dermatology, University Allergy Center, Medical Faculty Carl Gustav Carus, TU Dresden, Germany. She graduated in medicine in Dresden in 2003 and received a M.D. degree on her work on the metabolism of vitamin D3 in human keratinocytes in the year 2006. She specialized in dermatology in 2007 and in allergology in 2012 and received a scholarship for paediatric dermatology in the year 2015. She is principal investigator and deputy in several clinical studies, her research interests are chronic inflammatory skin diseases.

Hagen Malberg received the Ph.D. degree in 1999. From 1999 to 2010, he was the Head of Biosignal Processing Group (Karlsruhe Research Center and KIT Karlsruhe Institute of Technology). Since 2010, he is Chair of Biomedical Engineering, Technische Universität Dresden, Dresden, Germany. His research interests include biosignal analysis, particularly related to the regulation of the cardiovascular system.

Sebastian Zaunseder received the Ph.D. degree in Electrical Engineering from TU Dresden (TUD), Germany in 2011. Subsequently, he joined the Institute of Biomedical Engineering of TUD where he established the group Medical Sensing and Signal Processing. Since 2019, he is professor for Medical Engineering at FH Dortmund. His research interests cover novel sensing modalities and data processing including signal and image processing as well as machine learning.

Authors and Affiliations

Felix Gräßer[1] · **Falko Tesch**[2] · **Jochen Schmitt**[2] · **Susanne Abraham**[3] · **Hagen Malberg**[1] · **Sebastian Zaunseder**[4]

Sebastian Zaunseder
sebastian.zaunseder@fh-dortmund.de

[1] Institute of Biomedical Engineering, TU Dresden, Dresden, Germany

[2] Center for Evidence-Based Healthcare, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany

[3] Clinic and Polyclinic for Dermatology, University Hospital Carl Gustav Carus, TU Dresden, Dresden, Germany

[4] Department of Information Technology, FH Dortmund, Dortmund, Germany