TECHNISCHE UNIVERSITÄT DRESDEN

Data-based Therapy Recommender Systems

Dipl.-Ing.

Felix Magnus Gräßer

von der Fakultät Elektrotechnik und Informationstechnik der Technischen Universität Dresden

zur Erlangung des akademischen Grades

Doktoringenieur
(Dr.-Ing.)

genehmigte Dissertation

Vorsitzender:	Prof. Dr.-Ing. habil. Mehmet Ercan Altinsoy (TU Dresden)
Gutachter:	Prof. Dr.-Ing. habil. Hagen Malberg (TU Dresden)
Gutachter:	Prof. Dr. Thomas Neumuth (Universität Leipzig)
Gutachter:	Prof. Dr.-Ing. habil. Sebastian Zaunseder (FH Dortmund)

Tag der Einreichung:	29. September 2020
Tag der Verteidigung:	18. Juni 2021

Declaration

I hereby declare that this dissertation is the product of my original work and contains nothing which is the result of work carried out in collaboration with others except as declared in the acknowledgment and specified in the text. Thoughts taken directly or indirectly from external sources are acknowledged as references. I further state that no part of my thesis has already been or will be submitted to another university with the purpose of earning another degree or qualification.

Dresden, September 29, 2020 Felix Gräßer

Kurzfassung

Für viele Krankheitsbilder und Indikationen ist ein breites Spektrum an Arzneimitteln und Arzneimittelkombinationen verfügbar. Darüber hinaus stellen Therapieziele oft Kompromisse zwischen medizinischen Zielstellungen und Präferenzen und Erwartungen von Patienten dar, um Zufriedenheit und Adhärenz zu gewährleisten. Die Auswahl der optimalen Therapieoption kann daher eine große Herausforderung für den behandelnden Arzt darstellen. Klinische Entscheidungsunterstützungssysteme, die Wirksamkeit oder Risiken unerwünschter Arzneimittelwirkung für Behandlungsoptionen vorhersagen, können diesen Entscheidungsprozess unterstützen und Leitlinien-basierte Empfehlungen ergänzen, wenn Leitlinien oder wissenschaftliche Literatur fehlen oder ungeeignet sind. Bis heute sind keine derartigen Systeme verfügbar. Im Rahmen dieser Arbeit wird die Anwendung von Methoden aus der Domäne der Recommender Systems (RS) und des Maschinellen Lernens (ML) in solchen Unterstützungssystemen untersucht.
Aufgrund ihres erfolgreichen Einsatzes in anderen Empfehlungssystemen und der einfachen Interpretierbarkeit werden zum einen Nachbarschafts-basierte Collaborative Filter (CF) an die besonderen Anforderungen und Herausforderungen der Therapieempfehlung angepasst. Zum anderen werden ein Modell-basierter CF-Ansatz (SLIM) und ein ML Algorithmus (GBM) erprobt. Alle genannten Ansätze werden anhand eines exemplarischen Therapieempfehlungssystems evaluiert, das auf die Behandlung der Autoimmunkrankheit Psoriasis abzielt. Um das Risiko der Empfehlung kontraindizierter oder gar gesundheitsgefährdender Medikamente zu reduzieren, werden Regeln aus evidenzbasierten Leitlinien und Expertenempfehlungen implementiert, um solche Therapieoptionen aus den Empfehlungslisten herauszufiltern.
Insbesondere die Nachbarschafts-basierten CF-Algorithmen zeigen insgesamt kleine durchschnittliche Abweichungen zwischen geschätztem und tatsächlichem Therapie-Outcome. Auch die aus den Outcome-Schätzungen abgeleiteten Empfehlungen zeigen eine hohe Übereinstimmung mit der tatsächlich angewandten Behandlung. Die Modell-basierten Ansätze sind den Nachbarschafts-basierten Ansätzen insgesamt unterlegen, was auf den begrenzten Umfang der verfügbaren Trainingsdaten zurückzuführen ist und die Generalisierungsfähigkeit der Modelle erschwert. Im Vergleich mit menschlichen Experten sind alle untersuchten Algorithmen jedoch hinsichtlich Übereinstimmung mit der tatsächlich angewandten Therapie unterlegen.
Eine objektive und effiziente Bewertung des Behandlungserfolgs kann als Voraussetzung für ein erfolgreiches "Krankheitsmanagement" angesehen werden. Daher wird in weiteren Untersuchungen für ausgwählten klinische Anwendungen der Einsatz von ML Methoden zur automatischen Quantifizierung von Gesunheitszustand und Therapie-Outcome erprobt. Zusätzlich, als weitere Quelle für Informationen über Therapiewirksamkeiten, wird der Einsatz von *Sentiment Analysis* Methoden zur Extraktion solcher Informationen aus Medikamenten-Bewertungen untersucht.

Abstract

Under most medical conditions and indications, a great variety of pharmaceutical drugs and drug combinations are available. Beyond that, trade-offs need to be found between the medical requirements and the patients' preferences and expectations in order to support patients' satisfaction and adherence to treatments. As a consequence, the selection of an optimal therapy option for an individual patient poses a challenging task to prescribers. Clinical Decision Support Systems (CDSSs), which predict outcome as effectiveness and risk of adverse effects for available treatment options, can support this decision-making process and complement guideline-based decision-making where evidence from scientific literature is missing or inappropriate. To date, no such systems are available. Within this work, the application of methods from the Recommender Systems (RS) domain and Machine Learning (ML) in such decision support systems is studied.

Due to their successful application in other recommender systems and good interpretability, neighborhood-based CF algorithms are transferred to the medical domain and are adapted to meet the requirements and challenges of the therapy recommendation task. Moreover, a model-based CF method (SLIM) and a state of the art ML algorithm (GBM) are employed. All algorithms are evaluated in an exemplary therapy recommender system, targeting the treatment of the autoimmune skin disease Psoriasis. In order to reduce the risk of recommending contraindicated or even health-endangering drugs, rules derived from evidence-based guidelines and expert recommendations are implemented to filter such options from the recommendation lists.

Especially the neighborhood-based CF algorithms show small average errors between estimated and observed outcome. Also, the recommendations derived from outcome estimates show high agreement with the ground truth. The performance of both model-based approaches is inferior to the neighborhood-based recommender. This is primarily assumed to be due to the limited training data sizes, which renders generalizability of the learned models difficult. Compared with recommendations provided by various experts, all proposed approaches are, however, inferior in terms of agreement with the ground truth.

An objective and efficient assessment of treatment response can be regarded a prerequisite for successful "disease management". Therefore, the use of ML methods for the automatic quantification of health status and therapy outcome for selected clinical applications is investigated in further experiments. Moreover, as additional source of information about drug effectiveness, the use of *Sentiment Analysis*, in order to extract such information from drug reviews, is investigated.

Curriculum Vitae

Felix Magnus Gräßer was born on September 12, 1984 in Heidenheim a. d. Brenz. From 1993 to 2004 he attended the Freie Waldorfschule Heidenheim and from 2004 to 2005 the Schillergymnasium Heidenheim. From 2005 to 2006 he provided the obligatory civil service at Haus Freudenberg in Starnberg.

From 2006 he studied mechatronics with a focus on micromechatronics and biomedical engineering at the Technical University of Dresden. Between 2009 and 2010 he spent one year at the Royal Institute of Technology in Stockholm, Sweden. He continued his studies at Dresden University of Technology from 2010 and graduated in 2012.

Since 2012 he was employed at Linguwerk GmbH in Dresden as a development engineer until 2013. From 2013 to 2015 he worked full-time as development engineer at Sonovum AG, Leipzig, where he was responsible for software development and regulatory affairs related to medical software. He continued his work at Sonovum part-time until 2018.

In 2015, he joined the Institute of Biomedical Engineering (IBMT) at the Technical University Dresden as research associate. Since 2019 he leads a research group in the field of machine learning in biomedical engineering. At the IBMT he completed his dissertation by 2020.

Acknowledgements

I would like to thank the *Roland Ernst Stiftung* and the *Sächsisches Staatsministerium für Wissenschaft und Kunst* for funding the projects "Therapieempfehlungssystem" and "Data-based Therapy Recommender System (DARE)", respectively, which form the basis of this thesis.

Moreover, I thank the IBMT, namely Prof. Malberg and Prof. Zaunseder, for making this work possible. I especially thank Prof. Malberg for his supervision, support and guidance. Prof. Zaunseder I thank for the fruitful discussions, the intensive cooperation and his inspiration. My thanks also go to all colleagues and doctoral fellow students at the IBMT for the pleasant working atmosphere and the friendships that have grown. Last but not least, I would also like to thank all students who contributed to this work.

Special thanks also go to Prof. Schmitt and his team from the *Zentrum für Evidenz-basierte Gesundheitsversorgung* for the productive collaboration and the valuable input. Dr. Abraham from the *Klinik and Poliklinik für Dermatologie* I am grateful not only for providing the necessary data, but also the essential medical knowledge.

Finally, I would like to thank my family and friends. Particularly, I express my great gratitude to my parents, my sister and my brother for their unconditional support and their encouragement. But most of all I thank Sara for her endless patience, understanding, support and for spending her live with me. To finish, I thank my little daughter Maja for the joy she brought into my life.

Contents

List of Figures

List of Tables

Abbreviations and Symbols

List of Abbreviations

ADE Adverse Drug Event
ALS Alternating Least Squares
ANN Artificial Neural Network
ANS Autonomous Nervous System
AP Average Precision
CARS Context-aware Recommender System
CART Classification and Regression Tree
CBR Case-based Reasoning
CDSS Clinical Decision Support System
CF Collaborative Filter
CNN Convolutional Neural Network
CPOE Computerized Physician Order Entry
DLQI Dermatology Life Quality Index
DT Decision Tree
EbM Evidence-based Medicine
ECG Electrocardiogram
EEG Electroencephalogram
EHR Electronic Health Record
EOG Electrooculogram
EMG Electromyogram
ERD Entity Relationship Diagram
FNN Feedforward Neural Network
FWER Family-wise Error Rate
GBM Gradient Boosting Machine
GD Gradient Decent
GRU Gated Recurrent Unit
GUI Graphical User Interface
HEOM Heterogeneous Euclidean Overlap Metric
HIS Hospital Information System
HMM Hidden Markov Model
HR Heart Rate
HRS Health Recommender System
HRV Heart Rate Variability

ICD International Classification of Diseases
IR Information Retrieval
KNN K-Nearest-Neighbor
LDA Linear Discriminant Analysis
LIMS Laboratory Information Management System
LMNN Large Margin Nearest Neighbor
LOCF Last Observation Carried Forward
LogR Logistic Regression
LSI Latent Semantic Indexing
LSML Locally Supervised Metric Learning
LSTM Long Short-Term Memory
LR Linear Regression
MAP Mean Average Precision
MAR Missing At Random
MCAR Missing Completely At Random
MF Matrix Factorization
ML Machine Learning
MLP Multi Layer Perceptron
MSE Mean Squared Error
MTX Methotrexate
NLP Natural Language Processing
NMAR Not Missing At Random
NMF Non-negative Matrix Factorization
NB Naive Bayes Classifier
PASI Psoriasis Area and Severity Index
PCA Principal Component Analyis
PD Parkinson's Disease
PDF Probability Density Function
PSG Polysomnography
PUVA Psoralen and Ultraviolet A Light
RBA Relief-based Algorithm
RBF Gaussian Radial Basis Function
RCT Randomized Controlled Trial
ReLU Rectified Linear Unit
RF Random Forest
RMSE Root Mean Square Error
RNN Recurrent Neural Network
RS Recommender System
SGD Stochastic Gradient Decent
SLIM Sparse Linear Method
SMC Simple Matching Similarity Coefficient
SVD Singular Value Decomposition
SVM Support Vector Maschine
UV Ultra Violet

Mathematical Symbols

General Symbols

i, j, k, l, m, n	Index variables
$L(\cdot)$	Objective function
M	Number of attributes in $\mathbf{X}$
N	Number of instances in $\mathbf{X}$
p	Order of p-Norm
$R(\cdot)$	Regularization term
$\mathbb{R}$	Set of real numbers
r_s	Spearman's rank correlation coefficient
$\mathbf{X}$	Data matrix
$\mathbf{x}$	Element in $\mathbf{X}$
$\hat{\mathbf{x}}$	Preprocessed element in $\mathbf{X}$
y	Target label
$\hat{y}$	Prediction of y
$\hat{\mathbf{y}}$	Vector of predictions
β, λ	Regularization parameter
μ	Learning rate
σ	Standard deviation
$\|\| \cdot \|\|_p$	L_p-norm

Psoriasis Data

$\mathbf{A}, \mathbf{A}', \mathbf{A}''$	Previous outcome matrix, imputation stage 1, imputation stage 2
$\mathbf{A}^p$	Treatment history attributes of patient p
$\mathbf{a}_n^p$	Treatment history attributes of patient p in consultation n
$a_{n,m}$	*Affinity* score for consultation n and therapy option m
$\tilde{\mathbf{A}}^{all}$	Complete consultation-therapy outcome matrix
$\tilde{\mathbf{A}}^{hist}$	Historic consultation-therapy outcome matrix
$\tilde{\mathbf{A}}_{test}$	Test subset of $\tilde{\mathbf{A}}^{hist}$
$\tilde{\mathbf{A}}_{train}$	Training subset of $\tilde{\mathbf{A}}^{all}$
$\tilde{\mathbf{a}}^{all}$	Vector (row) in $\tilde{\mathbf{A}}^{all}$ associated with one consultation
$\tilde{\mathbf{a}}^{hist}$	Vector (row) in $\tilde{\mathbf{A}}^{hist}$ associated with one consultation
$\tilde{\mathbf{a}}^k$	Vector (row) in $\tilde{\mathbf{A}}_{train}$ representing a training consultation k
$\tilde{\mathbf{a}}_{test}^n$	Vector (row) in $\tilde{\mathbf{A}}_{test}$ representing a test consultation n
$\tilde{a}_m^k$	Element in $\tilde{\mathbf{A}}_{train}$ representing a training consultation k and therapy m
D	Number of attributes in $\tilde{\mathbf{X}}$
d	Attribute in $\tilde{\mathbf{X}}$
$f_{i,n,m}$	Outcome indicator i for consultation n and therapy option m
k, n	Consultation

M	Number of therapy options
M^{bio}	Number of biopharmaceutical drugs
M^{combi}	Number of combinations of pharmaceutical drugs
M^{conv}	Number of conventional pharmaceutical drugs
M^{other}	Other not specified systemic treatments
M^{UV}	UV therapies
m	Therapy option
N	Total number of consultations
N_p	Number of consultations of patient p
N^p_{train}	Number of consultations in $\tilde{\mathbf{X}}^p_{train}$
P	Number of patients
p	Patient
thr_{good}	*Affinity* threshold
w_i	Weight of outcome indicator i
$\mathbf{X}, \mathbf{X}', \mathbf{X}''$	Data matrix, imputation stage 1, imputation stage 2
$\mathbf{X}^p$	Patient describing attributes of patient p
$\mathbf{x}^p_n$	Patient describing attributes of patient p in consultation n
$\tilde{\mathbf{X}}^p_{test}$	Subset of $\tilde{\mathbf{X}}$ holding all consultations of patient p
$\tilde{\mathbf{X}}^p_{train}$	Subset of $\tilde{\mathbf{X}}$ holding no consultations of patient p
$\tilde{\mathbf{X}}^{p,i}_{test}$	Subset of $\tilde{\mathbf{X}}^p_{train}$ with the test partition of iteration i
$\tilde{\mathbf{X}}^{p,i}_{train}$	Subset of $\tilde{\mathbf{X}}^p_{train}$ with the training partition of iteration i
$\tilde{\mathbf{X}}^{j,hits}_{train}$	Subset of $\tilde{\mathbf{X}}^p_{train}$ with the nearest hits of consultation j
$\tilde{\mathbf{X}}^{j,misses}_{train}$	Subset of $\tilde{\mathbf{X}}^p_{train}$ with the nearest misses of consultation j
$\tilde{\mathbf{x}}$	Vector (row) in $\tilde{\mathbf{X}}$ representing a consultation
$\tilde{\mathbf{x}}^n_{test}$	Vector (row) in $\tilde{\mathbf{X}}_{test}$ representing consultation n
$\mathbf{Y}$	Outcome matrix, imputation stage 1, imputation stage 2
$\mathbf{Y}^p$	Treatment describing attributes of patient p
$\mathbf{y}^p_n$	Treatment describing attributes of patient p in consultation n
$\tilde{\mathbf{Y}}$	Consultation outcome matrix
$\tilde{\mathbf{Y}}_{test}$	Test subset of $\tilde{\mathbf{Y}}$
$\tilde{\mathbf{Y}}_{train}$	Training subset of $\tilde{\mathbf{Y}}$
$\tilde{\mathbf{Y}}^p_{train}$	Subset of $\tilde{\mathbf{Y}}$ holding no consultations of patient p
$\tilde{\mathbf{y}}^n_{test}$	Vector (row) in $\tilde{\mathbf{Y}}_{test}$ representing outcomes of consultation n
$\hat{y}^n_m$	Outcome prediction of therapy option m and consultation n

Patient Similarity

$\mathcal{C}_i$	Local compactness
$d(\cdot)$	General distance function
$d_{Cheb}(\cdot)$	Chebyshew metric
$d_{Euc}(\cdot)$	Euclidean metric
$d_{Hamm}(\cdot)$	Hamming distance

$d_M(\cdot)$	(Generalized) Mahalanobis metric
$d_{Man}(\cdot)$	Manhatten metric
$d_{Mink}(\cdot)$	Minkowski metric
$\mathcal{D}$	Set of dissimilar data points
K_σ	Gaussian kernel function
l,m,u	Metric learning objective function constraints
$\mathbf{M}$	Transformation matrix
$\mathcal{N}_i^o$	Homogeneous neighborhood of data point $\mathbf{x}_i$
$\mathcal{N}_i^e$	Heterogeneous neighborhood of data point $\mathbf{x}_i$
p	Order of *Minkoswki metric*
$\mathcal{R}$	Set of relative distant data points
$s(\cdot)$	General similarity function
$s^{n,k}$	Similarity between consultations n and k
$s_{Corr}(\cdot)$	Pearson correlation function
$s_{Cos}(\cdot)$	Cosine similarity
$s_{GSC}(\cdot)$	Gower similarity function
$s_{RBF}(\cdot)$	Gaussian radial basis function
$s_{SMC}(\cdot)$	Simple matching similarity coefficient
$\mathcal{S}$	Set of similar data points
$\mathcal{S}_i$	Local scatterness
thr_s	Similarity threshold
w_m	Weight of attribute m
δ_{ijm}	Availability of attribute m
ρ_{ijm}	Data-type specific similarity measure of attribute m
σ	Gaussian radial basis function spread parameter

Recommender Systems

$\mathcal{I}$	Set of items
$\mathcal{I}_{test}$	Set of test items
$\mathcal{I}_u$	Set of items rated by user u
$\mathcal{I}_{uv}$	Set of items mutually rated by users u and v
K	Number of nearest neighbors
k	Rank of $\hat{\mathbf{R}}$
M	Number of items
N	Number of users
N_i	Number of users which have given feedback on item i
n	Rank of $\mathbf{R}$
$\mathcal{N}_i(u)$	Neighborhood of u in which item i is rated
$\mathcal{N}_u(i)$	Neighborhood of i which is rated by user u
$\mathbf{P}$	User latent factor matrix
$\mathbf{p}_u$	Latent factors of user u

$\mathbf{Q}$	Item latent factor matrix
$\mathbf{q}_i$	Latent factors of item i
$\mathbf{R}$	User-item feedback matrix
$\mathbf{R}_{test}$	Test user-item feedback matrix
$\mathbf{r}_i$	Vector (column) in $\mathbf{R}$ with ratings on item i
$\mathbf{r}_u$	Vector (row) in $\mathbf{R}$ with ratings of user u
r_{ui}	Observed ratings user u on item i
$\hat{\mathbf{R}}$	Approximation of $\mathbf{R}$
$\hat{r}_{ui}$	Predicted ratings of user u on item i
$\bar{r}_i$	Average rating given to item i
$\bar{r}_u, \bar{r}_v$	Average rating of users u and v
$\mathbf{S}^*$	Aggregation coefficient matrix
$\mathbf{s}_i^*$	Vector (column) in $\mathbf{S}^*$ with aggregation coefficient of one item i
$\mathbf{s}_u^*$	Vector (row) of optimized similarities of user u
s_{ij}	Similarity between items i and j, i.e. item-based similarity coefficient
s_{ij}^*	Optimized item-based similarity coefficient
s_{uv}	Similarity between users u and v, i.e. user-based similarity coefficient
s_{uv}^*	Optimized user-based similarity coefficient
$\mathbf{U}$, $\mathbf{V}$	Matrix of left and right singular vectors of $\mathbf{R}$
u,v	User
$\mathcal{U}$	Set of users
$\mathcal{U}_{test}$	Set of test users
α	Case amplification coefficient
$\mathbf{\Sigma}$	Diagonal matrix of ordered singular values of $\mathbf{R}$
σ_i	Standard deviation of ratings on item i and j
σ_u	Standard deviation of ratings of user u and v

Relief-based Algorithm

J	Number of iterations
K_{RBA}	Number of nearest hits and nearest misses
thr_w	Attribute relevance threshold
$\mathbf{w}$	Attribute weight vector
$\mathbf{w}^{init}$	Initial attribute weight vector
w_d	Weight of attribute d
ρ_d	Data-type specific similarity measure of attribute d
$\bar{\rho}_d^{hits}$	Average distance of nearest hits of attribute d
$\bar{\rho}_d^{misses}$	Average distance of nearest misses of attribute d

Large Margin Nearest Neighbors Algorithms

$d_L(\cdot)$, $d_M(\cdot)$	Euclidean metric in transformed attribute space

K_{LMNN}	Number of included neighbors (*target neighbors* and *impostors*)
$\mathbf{L}$	Linear transformation matrix
$\mathbf{M}$	Square transformation matrix $\mathbf{M} = \mathbf{L}^T\mathbf{L}$
$\mathbf{y}^0$	Binary matrix defining matching labels
ϵ_{jkl}	Slack variable
ϵ_{pull}	Pulling term
ϵ_{push}	Pushing term
η_{nk}	Binary matrix defining neighborhoods
ν	ϵ_{pull} and ϵ_{pull} impact parameter

Decision Trees and Decision Tree Ensembles

A	Attribute in $\mathcal{S}s$
$\mathbf{D}_m$	Distribution of training samples $\mathcal{S}$ in iteration m
d_{max}	Maximum tree depth of Decision Tree
$F(\mathbf{x})$	Ensemble model
f	Fraction of training samples $\mathcal{S}$
$Gini(\mathcal{S}_i)$	Gini index of $\mathcal{S}_i$
$GiniGain(\mathcal{S}_i, A)$	Gini index reduction by splitting $\mathcal{S}_i$ by A
$H(\mathcal{S})$	Shannon entropy of $\mathcal{S}$
$h_m(\cdot)$	Base learner m
$IG(\mathcal{S}_i, A)$	Information gain by splitting $\mathcal{S}_i$ by A
$IGR(\mathcal{S}_i, A)$	Information gain ratio of splitting $\mathcal{S}_i$ by A
i	Decision tree node
J	Nominal categories of an attribute A
j	Nominal category in J
M	Number of base learner
m	Index of base learner
N	Number of observations in $\mathcal{S}$
n_{trees}	Number of Decision Trees in ensemble
r_m	(Pseudo-)residual of base learner m
$\mathcal{S}$	Set of training samples
$\mathcal{S}_i$	Set of training samples reaching node i
$\mathcal{S}_i^j$	Set of training samples reaching node i with nominal category j
p	Fraction of attributes from A
w_{child}	Minimum number of instance weights
w_m	Weight of base learner m
$\hat{y}_m$	Prediction of y of base learner m
β_m	*AdaBoost* weight coefficient of base learner m
γ_m	*Gradient Boosted Machine* weight coefficient of base learner m
ϵ_m	Error of base learner m

Hidden Markov Models

$\mathbf{A}$	Transition probabilities
$\mathbf{B}$	Emission probabilities
M	Number of features
N	Number of states
O	Observation sequence
S	State sequence
$\mathcal{X}$	State alphabet
$\mathcal{Y}$	Output alphabet
α	Hidden Markov Model
β	Forward variable
λ	Backward variable
$\mathbf{\Pi}$	Initial distribution

Artificial Neural Networks

a_j^l	Activation of neuron j in layer l
$\mathbf{b}^l$	Bias parameters in layer l
$\mathbf{f}$	Feature map
$\mathbf{g}$	Average gradients
$\mathbf{h}_{(t)}$	Recurrent Neural Network state vector at time step t
s	Step size (*stride*)
t	Time step in input sequence
$\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$	Network parameters
$\mathbf{w}$	Filter kernel
z_j^l	Input of neuron j in layer l
δ_j^l	Error fraction of neuron j in layer l
θ	Network parameters
τ	Length of input sequence
σ^l	Activation function of layer l
∇_θ	Partial derivatives with respect to the network parameters θ

Evaluation

FP_u	False positives for user u
N	Number of selected items from recommendation list
n	Position in recommendation list
p_0	Observed agreement
p_e	Expected agreement
TP_u	True positives for user u
$\delta(n)$	Include recommendation at position n into MAP@N computation
κ	*Cohen's Kappa* (Inter-rater reliability)

1 Introduction

1.1 Background and Motivation

The ability to make accurate and timely diagnostic and treatment decisions can be regarded as the core skill and the critical aspect of physician performance in medical practice [74, 139]. The purpose of diagnosis is to classify a patient into a category of patients which are believed to be similar in terms of clinical symptoms. Based on diagnosis and additional patient risk factors, such as demographic data and comorbidities, the attending physician is tasked to make an estimation on the course of the disease and derive management decisions. To do so, natural history of a disease and response to possible treatment options are predicted for a patient [85]. Outcome, however, is typically multifactorial [49], meaning that multiple aspects, such as benefits and harms, are to be considered and also additional factors such as costs and the way of application determine the treatment decision. Hence, the optimal treatment does not only differ among patients due to individual diagnosis and patient characteristics but also due to distinctions in individual patient values and objectives. Precise definition of the targeted outcome [165] and accurate prognosis are the foundation of optimal treatment decisions.

As health outcomes are probabilistic, clinical problem solving is characterized by making judgments and decisions under uncertainty [165]. In spite of still being difficult to trace, it is assumed that there are two distinct cognitive models or processes of clinical reasoning and deciding which are employed individually or complementary: *analytical* and *intuitive* reasoning [18, 74, 139, 358]. This distinction is in line with the dual-process theory which describes two systems of decision-making, the controlled, slow, and conscious *System 2* and the fast, automatic, and non-conscious *System 1* [163]. Analytical reasoning, on the one hand, relies on physiological and pathophysiological knowledge and is closely related to the *hypothetico-deductive* model [358]. This process assumes ideal conditions and tries to remove uncertainty by systematically incorporating all available information. Heuristics, such as rule sets or decision trees, but also probabilistic reasoning can be applied. Bayes rule, for examples, calculates the probability of a disease or outcome by revision of the prevalence of a disease or outcome by including further clinical information. This analytical decision process is regarded as the approach of novices but may also be employed when diagnoses are rare or difficult [74]. Making decision based on analytical reasoning only is most often not realizable in practice as too laborious and time-consuming and due to non-ideal conditions, i.e. the absence of information. Intuitive reasoning, on the other hand, is regarded as the approach of experienced clinicians as it relies heavily on the experience of the decision maker. This intuitive approach is based on recognizing patterns in the available information which are matched against templates to derive decisions (*recognition primed*

decision making [85]). Those templates develop by integrating clinical knowledge and personal experience with similar patients. By relying on instinctive first impressions (*thin-slicing* [9]) and mental shortcuts (e.g. *availability*, *representativeness* and *anchoring heuristic* [361]), intuitive decision-making can be very effective and efficient, but is also susceptible to suffer from risks and cognitive biases [361]. For example, experience is typically biased and especially exceptional cases can result in wrong decisions. Beyond that, incorporation of new diseases or treatment options, new evidence and patient values renders difficult if making decision only based on intuitive reasoning. Finally, the underlying reasoning of intuitive decisions is difficult to explain. As healthcare in general undergoes a shift from paternalistic care to an increasing interest and desire of patients to participate in decisions (*patient empowerment*) [321, 165, 20], explainability becomes an increasingly important factor. Regarding treatment options, physicians not only need to decide on one treatment but will be requested increasingly to clarify decisions and to provide detailed prognoses for the full range of options.

Depending on condition and indication, a great variety of pharmaceutical drugs and drug combinations may be available. Hence, the selection of the potentially most appropriate therapy option for an individual patient may pose a challenging task to prescribers and the reported uncertainties in turn result in treatment deficits [239, 12]. As was shown, making diagnoses but also the choice of treatment is often quiet subjective and underlies impacts such as biases [14, 74, 358] but also conflicts of interests [193] and is prone to errors ("To Err is Human" [175]). In various studies, considerable variability of diagnosis or treatment decisions regarding the same patient could be demonstrated [165]. A phenomenon that affects not only different physicians (*inter-rater reliability*), but also the same physician judging at different times (*intra-rater reliability*). Assuming one optimal treatment for a patient and time, this variance clearly indicates that many patients are not treated optimally. The potential outcome of alternatives is, however, "counterfactual" and unknown. Faulty diagnoses [74] or inappropriate treatment decisions can even cause "iatrogenic" complications such as medication errors and avoidable Adverse Drug Events (ADEs) (see figure 1.1) with, in the worst case, health endangering consequences. In particular, relevant patient characteristics, such as allergies and contraindications, are subject to be overlooked [366], but also polypharmacy is associated with an increased risk for medication errors and ADEs (*drug-drug interactions*) [328, 308]. According to [308], in Germany 57.000 deaths per year are caused by ADEs from which 28.000 are regarded as potentially avoidable. Responsibility for 81 % of the observed medication errors are with the attending physician. However, also the occurrence of less serious ADEs and *side effects* typically reduce adherence, especially if risks were not previously communicated to the patient. All stated concerns in turn increase healthcare costs. In the U.S., the number of deaths due to medication errors is estimated to 250.000 per year and hence is the third-most common cause of death according to [213]. Independent of the caused harm, the average cost of medication errors is estimated to 8.000 U.S.-Dollar [67].

To reduce medication errors and remedy the stated inconsistency of treatment choices, Evidence-based Medicine (EbM) was supposed to supplement a physician's opinion with the best available external evidence from the scientific literature. However, with always including the patient's val-

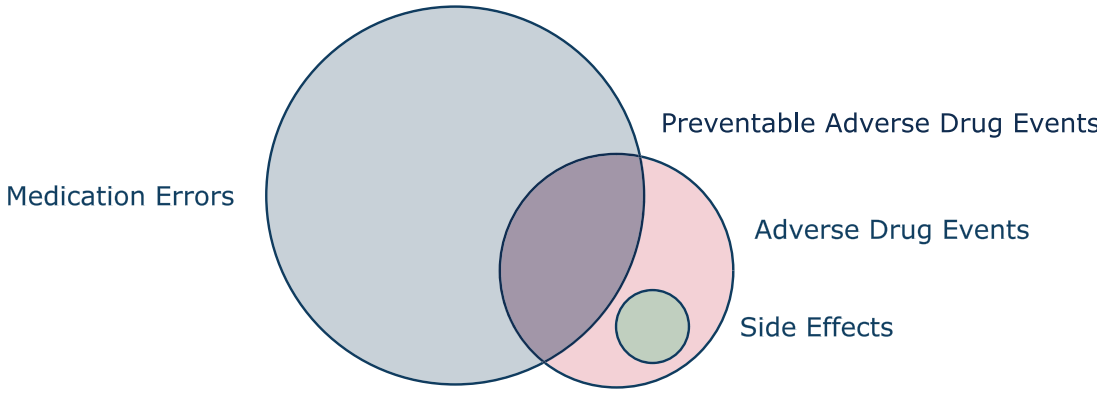


Figure 1.1: Relationship between ADEs, preventable ADEs and medication errors [231]. ADEs judged to be secondary to a main therapeutic effect are termed *side effects*.

ues and objectives into decisions. "EbM is the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients. The practice of EbM requires integration of individual clinical expertise and patient preferences with the best available external clinical evidence from systematic research" [296]. Hence, the initially introduced analytical or intuitive clinical decision-making approaches, based on clinical knowledge only or supplemented by personal experience, are extended by experience from empirical testing [20, 85]. The evidence-based practice comprises (1) the systematical search of the literature targeting a specific and well defined clinical problem or objective, (2) a critical assessment of the best found evidence concerning validity and applicability and (3) finally, estimation of the benefits and harms for the individual patient by incorporating his or her values and objectives.
To evaluate the scientific significance of the clinical studies described in the literature, six levels of evidence can be distinguished according to [220] as listed in table 1.1. The higher the ranking of the evidence class, the broader is its scientific foundation. Studies of class Ia have the highest evidence, whereas studies of class IV have the lowest. However, it must be noted that there is no European or international standard for classification.

Though, two counteracting trends hamper the application of evidence-based practice. On the one hand, the information explosion, i.e. the growing number of clinical studies, which are even often inconsistent and quickly out-dated [315], makes the application of EbM time consuming and laborious. Steadily increasing cost pressure and demographic changes, on the other hand, produce noticeable time constraints in everyday clinical practice. To date, no tools or technical means are available to automate or support the search for and retrieval of the relevant evidence. Evidence-based guidelines summarize the systematical search and assessment of clinical literature concerning a specific condition and potential harms and benefits of a treatment. The intention is to make EbM more readily and time-effectively accessible in clinical practice in order to improve quality and increase transparency. Committees of experts define recommendations for everyday decision-making. Depending on the way how consensus among experts is found, how evidence from published clinical studies is incorporated and the extend of systematics of the guideline development, four guideline levels are defined according to [15] and summarized in table 1.2. In contrast to S1 and S2k guidelines, only S2e and S3 guidelines are based on sys-

Table 1.1: Levels of evidence to assess the scientific significance of clinical studies described in the literature and associated levels of recommendations [220].

Evidence	Recommendation	Description
Class Ia	A	Evidence provided by a systematic review (meta-analyses) from several methodically high-quality RCTs.
Class Ib	A	Evidence provided by at least one sufficiently large and methodically high-quality RCT.
Class IIa	B	Evidence provided by at least one methodically high-quality but not randomized controlled trial, e.g. cohort study.
Class IIb	B	Evidence provided by at least one methodically high-quality trial of another type of quasi-experimental trial.
Class III	B	Evidence provided by one methodically high-quality, non-experimental descriptive study, as e.g. comparative studies, correlation studies or case-control studies.
Class IV	C	Evidence provided by reports of expert committees or expert opinions or clinical experience of acknowledged authorities.

tematic analysis of evidence from scientific literature. The strength of the guideline suggestions, i.e. the level of recommendation (A - shall, B - should, C - can), is directly associated with the evidence class from table 1.1. Guidelines, however, just provide a basic standard intended to give physicians orientation concerning therapy options but no therapy recommendations. Moreover, quality of guidelines varies and high quality guidelines are only available for common conditions. Also regarding easy application and seamless integration into the clinical work process, there are hardly any technical means available to date.

Table 1.2: Depending on the methodology applied for guideline development, four guideline classes are distinguished according to [15].

Class	Description	Method
S3	Evidence- and consensus-based	Representative committee, systematic search, selection, literature appraisal, structured consensus finding.
S2e	Evidence-based	Systematic search, selection, literature appraisal.
S2k	Consensus-based	Representative committee, structured consensus finding.
S1	Recommendation of expert committees	Informal procedure for consensus finding.

EbM and guidelines in general are susceptible to further issues. As individual patients' characteristics typically differ from the strict inclusion criteria which evidence is based on (*patient heterogeneity*), there are always exceptions to guidelines and evidence from literature [108, 112, 251]. For specific patients there might even be only inadequate studies available [53]. Lacking generalizability of clinical trials and especially the presence of multimorbidities can lead

to differing therapy outcome [209, 102] and increases the risk of drug interactions, adverse or unforeseen effects, or contraindications [50, 328]. These 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 [47]. Furthermore, also study endpoints frequently differ from the patient's actual values and objectives. Finally, EbM itself relies on the objectiveness of the available literature. However, it was shown that clinical studies often underlay conflicts of interests and pharmaceutical industry-sponsored and mixed-funding clinical trials are common [45, 251]. Study results are reported selectively [54] and favor specific (the sponsor's) products [30]. Meta-analyses additionally are subject to be influenced by a publication bias which favor studies with significant or positive results [342].

Firstly, to seamlessly integrate the evidence from literature and guidelines into the clinical work process and make them applicable in everyday clinical practice, appropriate technical tools are not yet available. Beyond that, however, the selection of patient-specific therapy options often cannot be provided on the basis of evidence from the literature and guidelines only. An obvious way to address this challenges is to complement this *external evidence* by clinical experience from past patient encounters, which is stored in local or global data bases such as Electronic Health Records (EHRs). Exploiting such *practice-based evidence* [321, 112, 197, 120, 53, 209] facilitates to support the attending physician with empirical experience and supplement external evidence where evidence from literature is missing, inappropriate, or inaccessible. Such data-based approaches can provide an essential basis for personalized therapy recommendations. Due to the large data volume, its high dimensionality and complex interdependencies within the data, however, an efficient integration of the available information cannot be expected from physicians or other health professionals without aids. Therefore, intelligent Clinical Decision Support Systems (CDSSs), which assist with exploiting such data to make treatment decisions, can be expected to play a significant role in future healthcare. Nevertheless, such data-driven CDSSs are rare in clinical practice to date which can be assumed to be closely related to lacking trust and interpretability of the underlying algorithms. Especially Recommender Systems (RSs), which are widely applied in other domains, such as e-commerce or music and movie streaming services, have many obvious analogies with the therapy recommendation setting an can be capable of overcoming such issues. Still, methods from RS research are hardly applied in clinical applications in general or for therapy recommendations in particular.

1.2 Aim of this Work

The overall aim of this work is to provide such a CDSS which supports with the clinical decision-making task by exploiting information stored in routine care data. The goal is to provide *patient-specific* therapy recommendations, i.e. recommendations which are optimized for a given patient and time considering his or her individual characteristics. These therapy recommendations, in turn, are expected to overcome uncertainties among health practitioners and improve patient satisfaction by improving outcome, reducing the risk of ADEs and *side effects* but also by help-

ing to avoid medication errors. A transparent and interpretable presentation of multifactorial outcome predictions, such as potential benefits and harms of therapy options, is supposed to strengthen confidence in such CDSSs and facilitate participatory decision-making. The incorporation of individual patients' values and objectives thus has the potential to increase patient adherence to the recommended treatment options.

Overall, this work can be considered as component of a CDSS as schematized in figure 1.2. This proposed CDSS integrates multiple sources of information such as (collective) clinical experience stored in health records (*practice-based evidence*), clinical evidence from scientific literature (EbM) stored in scientific databases (e.g. Pubmed[1]), information derived from pharmacovigilance (e.g. Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM)[2]), expert information (e.g. Rote Liste[3], Gelbe Liste[4]), or advisory platforms (e.g. Embryotox[5]). 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 as applied in section 7.4.
This vision of a CDSS implements a closed loop in order to feed back treatment decisions and outcome, on the one hand, and information on patient preference, on the other hand. This Interactive Machine Learning (iML) approach [158], encompassing a Doctor-in-the-Loop (DiL), facilitates a continuously learning therapy recommender system. Suchlike, such a system ideally continuously improves by extending the clinical experience databases and adapts to applied research and pharmacovigilance findings.

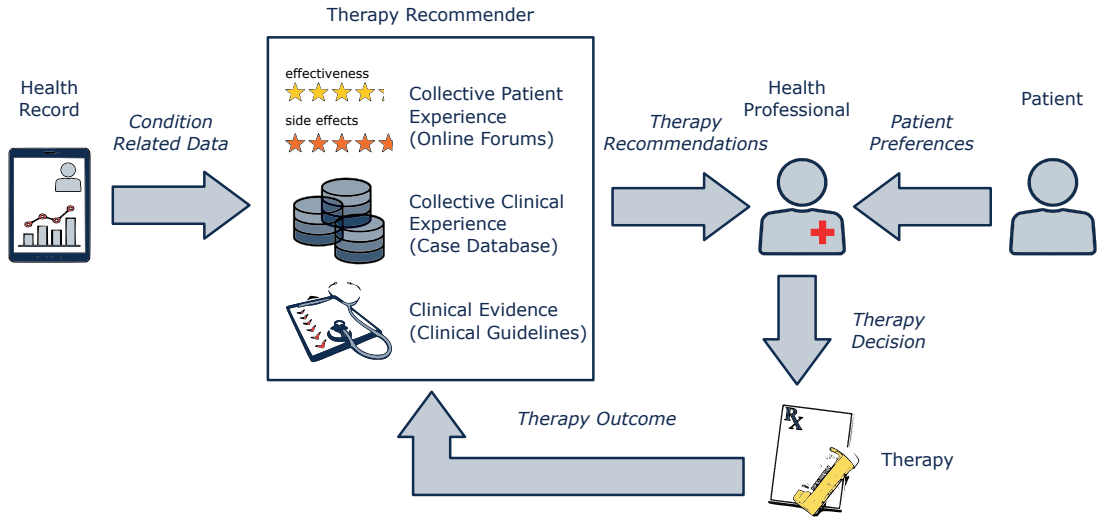


Figure 1.2: Therapy recommender system framework integrating multiple sources of information and encompassing a Doctor-in-the-Loop (DiL).

Within this thesis, particularly the development of a data-driven methodology is targeted, which exploits (phenotypic) patient characteristics and information on outcome of previously applied treatments. This data is considered to capture (collective) clinical experience concerning

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://www.bfarm.de/DE/Arzneimittel/Pharmakovigilanz/_node.html
[3] https://www.rote-liste.de/
[4] https://www.gelbe-liste.de/
[5] https://www.embryotox.de/

therapy options and can be regarded as *practice-based evidence*. On the basis of this clinical experience, the response to available drug options are to be modeled and predicted for a given target patient. These outcome predictions can be finally used to recommend the potentially most effective therapy options to the user with the lowest risk of ADEs. By means of this approach, the intention is to support the attending physician with empirical experience and supplement external evidence. Beyond that, however, such data-based therapy recommendations can also be expected to open up a new chapter of medicine if developed and evaluated on high-quality and sufficiently comprehensive data foundations. In order to promote acceptance among practitioners and to enable the extraction of scientific findings and knowledge, special emphasis is placed on the interpretability and transparency of the methods used.

It is hypothesized that such a data-driven approach is (1) capable of predicting outcome of treatments and (2) to provide appropriate personalized treatment recommendations which are in accordance with successful therapies recommended by the attending physician. Moreover, the hypothesis is made that (3) the proposed approach is, based on the given data, capable of obtaining at least as good *inter-rater reliability* with the attending physician than human experts. In order to assess the validity of the stated hypotheses, three fundamental research questions can be derived:

1. Is it possible, based on patient data and response of previously applied therapies, to predict outcome on therapy options for an individual target patient which are more accurate than the average outcome regarding this treatment option?

2. Is it possible to derive reasonable therapy recommendations from the outcome predictions, which are more in agreement with the effective recommendations of the treating physician than the recommendation of therapies according to their general popularity?

3. Is it possible to derive reasonable therapy recommendations from the outcome predictions, which agree with the effective recommendations of the treating physician to the same extend as the recommendations of human experts?

Two evaluation criteria (*endpoints*) of the experiments to be conducted within the scope of this work can be defined based on the research questions. The primary criterion, associated with the first research question, is the accuracy of outcome predictions. As an accurate outcome prediction is the foundation of appropriate therapy recommendation, primary focus is put on this endpoint in this work. The reference standard of this criterion is the measured outcome of actually applied therapies. The secondary evaluation criterion, related to the second and third research question, is the agreement between recommendations derived from outcome predictions and actually applied therapies, i.e. the recommendations from the attending physician. However, as the objective is rather to recommend potentially successful therapy options than imitating the attending physician, the reference standard for the secondary outcome are recommendations only, which have actually been applied and for which good outcome was observed.

1.3 Dissertation Outline

Primary objective of this work is the development and evaluation of an exemplary therapy recommender system.
Chapter 2 gives a comprehensive overview on the state of the art and the attempt of classification of CDSSs as well as figures concerning acceptance, application and evaluation of such systems. Moreover, in this chapter the results of a systematic literature review on *Therapy Decision Support Systems* in general is presented and related works to RSs in the medical domain, i.e. Health Recommender Systems (HRSs) are discussed.
The exemplarily targeted application of this work is the systemic treatment of the chronic autoimmune skin disease *Psoriasis*. Therefore, routine care data from the *Clinic and Polyclinic for Dermatology, University Hospital Dresden*, is collected, which is considered to represent local clinical experience. In chapter 4, a background on Psoriasis is given, the data extraction and preprocessing procedure is described and some descriptive statistics are given to summarize the available dataset. Preprocessing especially concerns strategies to handle missing values. Moreover, on a subset of the provided data a study on therapy recommendation *inter-rater reliability* is carried out involving dermatologists from different clinics in Germany as experts.
Within this work, various algorithms are developed and evaluated with regard to the above formulated hypotheses. These approaches are detailed in chapter 3 and their implementations described in chapter 5. As Collaborative Filters (CFs) are widely applied in other domains to recommend items, this work transfers methodologies from CF research to the therapy recommendation setting. The stakeholders involved, the definition of preferences and needs to be met, and the data and metrics used to identify homogeneous patient subgroups require modifications of such methods to be applied in the clinical domain. Additionally, also state of the art Machine Learning (ML) algorithms are adapted to the problem at hand to derive therapy recommendations. Finally, the proposed therapy recommender system incorporates an additional post-filtering layer which implements evidence-based and expert-based exclusion rules to reduce the risk of inappropriate or even harmful recommendations as also detailed in chapter 5.
The performance of the proposed algorithms and system variants is evaluated and compared in terms of prediction accuracy and recommendation quality in chapter 6. Moreover, in order to answer the research questions from above, the proposed recommender system's performance is compared with baseline results and with the recommendations of human experts on a subset of the available data.
Chapter 7 summarizes additional own studies, which intend to path the way for further applications and extensions of the proposed therapy recommender system approach. On the one hand, quantification of health status and outcome based on raw vital signs for various conditions is studied. On the other hand, sentiment analysis methods are applied to patient reviews to automatically assess experience with applied treatments.
In chapter 8, finally, the demonstrated results are discussed and practical recommendations for further applications are derived. Moreover, the overall work is summarized, limitations are named and potential future works and extensions derived.

2 State of the Art

In the following chapter, state of the art and related works are analyzed. Initially, an overview on definitions, background and scientific directions related to CDSSs in general are given. Hereafter, the findings of a systematic literature review, to identify related works dealing with CDSSs targeting treatment and therapy recommendations, are presented. Finally, an overview of regulatory considerations regarding therapy recommender systems is given.

2.1 Clinical Decision Support Systems

2.1.1 Definitions and Taxonomies

CDSS 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 [29, 343]. 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 [224], MYCIN [317], DXplain [19]) and already dates back to the 1970ies. Today, there is a large variety of CDSS described in the literature which also vary greatly in design, function, and use. In the following, some proposed general taxonomies are summarized.

An initial categorization of CDSSs was introduced in [318], which distinguished three basic types: (1) *Information-management models*, to provide information on patients or clinical knowledge, e.g. access to literature and educational material, (2) *Situation-awareness models*, to help clinical practitioners to focus attention on specific data, e.g. drug-drug-interaction, and (3) *Patient-data models*, to provide recommendations and customized information on an individual patient, e.g. guide diagnosis or treatment decisions.

The authors of [121], a review on effects of CDSS on health practitioner performance and patient outcomes, categorize the CDSS included into their study according to the clinical task they are supposed to support the physician with: (1) systems for diagnosis, (2) reminder systems for prevention, (3) systems for disease management, and (4) systems for drug dosing and drug prescribing.

In [386] a more detailed taxonomy of CDSS is studied and summarized. Here, CDSS are categorized into six distinct types: (1) *Medication dosing support*, to provide patient-specific drug or dosage recommendations, (2) *Order facilitators*, to support selecting appropriate diagnosis and treatment ordering, e.g. by providing order set templates, (3) *Point-of-care alerts/reminders*, e.g. to provide information about drug-drug interactions, (4) *Relevant information display*, to provide patient-specific data, (5) *Expert systems*, to provide complex decision support which

combines patient characteristics with other electronically available data, e.g. diagnostic or therapy suggestions, and (6) *Workflow support*, to provide tools such as process templates.
[29] and [255] distinguish CDSS approaches according to their implementation properties into *knowledge-based* and *non-knowledge-based* [29] or *knowledge-based* and *intelligent computing systems* [255], respectively.
Knowledge-based CDSSs, also known as expert systems, usually comprise three components. An inference or reasoning engine (1) extracts relevant information from a knowledge base (2), which is then communicated to the user via a communication interface (3). The underlaying knowledge-bases typically consist of compiled rules or probabilistic associations. The inference mechanism combines these rules or associations with actual patient data. The purpose of the communication mechanism is to input patient data into the system, either entered directly by the user or automatically extracted from EHR or other electronic data sources, and to output results to the user. [29, 7]
Non-knowledge-based CDSSs, or intelligent computing systems, on the other hand, typically apply machine learning or other statistical pattern recognition methods to automatically learn from past experiences stored in the clinical data [29, 255]. Such approaches rely on the extensive developments in pattern recognition and machine learning research. Here, no manually encoded expert or domain knowledge is needed. However, such systems require large amounts of data to build reliable models from, which facilitate to generalize sufficiently well on unseen cases. Furthermore, in comparison to rule-based and probabilistic approaches, such machine learning and pattern recognition approaches are often hardly interpretable and decisions not justified. However, insight into the decision-making process was reported to be important factor regarding acceptance of such systems [29].

2.1.2 Electronic Health Records

As introduced in chapter 1, healthcare decisions usually incorporate a wide range of potentially relevant data about a patient. However, as outlined above, also computer-based CDSS intended to facilitate patient-specific decision support require information about patient characteristics [29]. Manual data entry, as required by stand-alone systems, disrupts the patient care process, is time consuming, and subject to erroneous inputs, which limits usefulness and acceptance of such CDSS. Consequently, CDSS ideally use data already entered into an EHR, Hospital Information System (HIS), Laboratory Information Management System (LIMS), Computerized Physician Order Entry (CPOE) systems as detailed below, or even are capable of accessing additional sources with health related data [29].
In Germany, §10 of the professional code ((Muster-)Berufsordnung) of physicians working in Germany [281] urges practitioners to document diagnoses and treatment processes. Such records contain valuable information about medical knowledge and experience. In this context, especially EHRs can facilitate integration of multiple data sources, data exchange, and a reduction of documentation errors if data is provided in standardized formats [142]. Beyond that, selective retrieval of relevant information and automatic data processing open up new opportunities for data-driven CDSS [142, 120]. To date, however, EHRs mostly serve to digitize and manage

information rather than to leverage it and most data remains unused.

Missing data, high-dimensionality, noise, heterogeneity, diverse attribute scales, and often unstructured data is pervasive in the context of clinical data and hence in EHRs. Transforming data into a structured and standardized format, however, is not always easily possible. Especially, standardizing vocabulary [29] and integrating multiple data sources [53] is often challenging. Thought, concerning unstructured data as clinical notes, recent developments in Natural Language Processing (NLP), text mining, and machine learning techniques promise to facilitate powerful algorithms converting such data into standardized representation [120, 197, 316].

2.1.3 Computerized Physician Order Entry Systems

The motivation for the development of CDSS is application dependent. However, prevention of medical errors and improved patient safety can be regarded as the most important issues addressed with CDSS [25, 321, 21]. As the focus of this thesis is on pharmaceutical therapy recommendations, CDSS intended to prevent ADEs caused by inappropriate medication, dosage, duration or drug-drug interactions are of special interest. In this context, especially the coupling of CPOE systems with CDSS, also denoted as CPOE-CDSS, have shown to reduce the number of medication errors and preventable ADEs [308] and are closely linked to CDSS research [25, 175]. CPOE systems facilitate electronic entry and communication of instructions concerning patient treatment, such as pharmaceutical orderings, and have several advantages over manual, hand written ordering. Computer-based orders are typically communicated instantly, more legibly, accurately, and completely [385, 366, 308]. A CPOE system coupled with decision support can additionally improve safety of medication orders by considering individual patient characteristics and current medications but also assure compliance with guidelines and evidence-based knowledge sources. [29, 366]

In [366] such *medication-related* CDSSs are divided into two stages, i.e. *basic* and *advanced* decision support. Basic decision support includes features as drug-allergy checking, basic dosing and administration frequency guidance, formulary compliance checking (i.e. whether the selected medications comply with the institutional medication preferences), duplicate therapy checking (i.e. duplication of medications with similar therapeutic effects), and drug–drug interaction checking. Advanced decision support include features as advanced medication dosing, which take more detailed case variations into account (e.g. indication for the drug, patient characteristics, comorbidities, other medications the patient may be currently taking, or the patient's previous response to the drug), advanced guidance for medication-associated laboratory testing, which assist physicians with medication monitoring (e.g. monitoring related physiological parameters or drug levels), advanced checking of drug disease interactions and contraindications, which take comorbidities and other patient related conditions along with an adequate knowledge-base of drug–disease contraindication into account, and advanced drug–pregnancy alerting, which take potential pregnancy into account and evaluates potential contraindication.

2.1.4 Evidence-based Medicine

The development of CDSSs is also closely related to EbM. EbM, i.e. the practice of medicine based on the best available scientific literature, was widely promoted to improve clinical outcomes. Nevertheless, as also mentioned in chapter 1, EbM remains difficult to be actually practiced by physicians hampers EbM's implementation in most medical domains. One reason is the extensive quantity, complexity and dynamics of clinical research which constitute a great challenge for health practitioners. In order to make evidence-based medicine applicable in practice and to support real-time decision-making at the point-of-care, EbM must be seamlessly integrated into the workflow without causing any interruptions and requiring additional efforts. [29, 25]

The application of clinical guidelines, which intend to summarize the available scientific evidence, has shown to be beneficial for improved quality in practiced medicine. However, in spite of the wide acceptance and recognition of importance of evidence-based guidelines, clinicians are often not familiar with written guidelines and often apply them inappropriately during the actual care process [81]. On the one hand, the implementation of *formalized* clinical guidelines in CDSS promise to improve their acceptance and application, and hence the practice of EbM in the daily clinical routine [81, 217]. Approaches for development and implementation of computer-based guidelines are reviewed and compared in [81]. Several aspects are addressed, as guideline representation, acquisition, and verification but also consideration about execution and application of computer-interpretable guidelines are discussed.
On the other hand, the application of CDSS that help to retrieve and summarize patient-specific evidence-based information from the available scientific literature (*evidence-based decision support*) is assumed to substantially contribute to facilitate the practice of EbM and improve quality of health care [29, 321, 21, 168]. As mentioned in chapter 1, one major challenge concerning the application of EbM and clinical guidelines is to keep up with growing number of clinical studies. CDSS can help to automate literature search and appraisal in order to extract the relevant information and keep a knowledge-base update. Advances in NLP techniques in the medical domain are noteworthy [360, 249]. Such methods promise to support or automate systematic reviews (*Living Review* [100]) and efforts to encounter the challenge of keeping guidelines up to date (*Living Guideline* [6]).

2.1.5 Practice-based Evidence

As also motivated in chapter 1, selection of a patient-specific and personalized therapy options often cannot be provided on the basis of evidence from the literature and guidelines only. Various authors propose to complement this *external evidence* by local or global, *practice-based evidence* for individual and site-specific clinical decision-making [321, 112, 197, 120, 53, 209]. This approach involves systematic analysis, assessment and presentation of such local or global observational experience [209]. The intention is to provide decision support especially in cases where study evidence is lacking or inadequate (Classes I - III) and expert committees (Class IV) are not accessible of fail to find a consensus. One straightforward approach is to mimic a

personalized observational study by dynamically identifying a subgroup of similar patients in the database of past patient encounters [112, 197, 120]. 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 [120]. Clinicians using an EHR ideally generate such practice-based evidence as a by-product of routine care.
In [209], a *Green Button* is proposed which provides real-time and personalized practice-based evidence in the form of comparative effectiveness information for every patient at every visit [209] as visualized in figure 2.1. Such an approach was shown to be successfully applied in [112] using EHR data. Prerequisites, however, are correct identification of homogeneous subgroups [102] and sufficiently large data foundations to draw valid statistical conclusions and to guarantee acceptable power [120]. Nevertheless, such approaches can be able to lower uncertainties and increase transparency, as the underlaying data is accessible, and the in chapter 1 stated issues related to lacking objectivity of studies and meta-analyses can be counteracted. Moreover, such an approach can enable physicians to learn from each case by generation of new knowledge about the effectiveness of treatments and the prediction of outcomes [307].
However, the needed balance and caution when supplementing literature-based evidence with practice-based evidence in the clinical decision-making process is emphasized [351]. Drawing comparative inferences from observational research is also associated with risks such as the potential for unmeasured confounding and selection bias. As patients are not randomized to treatments, comparisons between treatment groups are subject to bias due to patient and physician factors that influence treatment selection [102]. Consequently, such approaches should not be intended to replace clinical judgment but rather increase the information available for physicians to be able to make accurate decisions [290].
The authors of [321] define the notion of *evidence-adaptive* CDSSs. This subclass of CDSSs are supposed to utilize a clinical knowledge-base which reflects the most up-to-date evidence from clinical research literature and practice-based sources. According to [321], such systems should incorporate an additional mechanisms which routinely updates this knowledge-base with new research findings.

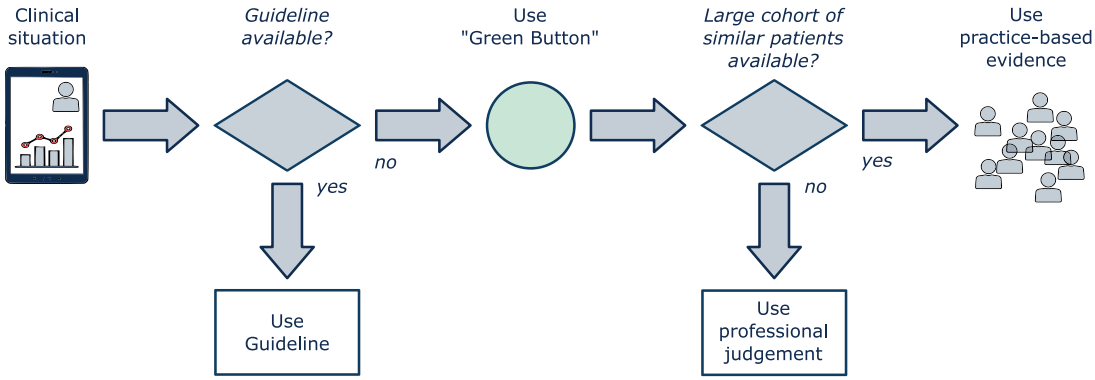


Figure 2.1: Implementing the *Green Button* proposed in [209] into clinical decision-making.

2.1.6 Acceptance, Evaluation, and Application of CDSSs

In [25], a review on CDSS, impacts and potential harms of CDSS are described. Concerning acceptance and successful application of CDSSs, several essential characteristics of successful implementations are mentioned: Timely decision support which don't induce treatment delays, interpretable supporting software instead of "Greek Oracle", notifications and support tailored to suit the current needs, and low false positive alerts, which increase the risk of alert fatigue.
The analysis of publications presented in [168] identified four essential characteristics of CDSSs: (i) decision support must be provided automatically as part of the clinician's workflow, (ii) actionable recommendations must be provided rather than just assessments, (iii) the decision support must be delivered at the time and location of decision-making, and (iv) decision support is supposed to be computer based.
Moreover, the growing engagement of patients in clinical decision-making is to be considered according to [321, 165], which inevitably is associated with improved clinical outcome. In order to ensure patient understanding and to support shared decision-making, CDSS should provide access to prognostic information not only to the clinical practitioner but also to the patient, in [321] denoted as *patient-directed evidence*.

Various measures for the quantification of the benefit of CDSSs were proposed. However, due to the different CDSSs' purposes, no generic metric is applicable to all such systems. A widely used measure is the consistency of outputs compared with physicians or other decision support systems or, in case of systems implementing evidence-based guidelines, the adherence to the respective guideline. Furthermore, evaluation of impact of CDSSs include improvement of the care process and patient health outcome in comparison with clinical care without using decision support. Finally, also organizational outcomes such as cost and efficiency may be evaluated.
Several reviews on the effects of CDSSs are available in the literature with different focuses. In the following, some of them are discussed.

The work published in [121], which dates back to 2005, identified 97 controlled trials assessing different types of outcomes, however, abstracted by the authors to improved practitioner performance and improved patient health outcomes. A wide variety of CDSSs were tested, where the majority (64 %) facilitate diagnosis, preventive care, disease management, drug dosing, or drug prescribing. From the 29 systems supporting with drug-dosing or prescribing decisions, 19, i.e. 66 % showed improved practitioner performance. Nevertheless, the authors state that further research is needed to verify effects of those systems on actual patient health. Those aspects are stated to remain understudied or are inconsistent. Derived factors hindering successful CDSS implementation are, according to the authors, failure of practitioners to use the CDSS, poor usability, lacking integration into the practitioners' workflow, and general practitioners' nonacceptance of computer recommendations.
Also in [168], the literature on CDSS is analyzed to identify relevant system features. Here, outcome is abstracted to improved clinical practice due to the application of CDSSs. As a result, 48 of the 71 included CDSS, i.e. 68 %, demonstrated significantly improved clinical practice. Additionally, as already introduced above, four critical features a CDSS should provide to increase

the likelihood for success. Out of 32 systems which provide all those four features, 30, i.e. 94 % were capable of significantly improving clinical practice.
In [43], a review dating back to 2012, 148 RCTs were included which are grouped with respect to evaluated outcomes: health care process (128 studies), patient health outcome (29 studies), and costs (22 studies). Also this review acknowledges positive effects of CDSS on improving health care processes. Especially, favorable effects of CDSSs dealing with prescribing treatments were shown. Clinicians utilizing the respective systems were more likely to chose the appropriate treatment or therapy. Evidence that demonstrate positive effects regarding clinical and economic outcomes, however, remain generally sparse.
The objective of [150] was to systematically review RCTs assessing the effects of CDSS which are especially designed for drug therapy management purposes. The included studies were grouped into systems evaluated regarding the two aspects process of care and patient outcome. Concerning process of care, in 37 of 59 included studies, i.e. 64 %, applying CDSSs demonstrated improvements, whereas only in 6 of 29 trials, i.e. 21 %, patient outcome could be improved. The authors conclude that lacking clear patient benefits but also lacking data on harms and costs of CDSSs for drug therapy management hinder a clear recommendation for application.
The authors of [229] especially evaluated the effectiveness of CDSSs linked to EHRs by systematically reviewing RCTs on the included systems. The effects of CDSSs on the three outcomes mortality or morbidity (18 studies) and cost (10 studies) are examined. According to the 28 included studies, no clear affect on mortality was evident (16 studies). However, a statistically significant effect regarding the prevention of morbidity could be shown (9 studies). Also some differences concerning costs could be observed, although they were mostly small in magnitude. As the various study results are very heterogeneous and the shown effects rather small, according to the authors, CDSSs overall don't result in substantial benefits or risks for patients in terms of mortality. However, the demonstrated effects largely depend on disease and setting.
Finally, in [248], a systematic literature review of CPOE systems and CPOE systems with decision support (CPOE-CDSS) is conducted. The 16 included systems especially address reduction of medication errors and preventable ADEs. The principal finding of this analysis is that CPOE systems are associated with a significant reduction in medication errors and ADEs. However, no statistically significant difference in effects could be shown between systems with or without CDSS. In the identified studies a reduction of preventable ADEs of more than 50 % was shown.
To conclude, CDSSs seem to enhance practitioner performance and the overall care process. Effects on clinical outcome and patient health remain insufficiently verified and further research is required. The same is true for economic effects. The application of CPOE systems, however, obviously has great potential to reduce medication errors and preventable ADEs. Considering application of CDSS in Germany, according to a survey in 218 hospitals from 2018 [159], CDSSs are not widespread. Considerable numbers state to have tools at their disposal which provide clinical knowledge at the point of care (48.8 %), alerting (34.7 %) or reminder systems (22.0 %). The application of systems which implement medical guidelines or clinical pathways or systems providing active decision support are, however, rarely present as summarized in figure 2.2.

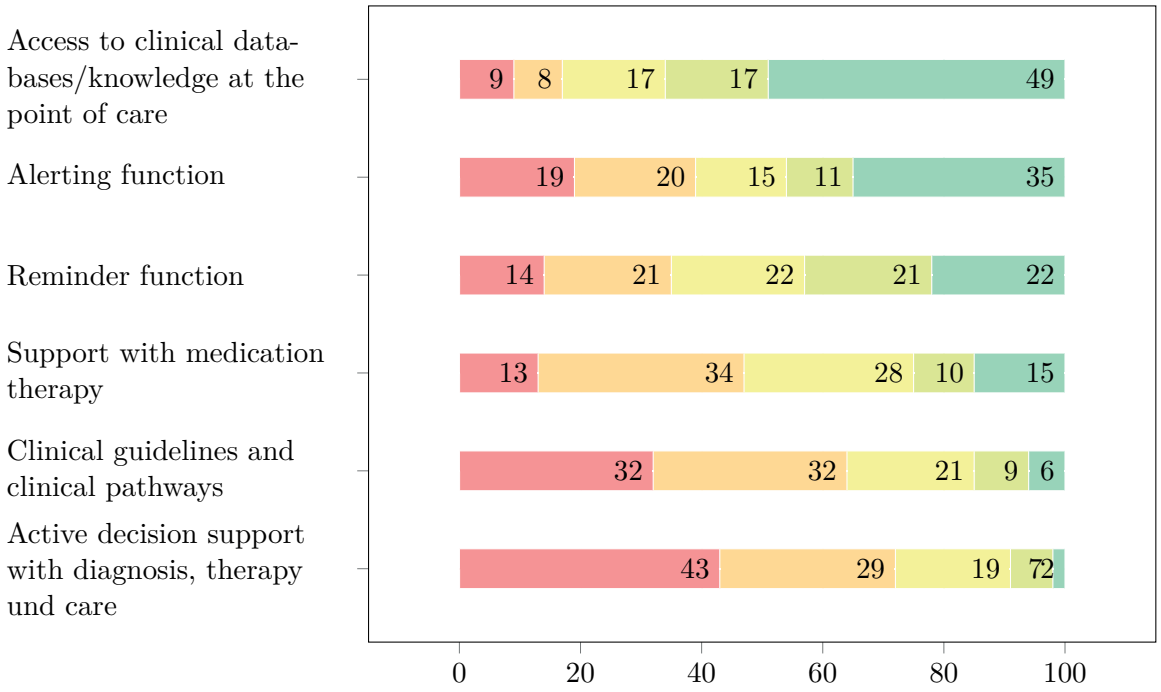


Figure 2.2: Application of CDSS in German hospitals (%) according to [159]. Implemented in all units (—), implemented in at least one unit (—), implementation started (—), planned, but implementation not started yet (—), not yet planned (—).

2.2 Therapy Decision Support Systems

The focus of this work is on CDSSs supporting decision-making concerning drug prescribing or recommendation. The system is supposed to provide a clinical practitioner with customized recommendations by incorporating patient characteristics [386], i.e. based on a *patient-data model* [318]. Both, practice-based evidence stored in the data basis and literature-based evidence captured by the relevant clinical guidelines, are intended to be incorporated. Thus, works describing *knowledge-based* and *non-knowledge-based* systems [29, 255], but also hybrid systems, are of particular interest. Finally, a special focus is put on approaches which consider the decisive CDSS features analyzed in [168], namely (i) automatically generated decision support, (ii) actionable recommendations instead of assessments, (iii) decision support at the time and location of decision-making, and (iv) the decision support is computer based.

To capture the relevant state of the art and identify a research gap regarding methodologies, which particularly meet the above stated requirements, a systematic literature review was conducted initially. As special focus is on medical applications, the search engine PubMed[1] was searched on November 20th, 2017, for studies on treatment, therapy, medication or drug decision support or recommender systems, i.e. according to the search term ((decision support [title]) OR (recommender* [title]) OR (decision aid [title])) AND (treatment [title] OR therapy [title] OR medication [title] OR drug [title]).

[1] https://pubmed.ncbi.nlm.nih.gov/

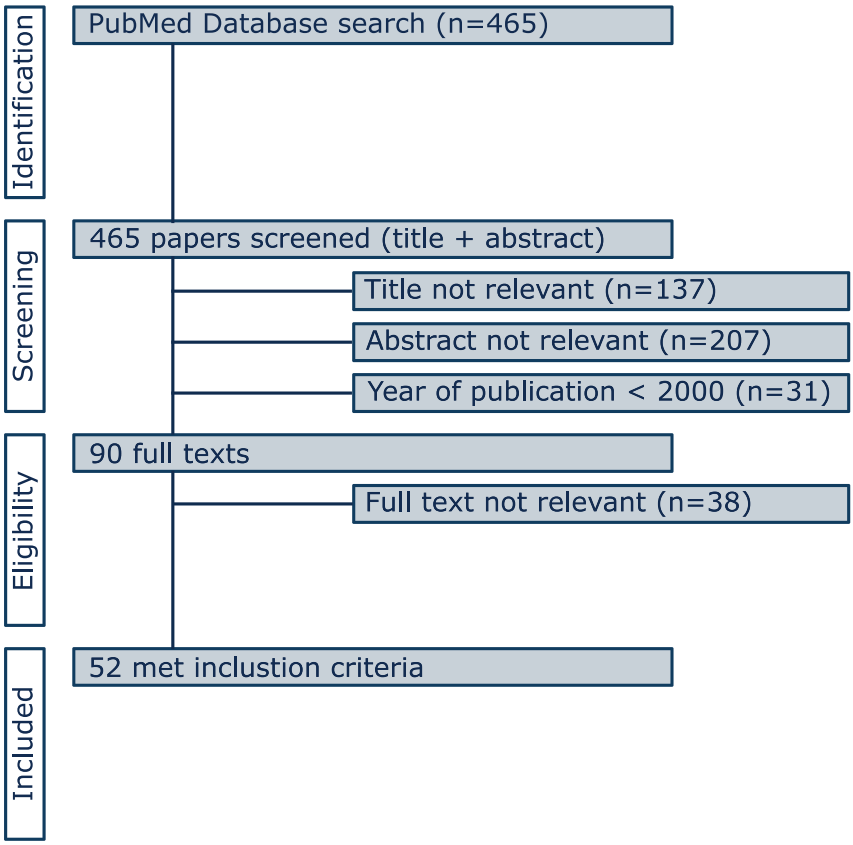


Figure 2.3: Flow diagram depicting the systematic literature review on therapy decision support systems based on the PRISMA statement.

As illustrated in figure 2.3, the search resulted in 465 hits without any duplicates. Following the PRISMA standard[1], the retrieved publications are stepwise evaluated in terms of (i) title, (ii) abstract, and (iii) full text evaluation according to further inclusion criteria.
Inclusion criteria are defined to meet the focuses of this work as stated above and to select works only, which are in accordance with the essential CDSS features for successful CDSSs development from [168]. To summarize, only studies are considered which either deal with (a) active recommendation or exclusion, or (b) prediction (and ranking) of outcome or risks related to potential treatments, therapies, medications or drugs and (c) recommendation, exclusion or prediction must be personalized for an individual target patient. All types of mainly alerting systems, e.g. monitoring for drug allergy or drug-drug-interaction, were neglected (d). Finally, as technical or methodological aspects are of interest, works which neither employ computers (e) nor describe an implementation or prototype or at least a detailed conceptual framework (f), were not considered.
Analyzing titles was conducted by three reviewers, who rated the compliance of each paper with the inclusion criteria according to the three ordinal attributes, *yes*, *no* or *maybe*. Due to the heterogeneous distribution of ratings and only moderate agreement (*Fleiss' Kappa*[2] κ =

[1] http://www.prisma-statement.org/
[2] *Fleiss's Kappa* measures the inter-rater agreement between a number of raters which can be, in contrast to *Cohen's Kappa*, more than two.

0.41), the only little restrictive inclusion scheme summarized in table 2.1 was applied for final selection. Analyzing titles resulted in the exclusion of 137 papers. The number of 328 remaining publications was further reduces to 297 by only including works which were published from 2000 to the date of the literature search.

Table 2.1: Literature search result inclusion scheme. Each listed combination of ratings led to inclusion of the respective publication into further analysis.

Rater 1	**Rater 2**	**Rater 3**
Yes	Yes	Yes
Yes	Yes	Maybe
Yes	Yes	No
Yes	Maybe	Maybe
Yes	Maybe	No
Maybe	Maybe	Maybe

In the following selection steps, abstracts and full text were evaluated by one reviewer for eligibility, i.e. if matching the defined criteria, which resulted in exclusion of 207 and 38 papers, respectively.

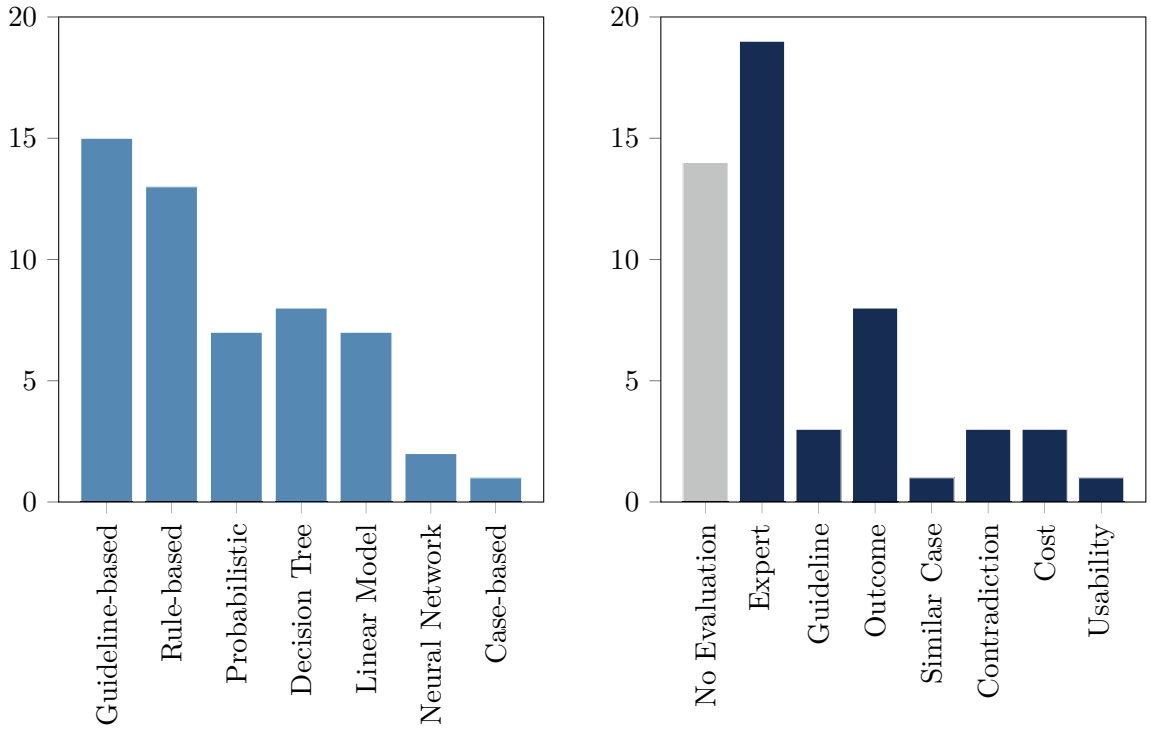


Figure 2.4: Algorithms (a) and evaluation schemes (b) applied in the identified publications.

In summary, our review examined 52 papers which were categorized regarding the type of therapy, the algorithms used and the type of study that was performed to evaluate the system. Furthermore, the condition addressed and the status of the development are discussed. Table A.1 summarizes the identified publications and findings.
From the 52 selected studies, 27 works address CDSSs for pharmaceutical treatments whereas the targeted conditions are very heterogeneous and vary widely. As shown in figure 2.4 (a), about half of the identified studies are knowledge-based (28). 15 implement clinical guidelines [58, 292, 354, 217, 130, 268, 250, 304, 80, 98, 222, 290, 101, 238, 392] and 13 rely on rules generated on the basis of other domain knowledge about the addressed application [32, 31, 144, 183, 279, 244, 185, 268, 64, 328, 218, 269, 11]. The latter category also includes systems which rely on guidelines supplemented with expert knowledge or knowledge from clinical trials [144, 244, 185, 328, 11]). The authors of [32, 31] propose an algorithm that is based on a rule set, which can continuously be updated by experts and is automatically evaluated using a database of previous cases. Regarding non-knowledge-based approaches, various ML or other data mining algorithms are utilized in the identified publications. Probabilistic (Bayessian) approaches (Bayessian Networks [334, 324, 207, 309], Causal Probabilistic Networks [154], Naive Bayes Classifier (NB) [208], others [84]) and decision trees [323, 216, 188, 340, 352, 272, 167] are the most popular choices. Whereas the first derive the likelihood of specific events, the latter learn rules automatically from given data. Both, Bayesian methods and decision trees are intuitive as they model human information processing and are furthermore able to represent causal relationships between variables. Moreover, the application of linear functions is very commonly proposed. Those can be constructed based on domain knowledge ([201, 245, 8]), or linear ([234]) or logistic ([260, 137]) regression models. In [252], a combination of a bagged decision tree ensemble with linear regression is proposed. Finally, two works employ Artificial Neural Networks (ANNs) [391, 263] and only one therapy recommendation algorithm utilizes an instance based algorithm which relies on similar cases in a database [88].
The majority of the suggested algorithms or systems can be considered to be standalone applications (29) and do not integrate existing patient records (EHR) or other data source (CPOE, LIMS) (23). Only one of the proposes systems [304] uses both, structured and unstructured data extracted from clinical notes by using NLP algorithms. Out of the 52 identified works, only for 37 proposed systems evaluation results from a retrospective (25) or prospective (12) study are described in the respective publication. As summarized in figure 2.4 (b), algorithms are typically evaluated concerning the accordance of CDSS recommendations with experts (19), clinical guidelines (3), or similar cases (1), concerning treatment outcome (8), concerning contradicting recommendations (3), concerning saved costs (3), and concerning efficiency of usage (1). For the remaining 16 publications, either only systems without performance characteristics are described (11) or no studies are performed yet (3).

Data-driven CDSSs are an active and dynamic area of research today. This is especially accelerated by advances in and popularity of machine learning research. A repeated search using PubMed on July 10th, 2021, yielded 692 results which are additional 227 (48.82%) publications compared to the initial search.

2.3 Health Recommender Systems

As already introduced in chapter 1, RSs are widely and successfully used in order to support users with the decision-making task in multiple, especially online applications, such as e-commerce, music and movie streaming services, news providers and social media. The underlying concept is to predict users' preference based on information about previous interaction with the system or other knowledge about the target user.

In spite of obvious analogies of typical RS applications and the therapy recommendation setting, the initial literature search on PubMed (section 2.2) did not retrieve any publications on RS methods in medical applications. An extended search including ACM Digital Library[1] and IEEE Xplore[2] and a backward search, which examines the bibliographies of the publications, identified various works employing RS techniques for health applications, namely HRSs, which are summarized in the following. The comprehensive literature review includes all works on HRSs employing typical RS techniques introduced in 3.2 and listed in table 3.1. Furthermore, only publications which focus on clinical applications and utilize data related to health records do derive recommendations are included. All works which deal with recommending hospitals, doctors or social networks, but also nutrition or lifestyle change and behavior recommendations are neglected. Hence, only 24 works from 19 research groups remained which can be broadly categorized into reviews or frameworks on the one hand, and approaches addressing the objectives adverse event prediction and prevention, outcome prediction and therapy recommendation and disease risk stratification, on the other hand. The identified publications on HRSs are analyzed regarding underlying algorithms, category of application, underlying data, and addressed user and are listed in table 2.2. 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. As can be seen, only 10 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 treatment exclusively. 4 works in this group use treatment or clinical order history and 6 works use patient data as basis for recommendations.

2.4 Regulatory Affairs

In the following, the regulatory requirements regarding a data-driven therapy recommender system are discussed. In Europe, basis for the regulations on clinical software in general is the Medical Device Regulation (MDR 2017/745) which is valid since Mai, 25th 2017. After a transition period of three years it finally replaces the EU directives Medical Device Directive (MDD 93/42/EWG) and the Active Implantable Medical Device Directive (AIMDD 90/385/EWG). In contrast to the directives, which were to be transfered into national law, the regulation is a

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp

Table 2.2: Related works regarding HRSs ordered by year of publication, analyzed in terms of underlying algorithms, category of application, underlying data, and addressed user.

Ref.	Year	Algorithm	Category	Data	User
[227]	2007	Memory-based	ADE prediction	Treatment history	Physician
[95]	2008	Association analysis	Treatment	Treatment history	Nurse
[148]	2010	Memory-based	Disease prediction	Patient data	Physician
[105]	2010	Association analysis	Disease prediction	Clinical history	Physician
[79]	2010	Memory-based	Disease prediction	Clinical history	Physician
[94]	2011	Association analysis	Treatment	Treatment history	Nurse
[212]	2012	Memory-based	Treatment	Patient data	Physician
[176]	2012	Memory-based	Treatment	Patient data	Physician
[310]	2013	-	Review	-	-
[325]	2013	Memory-based, content-based, knowledge-based	Treatment	Patient data	Physician
[57]	2013	Memory-based	Disease prediction	Clinical history	Physician
[60]	2013	Association analysis	Treatment	Order history	Physician
[381]	2014	Content-based	Information	Patient data	Physician, patient
[106]	2015	Association analysis, HMM	Disease prediction	Clinical history	Physician
[59]	2016	Association analysis	Treatment	Order history	Physician
[398]	2016	Memory-based, ANN	Treatment	Patient data	Physician
[396]	2016	Memory-based	ADE prediction	ADE data	Physician
[49]	2016	-	Framework	-	-
[145]	2016	Memory-based	Disease prediction	Patient data	Physician
[301]	2017	-	Framework	-	-
[136]	2017	Memory-based	Treatment	Patient data	Physician
[65]	2018	model-based, LR	ADE prediction	ADE data	Physician
[194]	2019	-	Review	-	-
[236]	2020	Memory-based	Treatment	Patient data	Physician, patient
[355]	2020	-	Review	-	-
[82]	2021	-	Review	-	-

binding legal act that all EU countries must fully implement. In order to ease product development complying with the requirements of the relevant EU legislation, European standardization bodies provide harmonized standards with more technical requirement descriptions. As long as the relevant standards are applied during product development and manufacturing, conformity to the MDR is assumed. Table 2.3 lists the relevant harmonized standards in the context of a therapy recommender system.

According to intended purpose and the risks associated with the respective device, the MDR includes 22 classification rules to classify medical devices into risk classes I, IIa, IIb, and III (MDR, Annex VIII). This risk class significantly determines the required efforts concerning conformity assessment and clinical evaluation. Apart from class I, which can be self-assessed, all risk classes require a notified body for conformity assessment according to the MDR. Rule 11 of the MDR especially deals with the classification of software. Software, which is designed to deliver information to support diagnostic oder therapeutic decisions, is at least classified as IIa. However, depending on the potential consequences of decisions, also class IIb or III may be applicable.

In Europe, there are no harmonized standards dealing particularly with data-driven or machine learning software applications and many regulatory issues remain unanswered to date. Nevertheless, also products using data-driven or machine learning algorithms must meet the already existent requirements regarding medical software. To do so, utility and performance must be verified. Furthermore, the product must be developed suchlike that repeatability, reliability and performance can be ensured and also the method how to ensure this verification must be described (MDR, Appendix I 17.1). In case clinical evaluation is based on a comparative product, this algorithms must be sufficiently technically equivalent (MDR, Appendix XIV, Part A, Paragraph 3). Finally, also the competences of the persons involved in the development must be determined and guaranteed (ISO 13485:2016, 7.3.2). In order to support development of such products in spite of the lacking standards, a guideline was published in July 2019 by a consortium of Johner Insitute, notified bodies, manufacturer and experts[1]. However, it must be kept in mind that this guideline is neither a legal requirement nor a harmonized standard.

[1] https://github.com/johner-institut/ai-guideline, accessed Dezember 12th, 2019

Table 2.3: Relevant harmonized standards.

Standard	**Description**
EN ISO 13485:2016	Medical devices — Quality management systems — Requirements for regulatory purposes
EN ISO 14971:2007	Medical devices — Application of risk management to medical devices
EN 62304:2006	Medical device software — Software life cycle processes
EN 62366-1:2015	Medical devices — Part 1: Application of usability engineering to medical devices
IEC/TR 62366-2:2016	Medical devices — Part 2: Guidance on the application of usability engineering to medical devices
IEC 82304-1:2016	Health software — Part 1: General requirements for product safety (not harmonized)

3 Fundamentals

Within this chapter, fundamentals concerning the methods and algorithms used in this work are described. In section 3.1, a detailed overview on similarity measures and distance metrics is provided, targeting different data types and properties. In the following section 3.2, a background on RS methods is given and especially CFs are detailed. Furthermore, in section 3.3, an introduction to ML with special focus on Decision Trees (DTs), DT ensemble techniques, Hidden Markov Models (HMMs), and ANN is provided and also aspects such as model interpretability are discussed. Finally, data preprocessing techniques and the evaluation metrics applied in this work are described in sections 3.4 and 3.5, respectively.

3.1 Patient Similarity

A straightforward approach to make personalized outcome predictions is to identify patients similar to a target patient and derive insights from his or her clinical data. To this end, instances, i.e. patients, are represented in an attribute space suchlike that they ideally form clusters in that space. Representatives of these instance-based and non-parametric methods are K-Nearest-Neighbor (KNN) classification and regression [73, 78, 61, 373, 51, 211, 155, 370, 371, 372, 341, 376, 206, 336], Case-based Reasoning (CBR) [26, 257, 346, 151, 129, 147], unsupervised and supervised *clustering* [254, 368], and structuring data into *patient similarity networks* [23, 369, 375, 356, 253]. However, also the family of memory-based CF algorithms, detailed in section 3.2.2 and applied in section 5.3, rely on the similarity between instances. Additionally, various works propose to train personalized, i.e. *case-specific*, machine learning models on similar patients or cases only, instead of global models, to provide customized and patient specific predictions [196, 243, 388, 356, 105].

3.1.1 Metric Space

The essence of all instance-based algorithms named in the previous section is the appropriate quantification of similarity or distance based on meaningful attributes.
Distance between two instance representations can be measured using a *distance function* which is defined for a metric space. Given the dataset $\mathbf{X}$ of dimension $N \times M$, a distance function

$d : \mathbf{X} \times \mathbf{X} \to \mathbb{R}$ is a *metric* on $\mathbf{X}$, if for any elements $\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k \in \mathbf{X}$ holds

(1) positive-definite:	$d(\mathbf{x}_i, \mathbf{x}_j) > 0 \;\Leftrightarrow\; \mathbf{x}_i \neq \mathbf{x}_j$
	$d(\mathbf{x}_i, \mathbf{x}_j) = 0 \;\Leftrightarrow\; \mathbf{x}_i = \mathbf{x}_j$
(2) symmetry:	$d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$
(3) triangle inequality:	$d(\mathbf{x}_i, \mathbf{x}_k) \leq d(\mathbf{x}_i, \mathbf{x}_j) + d(\mathbf{x}_j, \mathbf{x}_k)$

Suchlike, the pair $(\mathbf{X}, d)$ forms a *metric space* if d is a metric on $\mathbf{X}$. The elements of $\mathbf{X}$ are points in this metric space and $d(\mathbf{x}_i, \mathbf{x}_j)$ is the distance between the points $\mathbf{x}_i$ and $\mathbf{x}_j$.
A common generalization in the context of *metric learning* (section 3.1.6) is the application of *pseudo metrics*. A function $d : \mathbf{X} \times \mathbf{X} \to \mathbb{R} \cup \{+\infty\}$ is called a pseudo-metric (semi-metric) if the conditions (1) to (3) except $d(\mathbf{x}_i, \mathbf{x}_j) = 0 \;\Leftrightarrow\; \mathbf{x}_i = \mathbf{x}_j$ holds. In the pseudo-metric space, non-identical points can have zero distance.

A formal definition of the concept of similarity and the relationship between distance functions and similarity is given in [61]. Simplified, similarity can be defined as some inverse of a distance. Common approaches for converting a distance metric into a *similarity measure* is simply using the negative of a distance or, in case of distance metrics which quantify distance in the range 0..1, by computing the complement $s(\mathbf{x}_i, \mathbf{x}_j) = 1 - d(\mathbf{x}_i, \mathbf{x}_j)$. However, more definitions of *similarity functions* exist such as [35]

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{1 + d(\mathbf{x}_i, \mathbf{x}_j)} \tag{3.1}$$

In the following, selected distance and similarity functions for the various data types are shown as they are applied in section 5.3.1 and 5.3.2, depending on the utilized patient data. In the following, interval and ratio scaled attributes are summarized as *quantitative* and ordinal, nominal and dichotomous attributes as *qualitative* attributes. Ordinal attributes can be transformation to an interval scale under appropriate assumptions regarding the distance between adjacent ordinal categories. For the sake of simplicity, all ordinal variables are assumed to have equidistant categories in this work. Reviews especially addressing patient similarity can be found in [44], [312] and [256].

3.1.2 Quantitative Attributes

For distance computation between quantitative attributes, *Minkowski metrics* describe the family of metrics induced by the p-norms on a vector space. For the points $\mathbf{x}_i$ and $\mathbf{x}_j$ in the metric space $(\mathbf{X}, d)$, their distance is defined as

$$d_{Mink}(\mathbf{x}_i, \mathbf{x}_j) = ||\, \mathbf{x}_i - \mathbf{x}_j \,||_p = \left(\sum_{m=1}^{M} |\, x_{im} - x_{jm} \,|^p \right)^{1/p} \tag{3.2}$$

By the selection of p, *Minkowski metrics* with different order can be generated, however, with

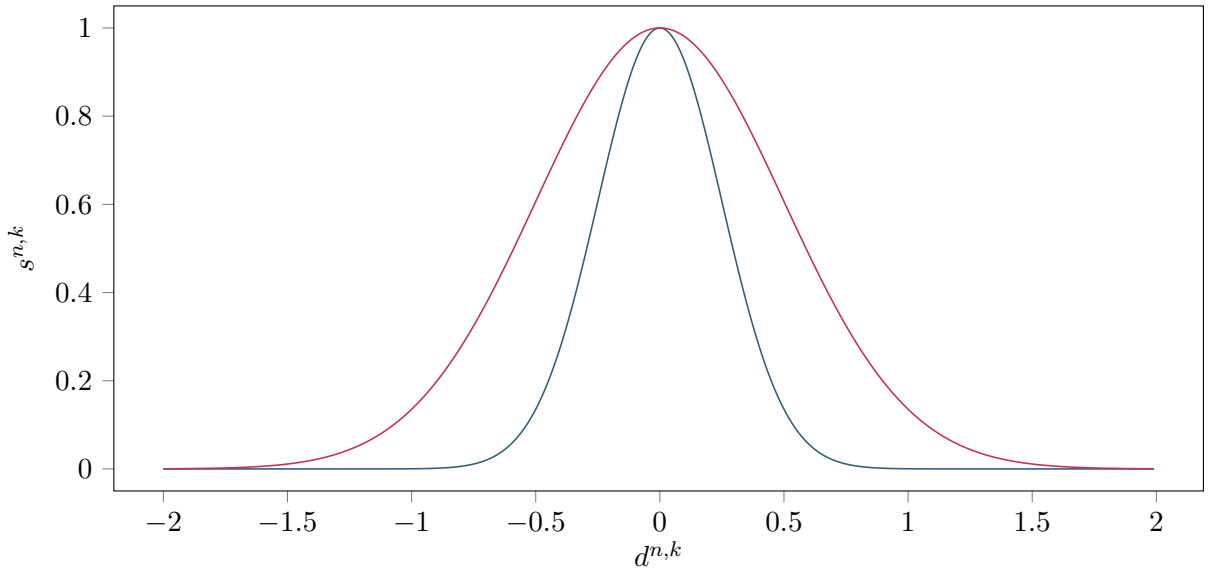


Figure 3.1: Gaussian RBF to convert distance metrics to similarity measures. Two exemplary spread parameters $\sigma = 0.25$ (—) and $\sigma = 0.5$ (—) are shown.

the restriction $p > 0$ in order to meet the triangle-inequality. Generally, with increasing p the impact of large elements increases whereas the impact of small components decreases. The most commonly applied *Minkowski metric* in the context of patient comparison are

$$\text{(1) Manhattan-Distance :} \quad d_{Man}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{M} |\, x_{im} - x_{jm} \,| \qquad (p = 1)$$

$$\text{(2) Euclidean Distance :} \quad d_{Euc}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{m=1}^{M} |\, x_{im} - x_{jm} \,|^2} \qquad (p = 2)$$

$$\text{(3) Chebyshev-Distance :} \quad d_{Cheb}(\mathbf{x}_i, \mathbf{x}_j) = \max_{m=1}^{M} |\, x_{im} - x_{jm} \,| \qquad (p \to \infty)$$

Figure 3.2 shows the respective *unit circles* of these *Minkowski metrics* which demonstrate all points equidistant to the origin of coordinates for the respective metric. Besides the aforementioned general approaches for converting distance metrics to similarity measures, applying the Gaussian Radial Basis Function (RBF) to convert *Euclidean distance* into a similarity is proposed in the context of quantitative attributes [365, 369]

$$s_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2}) \tag{3.3}$$

with the spread parameter σ being a free tuning parameter. In figure 3.1, exemplary RBFs with $\sigma = 0.25$ and $\sigma = 0.5$ are shown. shown.

In contrast to exploiting the relative positions of points in a quantitative vector space, also the relationships between the directions of vectors of data points can be revealed to measure

similarity directly. Typical examples are computing the cosine of the angle between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ [196, 199, 56], denoted as the *Cosine similarity*.

$$s_{Cos}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^T \mathbf{x}_j}{||\mathbf{x}_i|| ||\mathbf{x}_j||} = \frac{\sum_{m=1}^{M} x_{im} x_{jm}}{\sqrt{\sum_{m=1}^{M} x_{im}^2} \sqrt{\sum_{m=1}^{M} x_{jm}^2}} \tag{3.4}$$

Moreover, *Pearson correlation* can be regarded as the *Cosine similarity* between centered vectors. *Pearson correlation* quantifies the degree of linear relationship between two vector representations $\mathbf{x}_i$ and $\mathbf{x}_j$.

$$s_{Corr}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{m=1}^{M} (x_{im} - \bar{x}_m)(x_{jm} - \bar{x}_m)}{\sqrt{\sum_{m=1}^{M} (x_{im} - \bar{x}_m)^2} \sqrt{\sum_{m=1}^{M} (x_{jm} - \bar{x}_m)^2}} \tag{3.5}$$

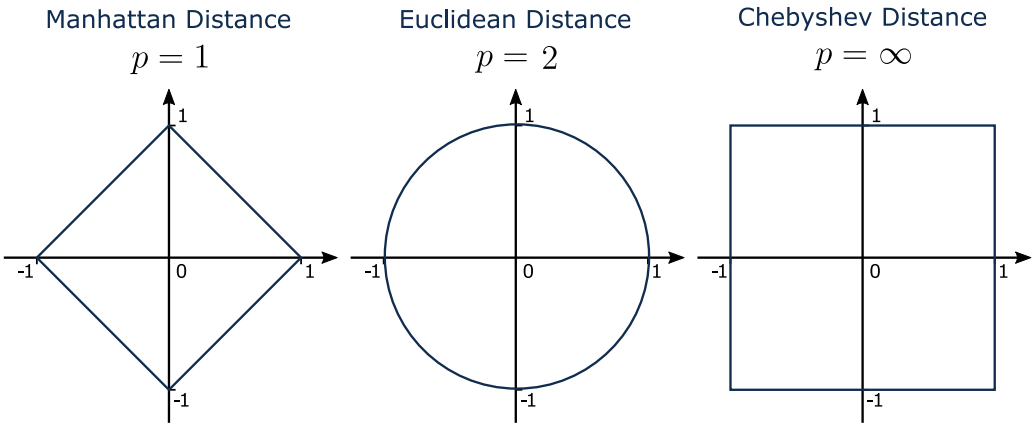


Figure 3.2: *Unit circles* of *Minkowski metrics*

3.1.3 Qualitative Attributes

To compute distance or similarity between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ of length M containing qualitative attributes only, various distance functions are defined which are all based on *overlap*, i.e. the co-occurrence of attribute values. *Hamming distance* generally measures the number of positions in which the symbols in a string differ, i.e. the disagreement between two vectors. Normalized by the number of attributes and transfered to a similarity function according to

$$s_{SMC}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{d_{Hamm}(\mathbf{x}_i, \mathbf{x}_j)}{M}, \tag{3.6}$$

the hamming distance is converted into the Simple Matching Similarity Coefficient (SMC) also denoted as *M-coefficient*. For the application of dichotomous attributes, SMC counts both, the mutual presence and absence of attributes. In contrast, the *Jaccard similarity coefficient* (S-coefficient) [395], intended to compare the similarity of finite sample sets, counts mutual presence only and normalizes by the number of attributes present in at least one vector. Hence, *Jaccard similarity* is beneficial in applications where the presence and absence of values do not

carry equal information, i.e. are asymmetric. In such cases, counting the mutual non-existence of values in both vectors provides no meaningful contribution to similarity. [16]

However, there are many more specialized distance and similarity functions for qualitative attributes proposed in the literature [51]. An extensive review and comparison can be found in [35] and [68].

3.1.4 Mixed-type Attributes

In many real world applications, quantitative and qualitative attributes need to be incorporated into a common distance or similarity function, known as *heterogeneous* distance or similarity measures, respectively.
One approach to combine both data types is discretization of continuous attributes at the expense of information loss and deterioration of generalization capability.
However, various coefficients are proposed and applied in the literature to handle heterogeneous data directly, as the *Gower similarity coefficient* [131, 376] and the Heterogeneous Euclidean Overlap Metric (HEOM) [383, 155].
According to the *Gower similarity coefficient* definition [131], as it is employed in this work, the similarity between the two instances $\mathbf{x}_i$ and $\mathbf{x}_j$ is computed as

$$s_{GSC}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{m=1}^{M} \delta_{ijm} \cdot w_m \cdot \rho_{ijm}}{\sum_{m=1}^{M} \delta_{ijm} \cdot w_m} \tag{3.7}$$

where ρ_{ijm} quantifies the similarity between two instances according to the mth attribute which can additionally be weighted with w_m. The coefficient δ_{ijm} controls whether to include ρ_{ijm} into the similarity computation. δ_{ijm} is set to 1 if the respective attribute is known for both instances and set to 0, otherwise. ρ_{ijm} is defined for three different data types [131]. For similarity computation between interval and ratio scaled (quantitative) values *Manhattan distance*, normalized to the individual attribute ranges, is used, whereas for nominal or dichotomous (qualitative) attributes SMC or *Jaccard similarity coefficient* are applied, respectively. Ordinal attributes are are transformation to interval scale and considered to be quantitative as described in section 3.1.1.

Also for mixed-type, i.e. heterogeneous data there are many more specialized distance and similarity measures proposed in the literature [129, 55, 75].

3.1.5 Feature Selection and Weighting

Individual attributes can be of varying importance for the task of distance or similarity computation or even bear irrelevant or redundant information. In general, large dimensionality of the attributes space is connected with various drawbacks. Large dimensionality is typically detrimental regarding sensitivity to noise and generalization performance. Secondly, the computational burden increases with dimensionality. Moreover, a model with increasing dimensionality

typically becomes less interpretable for users. Last but not least, the *curse of dimensionality* can have a crucial influence on the reliability of the calculated distance or similarity. With increasing dimensionality of the attribute space the representation of the local neighborhood is expanded and becomes imprecise and the concept of similarity increasingly meaningless [295]. As a consequence, attribute selection and weighting was identified as an key research issue in the context of CBR algorithms [26]. Also in the context of KNN approaches redundant, irrelevant, interacting, or noisy attributes were reported to degrade the performance of such algorithms substantially [378]. Works specifically dealing with patient similarity also address the issue of attribute selection and attribute weighting [155].

Generally, attribute or feature selection approaches aim at selecting a subset of the most discriminant attributes in order to minimize redundancy and maximize relevance. Consequently, attribute selection can improve accuracy and generalizability, lower computational complexity and required storage, and strengthen the ability of model interpretability. [161, 349, 160, 205]
In contrast to transformation based dimensionality reduction methods as Principal Component Analyis (PCA) or Linear Discriminant Analysis (LDA), attribute selection choses attributes from the original attribute space. As a consequence, the physical meaning of the original attribute can be maintained in favor of explainability and interpretability. [161, 349, 160, 205]
Three basic attribute or feature selection approaches are distinguished:

(1) *Filter methods* rely on general training data characteristics and are independent of the prediction model. They typically intend to rank the available attributes according to an univariate or multivariate criterion such as distance (Relief-based Algorithm (RBA) [171, 177, 363]), mutual information (mRMR [264]), or correlations with both, target value and other attributes (CFS [143], FCBF [390]). Based on such rankings, attributes can be selected or discarded using application specific thresholds.

(2) *Wrapper methods*, on the other hand, incorporate a classification or regression algorithm to determine the performance of a selected attribute subset in connection with this learning algorithm [174, 273, 326]. Due to this incorporated interaction between attribute subset and prediction model, wrapper methods can have superior performance to filter methods. Given a predefined learner, wrapper algorithms iteratively add or eliminate attributes to or from an attribute subset and evaluate the overall system performance in each iteration. The objective is to optimize the overall performance of the system by finding an optimal attribute subset. As the computational complexity for evaluating each attribute combination for N attributes is $O(2^N)$, an exhaustive search is impractical. To meet this challenge, various heuristics have been proposed which characterize the different wrapper models [273, 326, 327]. Nevertheless, in contrast to filter methods, wrapper own much higher computational cost and may suffer from generalization issues [210, 286, 285, 322].

(3) *Embedded methods* for attribute selection, finally, share the same idea with wrapper methods to perform attribute selection by incorporating the selected prediction model. However, in contrast to wrapper methods, those methods perform attribute selection directly during the model training process. [149]

In addition to selecting only the relevant and discarding redundant attributes, also the individual influence of attributes on the computed distance or similarity can vary in accordance with their importance. For this reason, incorporating attribute weighting schemes into distance or similarity measures was proposed in various non-medical [71, 228, 174, 378, 379, 305, 344] and medical contexts [78, 51, 172].
Attribute weighting strategies intend to assign a value to each attribute reflecting its relevance [228, 378]. Thus, attribute weighting can be considered as a generalization of attribute selection which assigns binary inclusion and exclusion weights. Finally, after scaling the attributes in accordance with their estimated importance, a distance or similarity function can be applied to two data points. Such an approach is utilized in section 5.3.2.1, by adapting and RBA from the domain of the *filter methods* to the given therapy recommendation task.

3.1.6 Metric Learning

Additionally to attribute importance, the multivariate distribution of observations can have crucial impact on the performance of an instance-based prediction algorithm. Accordingly, besides weighting attributes in accordance with their individual importance, more complex attribute space transformations can be beneficial. *Metric learning* deals with automatically learning specialized distance functions that include patterns underlying the data at hand and incorporate feedback from one [337, 336, 374, 371, 370] or several physicians [338, 373] into the distance computation.

The most common approach proposed in the literature is learning a linear transformation before applying a distance function such as the *Euclidean distance*. This linear metric learning approach can be formulated as a generalized quadratic or *Mahalanobis* pseudo-metric or metric

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)} \tag{3.8}$$

with the symmetric positive semi-definite or definite transformation matrix $\mathbf{M}$, respectively. [28, 189, 387, 377, 337, 336, 373, 243, 371, 206]
Unlike unsupervised linear transformations, such as PCA or *Mahalanobis distance* [211, 254], which both decorrelate and standardize the data by rotating the original basis and scaling the data by its standard deviation, metric learning algorithms incorporate application specific supervised information as ground truth or physicians' feedback. This is typically done in the form of similarity or dissimilarity [28, 189, 370, 371] constraints for the instances $\mathbf{x}_i$, $\mathbf{x}_j$ and $\mathbf{x}_k$

and with the application specific boundaries $u,\ l$ and m

$$\begin{aligned} \mathcal{S} &= \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are supposed to be similar}\} \\ d_M(\mathbf{x}_i, \mathbf{x}_j) &\leq u \qquad\qquad (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \end{aligned} \tag{3.9}$$

$$\begin{aligned} \mathcal{D} &= \{(\mathbf{x}_i, \mathbf{x}_j) : \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are supposed to be dissimalar}\} \\ d_M(\mathbf{x}_i, \mathbf{x}_j) &\geq l \qquad\qquad (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \end{aligned} \tag{3.10}$$

or relative distance constraints [28, 189, 274]

$$\begin{aligned} \mathcal{R} &= \{(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) : \mathbf{x}_i \text{ is more similar to } \mathbf{x}_j \text{ than to } \mathbf{x}_k\} \\ d_M(\mathbf{x}_i, \mathbf{x}_j) &< d_M(\mathbf{x}_i, \mathbf{x}_k) \qquad\qquad (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \in \mathcal{R} \\ \text{or } d_M(\mathbf{x}_i, \mathbf{x}_j) &< d_M(\mathbf{x}_i, \mathbf{x}_k) - m \qquad\qquad (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \in \mathcal{R} \end{aligned} \tag{3.11}$$

The objective of the metric learning algorithm is to find the parameters $\mathbf{M}$ of the metric such that it meets those constraints as closely as possible. This is typically formulated as an optimization problem which comprises an objective function $L(\mathbf{M}, \mathcal{S}, \mathcal{D}, \mathcal{R})$, encoding the respective constraints, and a *regularization term* $R(\mathbf{M})$, which is controlled by the regularization parameter $\lambda \geq 0$. The target is to minimize the overall loss, which is increased by every violation of the given constraints, by simultaneously finding the least complex solution [28, 189].

Particularly to learn patient similarity, the Locally Supervised Metric Learning (LSML) algorithm was proposed [337, 336, 338, 373, 374, 243, 96, 341]. The LSML algorithms is intended to learn a linear metric which can be formulated as a generalized *Mahalanobis metric*. The LSML algorithm targets to – based on supervised information – simultaneously minimize local compactness $\mathcal{C}_i$ and maximize local scatterness $\mathcal{S}_i$ of each instance $\mathbf{x}_i$ by optimizing the objective function

$$L(\mathbf{M}) = \frac{\sum_{i=1}^{N} \mathcal{C}_i}{\sum_{i=1}^{N} \mathcal{S}_i} \tag{3.12}$$

with local compactness $\mathcal{C}_i$ and local scatterness $\mathcal{S}_i$ being defined as

$$\mathcal{C}_i = \sum_{\mathbf{x}_j \in \mathcal{N}_i^o} d_M^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} \sum_{\mathbf{x}_j \in \mathcal{N}_i^o} (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j) \tag{3.13}$$

and

$$\mathcal{S}_i = \sum_{\mathbf{x}_k \in \mathcal{N}_i^e} d_M^2(\mathbf{x}_i, \mathbf{x}_k) = \sum_{i=1}^{M} \sum_{\mathbf{x}_k \in \mathcal{N}_i^e} (\mathbf{x}_i - \mathbf{x}_k)^T \mathbf{N} (\mathbf{x}_i - \mathbf{x}_k). \tag{3.14}$$

$\mathcal{N}_i^o$ is the homogeneous and $\mathcal{N}_i^e$ the heterogeneous neighborhood of an instance $\mathbf{x}_i$, i.e. the K nearest neighbors of $\mathbf{x}_i$ with equal and different label, respectively.

Another successfully applied algorithm belonging to the class of linear metric learning algorithms is the Large Margin Nearest Neighbor (LMNN) algorithm proposed in [377]. The LMNN algorithms learns a *Mahalanobis metric* $\mathbf{M}$ by taking *a priori* information regarding

similarity and dissimilarity into account. Within the neighborhood of a target instance, samples which are intended to be similar are pulled towards this target instance whereas samples which are considered to be dissimilar are pushed outside the boundary of the neighborhood. This metric learning method is particularly intended for the application in neighborhood-based classifiers and is adapted to the therapy recommendation algorithm proposed in section 5.3.2.2.

In order to find optimal solution for formulated objective functions $L(\cdot)$, Gradient Decent (GD) but also specialized optimization algorithms adapted to the metric learning model at hand are proposed in the literature [189, 337, 338, 373, 374, 370].

3.2 Recommender Systems

3.2.1 Overview

RS describe software tools and techniques which intend to support users with the decision-making task in versatile applications by providing personalized suggestions [288, 333]. Due to information overload and overwhelming alternatives, especially in online applications as e-commerce, music and movie streaming services, news providers and social media, RS have gained increasing popularity within the preceding years and are an active topic of research. From the service providers' perspective, the benefit of RS is, on the one hand, increasing the number of items sold. On the other hand, a well designed RS can also be capable of increasing user satisfaction and give insight into user needs and preferences [288, 333].

The field of RS has evolved considerably over the recent years yielding sophisticated and specialized methods depending on domain, purpose and personalization level [333]. The typical approach is to predict a user's preference for items and convert these estimates into personalized recommendations [288, 4, 333]. RS types differ in the incorporated knowledge about users and items and concerning the algorithm for computing preference estimates [288].
Specific knowledge and information about users and items, but also feedback on previous purchases or recommendations can be leveraged to personalize recommendations [288]. Such feedback on items can either be implicit user behavior, e.g. simple mouse clicks, browsing behavior, or actual purchases, or explicit in the form of numerical ratings, e.g. a five-star rating system [4, 333]. The observed feedback of the N users $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$ on the M items $\mathcal{I} = \{i_1, i_2, \dots, i_M\}$ are typically stored in a $N \times M$ *user-item feedback matrix* $\mathbf{R}$. Consequently, feedback of users on items are represented in $\mathbf{R}$ as vectors in the user or item space, respectively. A taxonomy of recommendation algorithms can be made differing between 5 basic types of RS models as summarized in table 3.1 [48]. *Collaborative Filtering* (CF) techniques employ the feedback history of multiple users with similar taste to derive personalized preference estimations [298]. *Content-based* RS exploit user feedback along with attributes describing an item to make recommendations. Here, the basic idea is to link user interests, i.e. feedback associated with the item attributes, with items [261]. *Demographic-based* RS use personal attributes of an user to identify demographic classes which are associated with certain preferences [186]. The idea is to identify users with similar demographic properties. *Knowledge-based* RS leverage

users' explicitly specified interests or requirements, which are then combined with domain knowledge and constraints to derive recommendations. *Hybrid* approaches (section 3.2.3) combine the complementary advantages and disadvantages of the aforementioned methods to facilitate more robust recommendations in a wider settings variety [48]. Finally, more advanced models incorporate additional context information, such as location and time, into the recommendation process [288].

There are strengths and weaknesses of the various RS techniques. Especially the *cold start* problem constitutes challenges to most adaptive RS approaches such as CF, content-based and demographic-based recommender. For users for whom no or only little feedback is available, deriving preference pattens and comparison to other users renders difficult or is impossible, respectively [4]. Content-based but also demographic-based methods rely on the explicit association of items with specific content features which needs to be implemented prior to runtime and are static. Demographic-based recommender additionally rely on the willingness of users to provide personal information which potentially limits the popularity of those approaches. The drawback of knowledge-based systems is the limiting factor of acquiring domain knowledge, which also needs to be provided prior to runtime. Furthermore, in comparison with the aforementioned adaptive approaches, this knowledge and hence the association between user and items is static and doesn't evolve over time. However, such knowledge-based algorithms are independent from feedback history and thus, are not having cold start or sparsity problems. [48]

3.2.2 Collaborative Filtering

As stated beforehand, CF models utilize the captured feedback history from multiple users to predict the feedback, i.e. preference of a target user u. The underlying assumption is that unknown feedback can be imputed by exploiting correlations of the observed feedback across various users and items [4]. As CF approaches rely on the feedback of other users only, limitations of content-based methods, such as lacking content information or domain knowledge, can be overcome. There are two basic CF types distinguished, namely *memory-based* and *model-based* methods. [288, 333]
Whereas memory-based techniques use the user feedback directly, model-based methods employ typical machine learning algorithms, as introduced in section 3.3, to develop regression or classification models from user feedback. However, also dimensionality reduction approaches, which uncover latent factors that explain observed feedback, can be regarded as model-based methods. In the following, memory-based CF, as well as dimensionality reduction techniques and linear regression methods are introduces, as they are widely applied in the context of CF. Within this work, various memory-based CF variants and extensions will be contrasted in section 5.3, a model-based linear regression method applied in section 5.4 and an approach utilizing a machine learning model is demonstrated in section 5.5.

Table 3.1: Taxonomy on RS methodologies based on [48]

Type	**Concept**	**Advantages**	**Disadvantages**
Collaborative filtering (CF)	Identification of users with similar preference according to their purchase or feedback history (*user-based* CF). Items preferred by those similar users are recommended.	• No domain knowledge needed • Adaptive and improving over time • Easy implementation	• New user and new item *cold start* problem • Dependent on historic data
Content-based	Identification and recommendation of items which meet the user's preference based on item features (*content*).	• No domain knowledge needed • Adaptive and improving over time	• New user *cold start* problem • Dependent on historic data • Item features are static
Demographic-based	Recommendations based on personal attributes of a user and the identification of demographic classes.	• No domain knowledge needed • Adaptive and improving over time	• *Cold start* problem concerning new users • Dependent on historic data • Dependent on demographic information
Knowledge-based	Match between the user preferences and item features are computed based on domain knowledge.	• No *cold start* problem • Directly maps from user's preference to items	• Static knowledge base • Knowledge engineering required
Hybrid Recommender System	Techniques above are combined to compensate for complementary disadvantages of one techniques by advantages of another one (section 3.2.3).		

3.2.2.1 Memory-based Collaborative Filtering

Memory-based CF models, also widely denoted as *neighborhood-based* CF, belong to the earliest and most popular RS techniques. The popularity of this type of algorithms can be associated with their simplicity, efficiency and ability to provide accurate and personalized recommendations [284, 178, 314, 288]. In chapter 5, such memory-based CF approaches are adapted and applied to the therapy recommendation setting.
In the memory-based CF setting, the user-item feedback matrix $\mathbf{R}$ is directly utilized to predict feedback on items. Two approaches are common. *User-based* CF [152, 36] predict the feedback $\hat{r}_{ui}$ of a target user u on a target item i based on feedback given by similar users on i, whereas *item-based* CF [298, 87] predict $\hat{r}_{ui}$ based on feedback given to similar items by user u, respectively. This work will, however, focus on user-based CF approaches in the following. The transfer to the item-based approach is straightforward.

In case of the user-based approach, the assumption is that feedback of a user u regarding an item i is likely to be similar to the feedback of another user v on i, if u and v have rated other items similarly [126]. The feedback $\hat{r}_{ui}$ of a target user u on an unknown item i is estimated by averaging the observed feedback in the user's neighborhood $\mathcal{N}_i(u)$. If s_{uv} quantifies the similarity between two users u and v, the set of nearest neighbors $\mathcal{N}_i(u)$ having the largest similarity s_{uv} and which provide feedback on item i can be identified. To additionally take the level of similarity between u and a neighbor v into account, the contribution of v to the feedback prediction is typically weighted by its similarity s_{uv}. Hence, $\hat{r}_{ui}$ is defined as the linear combination of the feedback on i observed in the neighborhood $\mathcal{N}_i(u)$ of a target user u with coefficients being the similarity s_{uv}. The impact of s_{uv} can additionally be emphasized using an exponential *case amplification coefficient* $\alpha > 0$ [36] such that

$$\hat{r}_{ui} = \frac{\sum_{v \in \mathcal{N}_i(u)} (s_{uv})^{\alpha} \cdot r_{vi}}{\sum_{v \in \mathcal{N}_i(u)} |(s_{uv})^{\alpha}|} \tag{3.15}$$

The fashion how similarities s_{uv} are computed has crucial impact on performance and recommendation quality of memory-based CF. Both, selection of the nearest neighbors of a target user or item and the impact a neighbor induces into the feedback prediction depend on this similarity [288]. As detailed in the previous section 3.1, there are numerous functions which quantify similarity between user representations. In the context of RS, especially *Cosine similarity* and the *Pearson correlation* coefficient are widely used similarity measures.

As each user typically has its individual scale for explicit feedback normalization schemes have been proposed in the context of CF[288, 152].
Mean-centering refers to transforming a given rating to the polarity of this feedback and the extent to which it is associated with positive or negative sentiments. Therefore, the given feedback r_{ui} is compared to the mean of all available ratings for this user $\bar{r}_u$ or item $\bar{r}_i$, respectively.

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{v \in \mathcal{N}_i(u)} (s_{uv})^{\alpha} \cdot (r_{vi} - \bar{r}_v)}{\sum_{v \in \mathcal{N}_i(u)} |(s_{uv})^{\alpha}|} \tag{3.16}$$

In order to additionally consider the variance of individual rating scales, it is also common to divide the mean-centered feedback by the standard deviation of ratings of a user σ_u or item σ_i. This yields the standardization for user-based predictions

$$\hat{r}_{ui} = \bar{r}_u + \sigma_u \frac{\sum_{v \in \mathcal{N}_i(u)} (s_{uv})^\alpha \cdot (r_{vi} - \bar{r}_v)/\sigma_v}{\sum_{v \in \mathcal{N}_i(u)} |(s_{uv})^\alpha|} \tag{3.17}$$

Besides the utilized similarity measure, also the selection on an appropriate neighborhood size K has significant impact on the recommendation quality. If too few neighbors are incorporated into the feedback estimation, the prediction accuracy is usually low. On the other hand, if the neighborhood is too large, patterns are blurred and noise is induced which leads to a drop in prediction accuracy [288]. In contrast to determining a fixed number of neighbors K, the neighborhood size can also be chosen by determining a similarity threshold thr_s. Here, only neighbors which are localized within this neighborhood are incorporated into the feedback prediction computation. Both K and thr_s are typically determined by cross-validation.

As stated beforehand, sparsity can be challenging to memory-based CF approaches and significantly impact recommendation quality. Furthermore, similarity becomes less meaningful with increasing dimensionality due to the *curse of dimensionality* as mentioned before. One strategy to tackle those problems is to apply dimensionality reduction methods, which can additionally increases efficiency and reduces the impact of noise [345]. The underlying concepts of dimensionality reduction are matrix decompositions, in the context of RS often referred to as Matrix Factorization (MF). Numerous MF algorithms featuring varying constraints regarding the factorization and their objective function have been proposed [345, 4]. In appendix F.2, an overview on MF techniques as typically applied in the context of RS is given.

3.2.2.2 Linear Regression

As feedback predictions in user- and item-based CF algorithms are computed as linear combinations of observed feedback, memory-based CF algorithms can be regarded as a regression task. The applied methodology, however, deviates from conventional linear regression approaches and has particular constraints.
Model coefficients are determined using heuristic similarity measures. Also, as described in section 3.2.2.1, one general regression model is employed for all items or users, respectively. In case of the user-based CF algorithm, similarity between a target user and its neighboring users is utilized as linear coefficients to model feedback of this particular user regarding all possible items. In case of item-based CF, similarity between a specific item and its neighboring items is utilized as linear coefficients to model feedback of all users regarding this particular item. Those similarity measures, however, are not optimal regarding the observed feedback and do not account for any interdependencies among items or users. Furthermore, only a preselected subset of all available information is included into the regression model – only the observed feedback of the K nearest neighbors of a target user or item are incorporated.

If model coefficients were determined using an optimization algorithm, the memory-based CF approach equals a multiple linear regression model [4]. Suchlike, one model for each user or item is learned from the available feedback. Ideally, those models are capable of generalizing the given observed feedback. In case of the user-based algorithm, a user model is learned using observed feedback of this particular user on training items as dependent variables and feedback of other users on the respective items as independent variables. In case of the item-based algorithm, an item model is learned using observed feedback on this particular item of all training users as dependent variables and feedback on other items as independent variables. Such models are expected to reveal the optimal linear relationships between the numeric feedback of users or items while exploiting the interdependency among items or users, respectively [4]. When learning a user model in case of the user-based algorithm, similarly rating users are expected to yield coefficients which are related as well. Whereas when learning an item model in case of the item-based algorithm, similarly rated items are expected to result in related coefficients. Finally, the memory-based CF algorithm can be extended to not just incorporate a subset but all users or items into the regression model. This extension bears the potential to incorporate more valuable information but is subject to the risk to introduce noise into the model.

Optimization-based neighborhood models, intended to learn the model coefficients by only incorporating the local neighborhood of a target user or item, are proposed in several works [180, 27, 259, 3]. By replacing the normalized similarity coefficient $s_{uv}/|s_{uv}|$ with the unknown parameter s^*_{uv}, the feedback $\hat{r}_{ui}$ of target user u on item i can be modeled in the user-based approach as

$$\hat{r}_{ui} = \bar{r}_u + \sum_{v \in \mathcal{N}_i(u)} s^*_{uv}(r_{vi} - \bar{r}_v) \tag{3.18}$$

Analogously to memory-based approaches, the neighborhood $\mathcal{N}_i(u)$ of a target user u are determined using *Cosine similarity* or *Pearson correlation*. The similarity weights s^*_{uv}, however, are proposed to be learned for this particular neighborhood. The objective function $L(\mathbf{s}^*_u)$ introduced in [180] uses the aggregated least-squares between observed r_{ui} and predicted feedback $\hat{r}_{ui}$. As the errors can be added over all items i rated by a user u, one objective function is set up for each individual user u. Assuming $\mathcal{I}_u$ to be the subset of items rated by user u, the objective function can be written as

$$L(\mathbf{s}^*_u) = \sum_{i \in \mathcal{I}_u} (r_{ui} - \hat{r}_{ui})^2 = \sum_{i \in \mathcal{I}_u} (r_{ui} - [\bar{r}_u + \sum_{v \in \mathcal{N}_i(u)} s^*_{uv} \cdot (r_{vi} - \bar{r}_v)])^2 \tag{3.19}$$

which is minimized using optimization solvers as, e.g., GD or Stochastic Gradient Decent (SGD). To cope with overfitting, the optimization variables s^*_{uv} are proposed to be penalized and model complexity reduced by adding a regularization term to each objective function $L(s^*_{uv})$ [180], yielding

$$L(\mathbf{s}^*_u) = \sum_{i \in \mathcal{I}_u} (r_{ui} - \hat{r}_{ui})^2 + \lambda ||s^*_{uv}||_2^2. \tag{3.20}$$

Regularization can be controlled with the user-defined parameter λ. Using the L_2-norm of s^*_{uv}

as regularization term is referred to as *ridge regression* in the context of linear regression [149].

The item-based application can be implemented analogously to the user-based approach. In this case, the regression coefficients $\hat{s}_{ij}$ to be learned represent the correlation between an item i and the nearest neighboring items $\mathcal{N}_u(i)$ which are rated by user u. A popular regression approach in relation to this item-based algorithm is the Sparse Linear Method (SLIM) proposed in [246]. SLIM is intended to learn a sparse coefficient matrix to model feedback given to individual items. In particular, SLIM is designed to work with non-negative feedback values which do not require mean-centering and additionally promote interpretability.
The main difference to equation 3.20 is the extension to *elastic-net regularization*, combining L_2- and L_1-norm regularization. Additionally to L_2-norm regularization, penalizing overall large coefficients, L_1-norm regularization favors sparse solutions for s^*_{ij}, leading to many coefficients having zero value [149]. This property can be regarded as an embedded attribute selection method (section 3.1.5). The additional advantage of sparse solutions is twofold: predictions can be expressed as more interpretable linear combination of only a small number of related items and the computational expenses are reduced. Additionally, SLIM does not restrict the regression coefficients to only the neighborhood $\mathcal{N}_u(i)$ of target item i but includes all available items $\mathcal{I}$.

Here, the feedback $\hat{r}_{ui}$ of user u on item i is predicted as aggregation of all available feedback $\mathbf{r}_u$ of u on all other items.

$$\hat{r}_{ui} = \mathbf{r}_u^T \mathbf{s}_i^* \tag{3.21}$$

Hence, the regularized optimization problem to be solved can be formulated as independent objective functions

$$L(\mathbf{s}_i^*) = \frac{1}{2}||\mathbf{r}_i - \mathbf{R}\mathbf{s}_i^*||_2^2 + \frac{\beta}{2}||\mathbf{s}_i^*||_2^2 + \lambda||\mathbf{s}_i^*||_1 \tag{3.22}$$

which can be added over all items $i \in \mathcal{I}$. $||\mathbf{s}_i^*||_2$ and $||\mathbf{s}_i^*||_1$ are the entry-wise L_1- and L_2-norms of vectors $\mathbf{s}_i^*$, respectively. The additional constraint $s^*_{ii} = 0$ facilitates the target item i to be excluded and trivial solutions to be avoided. Furthermore, the non-negativity constraint $s^*_{ij} \geq 0$ ensures the learned coefficients to represent positive relations between items, i.e. the impact of each feedback, which additionally enhances interpretability.
Typically, $\mathbf{R}$ is employed as training and test data to find an optimal *aggregation coefficient matrix* $\mathbf{S}^* \in \mathbb{R}^{|\mathcal{I}|\times|\mathcal{I}|}$ concatenating all $\mathbf{s}_i^*$ and which minimizes the error on reproduced user feedback $\hat{\mathbf{r}}_i = \mathbf{R}\mathbf{s}_i^*$. Therefore, the available feedback on the $|\mathcal{I}|$ items stored in $\mathbf{R}$ is divided into training and test instances. The training instance are utilized for model training, whereas the remaining observed feedback is used for performance evaluation. Each column $\mathbf{s}_i^*$ of $\mathbf{S}^*$ corresponds to the regression coefficients of one item model.

An extension to SLIM published in [247] focuses on incorporating item side information in order to improve recommendation performance. Here, a number of approaches are compared which constrain the SLIM model by the relations between items or link side information into the model. A modification of the SLIM algorithm is applied for therapy recommendation in chapter 5.

3.2.2.3 Advantages and Disadvantages

When comparing model-based and memory-based CF methods, the main advantage of memory-based methods is their simple and intuitive approach. Recommendations can intuitively be justified, interpreted and explained. In the user-based setting, the list of neighboring users along with the observed feedback a recommendation is based on can be presented to the target user. Moreover, in contrast with model-based methods, memory-based CF do not require extensive data collections as training data [288, 333].
However, as mentioned above, too sparse user-item feedback matrices $\mathbf{R}$ can results in unreliable and biased recommendations or cold start issues [333]. Feedback overlap across users is essential to be capable of comparing users and computing similarity reliably. Additionally, as only items can be recommended for which feedback is provided by neighboring users, large sparsity can also limit the item *coverage*, i.e. the percentage of items the algorithm can provide recommendations for [152].
Furthermore, for large datasets memory-based CF algorithms suffer from scalability issues which directly depends on the number of correlations between users or items to be computed [203]. With increasing number of users, memory and time expenses increase exponentially for the user-based approach. The same is true for increasing items in case of the item-based approach. Compared to model-based approaches, a basic memory-based method doesn't require computationally expensive training. However, calculating feedback predictions during runtime requires to load and search the entire database. In case of large datasets this search can become very expensive or infeasible. This is particularly crucial as such systems typically are supposed to react immediately to online requirements and make fast recommendations. To overcome this issue, nearest neighbors are typically pre-computed in an offline training phase which reduces the complexity of the computation during the recommendation phase. However, this calculation needs to be updated depending on the frequency of dataset changes. And, even though offline, the neighborhood computation can become very complex if the number of users or items becomes large.
Besides model-based methods, another approach to handle the computational burden during runtime is to apply clustering algorithms in a preprocessing phase. Pre-computed clusters allow to search for similar users in a reduced and highly similar space. Here, however, a trade-off needs to be found. With decreasing number of clusters, i.e. granularity, this clustering approach is capable to improve scalability but at the expense of recommendation quality and vice versa. [333, 203, 4]

3.2.3 Hybrid Recommender Systems

To cope with the aforementioned disadvantages of the introduced RS methods and improve recommendation performance, it is a powerful strategy to combine different types of RS algorithms or models of the same type, respectively. Such combinations of two or more recommendation techniques are denoted as *hybrid recommender systems* in the literature [48, 4]. Table 3.2 summarizes a general taxonomy of recommender system hybridization techniques.

Table 3.2: Taxonomy of hybrid RS based on [48]

Method	Description
Weighted	Preference predictions of several RS are combined to produce a single unified rating prediction by computing a weighted aggregate. The results of multiple RS can either be averaged or linear regression-based algorithms can be applied [320, 179] to determine appropriate weights.
Switching	The algorithm switches between various RS depending on the current requirements. This approach was originally motivation to cope with the *cold start* problem, where content-based and collaborative recommender were combined for switching systems [48].
Mixed	Recommendations from several different RS are presented at the same time.
Feature combination	Features from different data sources are combined and utilized in a single recommender algorithm. One approach is to enhance the feedback matrix $\mathbf{R}$ by adding side information, i.e. user or item features [330, 247].
Cascade	One RS refines the recommendations given by another RS.
Feature augmentation	The output of one RS is used to create input features for another RS.
Meta-level	The entire model learned by one RS is used as input to another RS.

3.3 Machine Learning

3.3.1 Overview

In typical classification problems the task is to classify an unknown observation $\mathbf{x} \in \mathbf{X}$ comprising M attributes into a defined category, i.e. class, expressed as target label y. In applications where the target label y is an continuous variable, the task is denoted as regression problem. The objective of ML techniques is to automatically learn the parameters of a model that is capable of generalizing the given set of training observation. The learned model is supposed to be capable of mapping any input vector $\mathbf{x}$ to the appropriate category or continuous target value y, i.e. performing the stated classification or regression task. As training observations are given as pairs of attribute vectors and known target labels $(\mathbf{x}, y)$, such applications are also known as supervised learning problems. [33]

There exists a variety of classification or regression algorithms applied in the medical domain [17] and which can also be transfered to the aforementioned CF setting. Classification or regression algorithms can be contrasted depending on whether classes can be separated or continuous target value computed using linear or non-linear functions. Linear classification approaches, on the one hand, e.g. Logistic Regression (LogR), *Naive Bayes classifier* (NB), or Support Vector Maschine (SVM) with linear kernel, compare linear combinations of input attributes with a threshold to decide on class membership, i.e. they have a linear decision boundary. Linear Regression (LR), as already introduced in section 3.2.2.2, computes the dependent variables value as linear combination of the input attributes. Non-linear approaches, on the other hand, are

capable of learning more complex non-linear relationships between input attributes to determine output value or class membership. One approach is to transform data to a new representational spaces (based on the kernel functions) to apply linear classification techniques, e.g. Multi Layer Perceptron (MLP) or SVM with non-linear kernel. However, data transformation hampers interpretability of the classification process and entails black-box characteristics.
Another way to categorize ML algorithms is the differentiation between *parametric* and *non-parametric* models. Whereas the first abstract the training data by adjusting an *a priori* defined finite set of parameters, the latter cannot be described by a fixed set of parameters but the structure is determined by the data. Whereas all the aforementioned algorithms can be categorized to the group of *parametric* models, the instance based KNN algorithm and the family of DT classifiers and regressors are representatives of the *non-parametric* approaches. [33]
Especially in the context of medical application, besides the ability to generalize the given training observations, explainability and interpretability of classification or regression results but also the capability of handling missing values can determine the algorithm selection.

For this work, DTs and ensembles of DTs, HMMs, and ANN are of particular interest and are explained more in detail in the following sections.

3.3.2 Decision Trees

One of the most popular and successfully applied family of classification and regression algorithms are classification and regression trees, also termed decision trees (DTs). Depending on the algorithm used, such DTs are capable of handling both, quantitative and qualitative data types, i.e. heterogeneous input data, and can even be able to cope with missing values. Additionally, DTs embed feature selection (section 3.1.5) during training which makes them rather robust to irrelevant input variables. Dependent on their complexity, DTs can facilitate a comprehensible, i.e. interpretable decision making process. [149]

A DT can be considered as a hierarchically arranged number of nodes and edges. Each node represents a decision rule which is applied to any observation vectors $\mathbf{x}$ passing through the tree. Suchlike, complex decision problems are divided into a hierarchy of simpler tasks. Depending on the input variable, the observation will end in a leaf node. The empirical class distribution from the training process are stored for each leaf. Suchlike, for each sample to be classified a probabilistic class membership can be determined [226, 275]. In case of a regression task, the actual prediction value for each leaf is the weighted mean of the training data stored in the respective leaf. Hence, complex decision problems are divided into a hierarchy of simpler tasks.

Finding suitable decision rules at the tree nodes is part of training a DT. This *decision tree induction* is a recursive process which develops the tree top-down, i.e. starting from the root node by generating new nodes until a stop criterion is reached. For each node an attribute along with a threshold or rule, depending on the data type, needs to be selected. This discrete splitting criterion is one major characteristic of different DT algorithms. However, all those splitting criteria have in common to aim at increasing a measure the purity or homogeneity by dividing the training data. In appendix F.1, decision tree induction techniques are explained. Additionally, pruning methods to prevent overfitting and how DTs handle missing values is

detailed.

Besides being one of the most studies machine learning algorithms a key advantage of decision trees is the fact that they are simple to understand and to interpret. DT models can be easily visualized or transfered into if-then-rules. However, they are prone to overfitting and lack generalization capabilities if grown to deeply. A further significant issue with decision trees is their high variance resulting in a high tendency to suffer from instability. Small variations in the data can provoke completely different trees. [149]

3.3.3 Decision Tree Ensembles

Combining different models learned from the data, so called *ensemble models* have proven to be a powerful strategy to generate models that improve classification or regression performance and generalization capabilities of single models.
Classifiers or regression models of same type or different models (*stacking*) can built up an ensemble. Generally, there are two types of combining models. *Model selection*, i.e. training and selecting the appropriate model for specific local areas of the feature space and *model fusion*, i.e. training individual models on the entire attribute space and combining their output [190, 271]. The latter approach combines multiple biased or high variance classifiers, i.e. *base learner* which fail on proportions of the data, to form one powerful predictor. Combining the outputs of those diverse models into one single prediction, e.g. taking (weighted) votes for classification or computing (weighted) averages for continuous outputs the individual classifier or regressor provides, more reliable and accurate decisions can be yielded [39].

A crucial keystone shared by those ensemble generation strategies is the concept of *diversity*. The intuition is, by combining diverse models which fail on different data, to reduce the overall error. This overall error of a classifier or regressor can, according to [149], be decomposed into the two sources, the algorithm's *bias* and *variance*. The intrinsic *bias*, on the one hand, is regarded as the error rate on a hypothetical, infinitely large training data. The variance, on the other hand, is associated with errors stemming from variations in a given dataset of finite size.

In the following, the two most widely used ensemble strategies are introduced: *bagging* and *boosting*. Both strategies can be assigned to the classifier fusion approach but differ in the algorithms generating the individual models and in the way how they are merged.
In case of *bagging*, the intention is to combine base models which are characterized by high variance but low bias. Suchlike, the overall variance component and consequently the overall ensemble error can be reduced. Consequently, assuming a base model which is characterized by high variance but low bias, the most popular approach to achieve classifier diversity is to use different training sets derived from the overall training data for training. [271]
In case of *boosting*, the bias is intended to be decreased. Here, multiple individual classifiers, having high bias, are combined into a more powerful ensemble of classifiers.

Building classifier ensembles has proven to be effective for a wide variety of base learner [198]. The focus of this work, however, is on ensemble generated from DTs. *Tree bagging* and *boosting* have proven to be both very effective and to provide state-of-the art results on a wide range of problem such as classification but also regression tasks [271, 62]

As described above, the DT induction process is characterized by high variance but low bias if building deep DTs. As a consequence, small variations in the dataset, features or any parameters used for growing a tree typically lead to differing decision trees. On the other hand, only shallow DTs are characterized by high bias. These features make DTs a suitable choice for employing both ensemble building strategies, *bagging* and *boosting*, to improve the performance of single base classifiers. *Bagging* and *boosting* techniques for DTs are detailed in appendix F.1.4 and also the interpretability of DT ensembles is discussed.

3.3.4 Hidden Markov Models

Markov models are capable of modeling a stochastic processes in which the current state s_t at a discrete point in time t depends on previous system states. In case of first-order Markov processes, s_{t+1} depends on the current state only, rather than on the entire history of the past process (*limited horizon*). The transition into the subsequent state s_{t+1} is defined by the transition probabilities $a_{ij} = P(s_{t+1} = x_j | s_t = x_i)$. HMMs are characterized by an underlying stochastic process of *hidden states* which, however, can only be viewed by emitted *observable symbols* o_t. In each state x_j and point in time t, a symbol o_t is emitted with the probability $b_j(o_t) = P(o_t | s_t = x_j)$. Both, transition probability a_{ij} and emission probability $b_j(o_t)$ depend on the current state only and do not change over time (*time-invariance*). Overall, a HMM is defined by the five-tuple $(\mathcal{X}, \mathcal{Y}, \mathbf{A}, \mathbf{B}, \mathbf{\Pi})$ with the model parameters summarized in table 3.3. Figure 3.3 visualizes a HMM schematically. HMMs are applied in chapter 7 in the context of sleep stage classification.

Table 3.3: HMM model parameters [277]

Description	**Parameter**
Number of states	N
Number of features	M
State alphabet	$\mathcal{X} = \{x_1, x_2, ..., x_N\}$
Output alphabet	$\mathcal{Y} = \{y_1, y_2, ..., y_M\}$
State sequence	$S = (s_1, s_2, ..., s_T), s_t \in \mathcal{X}$
Observation sequence	$O = (o_1, o_2, ..., o_T), o_t \in \mathcal{Y}$
Transition probabilities	$\mathbf{A} = \{a_{ij}\}, 1 \leq i, j \leq N$
Emission probabilities	$\mathbf{B} = \{b_j(o_t)\}, 1 \leq j \leq N$
Initial distribution	$\mathbf{\Pi} = \{\pi_i\}, 1 \leq i \leq N$

The literature distinguishes three fundamental problems related to HMMs [277]:

(1) *Evaluation problem*: Given a model $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$ and an observation sequence $O = (o_1, o_2, ..., o_T)$, how to compute the probability of the observation sequence $P = (O|\lambda)$ efficiently?

(2) *Decoding problem*: Given a model $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$ and the observation sequence $O = (o_1, o_2, ..., o_T)$, what is the most likely underlying state sequence $S = (s_1, s_2, ..., s_T)$?

(3) *Training problem*: Given an observation sequence $O = (o_1, o_2, ..., o_T)$, how to adjust the model parameter $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$ in order to maximize $P = (O|\lambda)$?

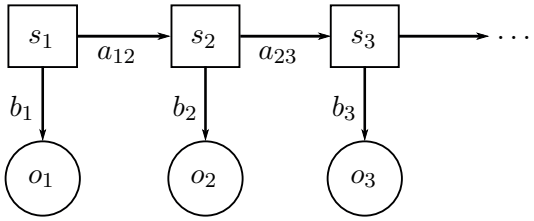


Figure 3.3: Schematic structure of a Hidden Markov Model.

3.3.4.1 Decoding Problem

In order to deduce reliably the hidden states on the basis of a given emission sequence, both the transition probabilities a_{ij} and the emission probabilities $b_j(o_t)$ must be taken into account. The most common approach to solve the decoding problem is the *Viterbi* algorithm [367] which determines the most probable state sequence S using the model parameters λ and the given emission sequence O

$$\underset{S}{\operatorname{argmax}} P(S|O,\lambda). \tag{3.23}$$

Under consideration of the previous state and the emitted symbols, the Viterbi algorithm calculates for each point in time the probability of each state. Since each state probability indirectly depends on all states that have been propagated, the most probable path is recursively selected.

3.3.4.2 Evaluation Problem

In order to compare different models, the probability $P = (O|\lambda)$ is determined. An efficient method of calculating this probability offers the *forward-backward* algorithm. The *forward* variable

$$\alpha_i(t) = P(o_1, o_2, ..., o_t, s_t = x_i|\lambda) \tag{3.24}$$

is the probability that the process is at time t in state x_i, with respect to the previous emission sequence. The *backward* variable

$$\beta_i(t) = P(o_{t+1}, o_{t+2}, ..., o_T|s_t = x_i, \lambda), \tag{3.25}$$

on the other hand, takes the subsequent emission sequence into account. Both variables are combined to yield the requested probability

$$P(O|\lambda) = \sum_{i=1}^{N} P(O, s_t = x_i|\lambda) = \sum_{i=1}^{N} \alpha_i(t) \cdot \beta_i(t). \tag{3.26}$$

3.3.4.3 Training Problem

In case of a supervised learning task, i.e. given labeled training data, the transition probabilities $\mathbf{A}$ can be derived from observed statistics and the emission probabilities $\mathbf{B}$ for each state from

Probability Density Functions (PDFs), such as the multivariate Gaussian distribution function for continuous outputs [226].
In case of unsupervised learning tasks, the training problem can be solved using the *Baum-Welch* algorithm [277]. This variant of the Expectation–Maximization (EM) algorithm estimates the model parameters suchlike that the likelihood of the emission sequence in the training samples is maximized

$$\underset{\lambda}{\operatorname{argmax}} P(O|\lambda). \tag{3.27}$$

3.3.5 Artificial Neural Networks

3.3.5.1 Feed Forward Neural Network

ANNs are composed of several simple computing units, i.e *neurons*, typically arranged in layers but differing in the network's architecture. The Feedforward Neural Network (FNN) is considered the most fundamental network architecture, whose neurons are unidirectionally connected layer by layer from the network input to the network output. Both, the number of *hidden layers* and neurons per layer are hyperparameters. FNNs composed of at least one hidden layer are also denoted as MLPs. [127]
For each neuron j of layer l, as pictured in figure 3.4, the *activation* a_j^l is computed as

$$a_j^l = \sigma^l(z_j^l) = \sigma^l(\sum_n x_n \, w_{n,j}^l + b_j^l) \qquad \text{with} \quad l = 1 \tag{3.28}$$

$$a_j^l = \sigma^l(z_j^l) = \sigma^l(\sum_h a_h^{l-1} \, w_{h,j}^l + b_j^l) \qquad \text{with} \quad l > 1 \tag{3.29}$$

According to equation 3.28, the activation state of a neuron is determined by an *activation function* applied to the network input z_j^l, i.e. the weighted sum of neuron inputs and additional threshold value (*bias*). The activation function, such as sigmoid function, tanh function, or Rectified Linear Unit (ReLU), is essential to introduce non-linearity into the model and enable the approximation of nonlinear functions. [127]

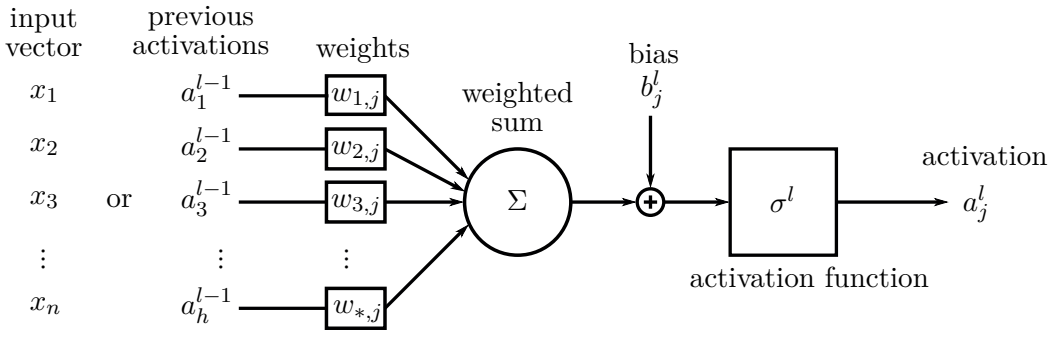


Figure 3.4: Schematic structure of an artificial neuron.

In total, the neural network defines the representation function $\hat{\mathbf{y}} = f(\mathbf{x}, \theta)$ of an input vector $\mathbf{x}$ to an output $\hat{\mathbf{y}}$ and with the network describing parameters, i.e. connection weights, θ. The input vector $\mathbf{x}$ is layer-wisely propagated through the network. Whereas the first layers activations

$\mathbf{a}^1$ ($l = 1$) are computed on the basis of $\mathbf{x}$, activation states of all following layers ($l > 1$) are computed sequentially on the basis of the activations $\mathbf{a}^{l-1}$ of upstream layers. Activation $\mathbf{a}^L$ of the final layer forms the output $\hat{\mathbf{y}}$ of the network. The output layer's activation function depends on the ML task (regression, binary or multi-class classification). In case of a regression problem, a simple linear function or ReLU can be applied. Whereas in case of binary classification, the sigmoid function as introduced is typically applied, in case of a multi-class classification problem, the *sofmax* function maps the output layer's input to a probability distribution over K classes, each represented by an output neuron j. [127]

$$\sigma_{softmax}(z_j) = \frac{e^{z_j}}{\sum_k^K e^{z_k}} \tag{3.30}$$

The aim of network training is to optimize its parameters θ suchlike that an objective function $L(y_i, \mathbf{x}_i, \theta) = L(y_i, \hat{y}_i)$ is minimized for a set of training samples $(\mathbf{x}_i, y_i)$, i.e. the loss between predictions $\hat{y}_i$ and target labels y_i is reduced. Also the applied objective function is determined by the ML task. The loss of a regression task can e.g. measured by Mean Squared Error (MSE). The loss of a binary or multi-class classifier is typically measured by the cross-entropy between predicted and actual distribution *one-hot-encoding* over K classes.

$$L(y_i, \hat{y}_i) = \sum_{k=1}^{K} y_k - \log(\sigma_{softmax}(z_k)) = \sum_{k=1}^{K} y_k - \log(\hat{y}_k) \tag{3.31}$$

For parameter optimization of neural networks, typically SGD variants are applied. In case of the mini-batch SGD, the objective function's average gradients regarding the parameters to be optimized are determined according to equation 3.32 using subsets of size m of the training data (*mini-batch*). For each mini-batch, the resulting gradients are used to update the network parameters according to equation 3.33 with the learning rate μ. The respective parameters are updated according to this error fractions δ_j^l. Within each *epoch*, all training samples are applied once for model training.

$$\mathbf{g} = \nabla_\theta \frac{1}{m} \sum_{i}^{m} L(\mathbf{x}_i, \mathbf{y}_i, \theta) \tag{3.32}$$

$$\theta \longleftarrow \theta - \mu\, \mathbf{g} \tag{3.33}$$

The SGD algorithms are realized by means of the *backpropagation* of errors. The loss is propagated backwards layer by layer through the network and the error fraction δ_j^l of each layer l and neuron j is computed as partial derivative of the objective function with respect to the parameters to be optimized. The respective parameters are updated according to δ_j^l.

3.3.5.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) represent a particular type of FNNs, which is specialized in processing data with a grid-like structure, such as images (2D grid of pixel values) or time series (1D grid of samples) [127]. CNNs differ from FNNs by applying a *convolution* op-

eration (i.e. cross-correlation) instead of ordinary matrix multiplication in at least one of its layers. Figure 3.5 details the procedure of a convolutional layer with a 1D input vector $\mathbf{x}_i$ and 1D filter (*kernel*) comprising three weighting parameters. By shifting the kernel over the input sequence (plus optional zero padding) with defined step size s (*stride*) and cross-correlation computation, a *feature map* is created. Within a convolutional layer, a number of different filters are usually applied which yield different feature maps. Within a training process as described in section 3.3.5.1, the kernel parameters are optimized to extract meaningful features from the input sequence. Downstream *pooling layers* reduce the number of attributes to be processed by local application of a statistical function (maximum, mean) to the feature maps. The feature maps of the final CNN layer are usually concatenated to form an input vector for an MLP for classification. The typical structure of a complete CNN is shown in figure 3.6.

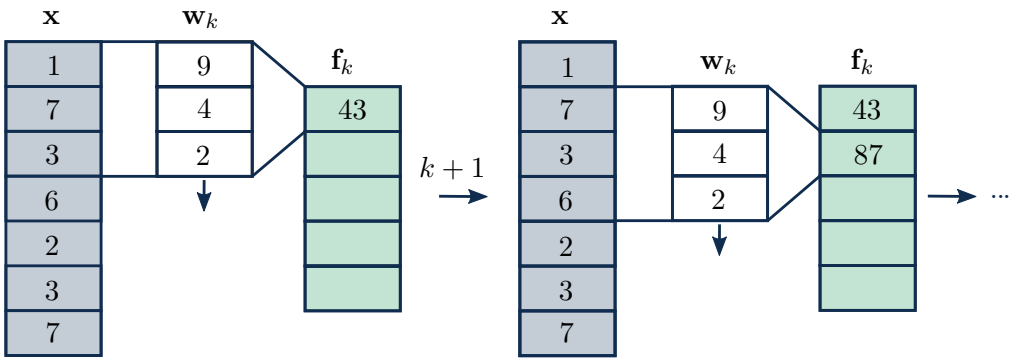


Figure 3.5: Discrete convolution of an input sequence $\mathbf{x}$ with the filter kernel $\mathbf{w}_k$ and the result sequence, i.e. *feature map*, $\mathbf{f}_k$.

The motivation behind CNNs can be summarized by the following three main concepts [127]:

(a) *Sparse connections*: Convolution of a small kernel with a larger input sequence requires both, a smaller amount of parameters and a smaller number of operations to process the input compared to fully connected FNNs.

(b) *Shared parameters*: The kernel parameters only need to be learned once to detect a specific feature within the entire input sequence.

(c) *Translation invariance*: Due to the shared parameters, a specific feature can occur at different places in an input sequence and will generate the same output representation, even though at different position in the output. This invariance is additionally supported by the use of pooling operations.

In this work, CNNs are applied in chapter 7 in the context of sleep stage classification and Parkinson's Disease (PD) patient assessment.

3.3.5.3 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) represent another type of neural networks, which are specialized in processing sequential data. RNNs are characterized by the extension of the computational

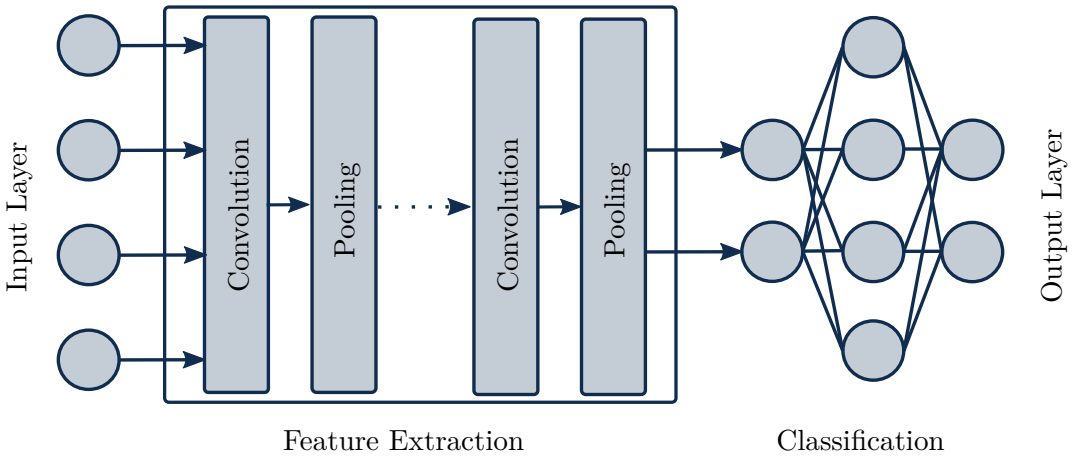


Figure 3.6: Schematic structure of a Convolutional Neural Network.

model by feedback cycles and states, which allows the incorporation of temporal relations. [127] Figure 3.7 shows the exemplary computational model of a RNN with a recurrent hidden layer and an ordinary output layer for processing of an input sequence $\mathbf{x}$ consisting of $t \in \{1, ..., \tau\}$ vectors. Starting from an initial state $\mathbf{h}_0$, the forward propagation of the input sequence $\mathbf{x}$ generates a prediction per time step t according to

$$\mathbf{h}_{(t)} = \sigma^{l_h}(\mathbf{W}\,\mathbf{h}_{(t-1)} + \mathbf{U}\,\mathbf{x}_{(t)} + \mathbf{b}^{l_h}) \tag{3.34}$$

$$\hat{\mathbf{y}}_{(t)} = \mathbf{o}_{(t)} = \sigma^{l_o}(\mathbf{V}\,\mathbf{h}_{(t)} + \mathbf{b}^{l_o}) \tag{3.35}$$

with the respective weight matrices $\mathbf{U}$, $\mathbf{V}$, $\mathbf{W}$, and $\mathbf{b}^l$. A hidden layer neuron's state can be seen as summary of the past input sequence and the current input value. If it is assumed that the prediction $\hat{\mathbf{y}}_{(t)}$ benefits from the incorporation of past and future values, even a bi-directional RNN can be applied. To do so, the recurrent hidden layer is extended by a parallel sub-layer, which is operating in the reverse direction. [127]
Parameter optimization of RNNs is done analogously to FNNs using the backpropagation algorithm described in section 3.3.5.1, however, on the basis of the unfolded model (*backpropagation through time*).

The unfolded RNN representation illustrates two advantages regarding the processing of sequences:

- The RNN's input size is not limited to a predefined sequence size as the model is defined in terms of state transitions.
- For each time step t of a sequence, the processing of an input is realized on the basis of the same model instead of learning separate parameters for each time step.

Hence, the required number of parameters to extract time-dependent features from an input sequence is significantly reduced and extraction of sequential patterns is generalized to be applied to any input sequence length. [127]

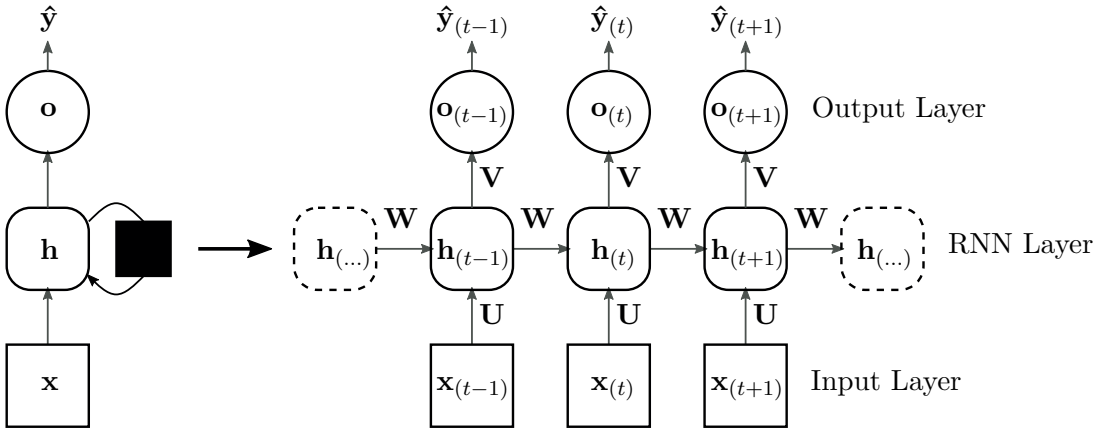


Figure 3.7: RNN in compact and unfolded representation.

A RNN architecture, which is particularly developed to overcome training-related difficulties (*exploding* and *vanishing* gradients [157]) and which facilitates to learn long-term dependencies, is the Long Short-Term Memory (LSTM) [157, 123, 127]. The comparison of an ordinary RNN neuron with a LSTM cell is shown in figure 3.8. In contrast to ordinary RNN neurons, LSTM cells comprise multiple neurons which interact with each other and control information exchange. The internal cell state $C(t)$ represents the key element of the LSTM cell. With the help of *forget gate* and *input gate*, information can be extracted or added from the respective input in a controlled manner. The *output gate* finally determines which portions of the cell state are included in the output $h(t)$ of the LSTM cell. The inner cell state $C(t)$ itself forms a path along the time steps, which allows an information flow with few simple interactions and thus simplifies the feedback of the error signal through the network.

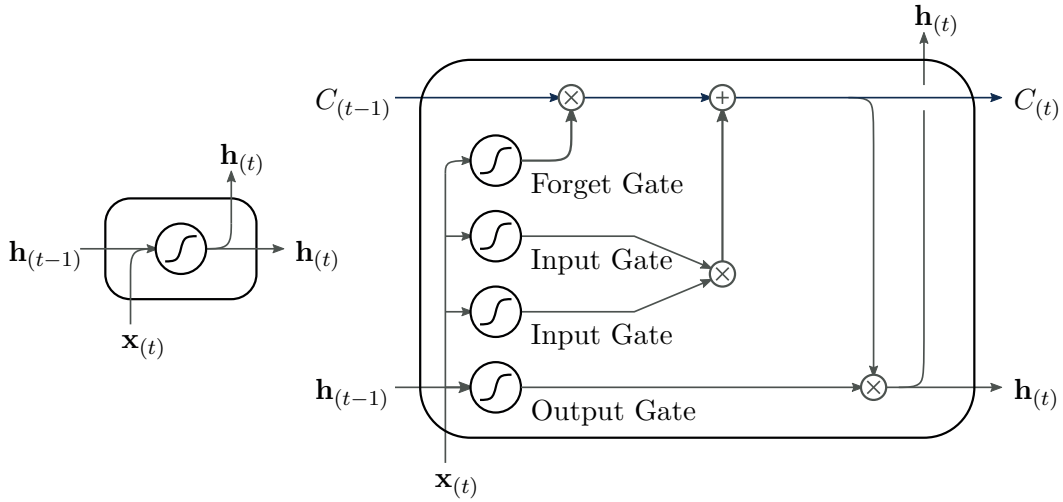


Figure 3.8: Comparison of a RNN neuron with a LSTM cell.

Also LSTMs networks are applied in chapter 7 in the context of sleep stage classification.

3.4 Data Preprocessing

3.4.1 Data Normalization

In many applications, attributes are measured in various different units or are characterized by significantly different variability. When computing distance or similarity in such a vector space, but for the application of most ML methods, attribute rescaling is an essential pre-processing step to equalize the attributes' impact.
Min-max normalization rescales each element in an attribute vector $\mathbf{x}$ to the range $[0, 1]$, according to

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})} \tag{3.36}$$

Standardization linearly rescales each attribute to have zero mean and unit variance

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - \bar{x}}{\sigma} \tag{3.37}$$

where $\bar{x}$ is the mean value and σ the standard deviation of the attribute vector $\mathbf{x}$, respectively. Standardization results in the covariance matrix of the data to be equal to its correlation matrix.

3.4.2 Missing Value Imputation

In the medical context, missing data is a pervasive challenge. The causes range from patients' lacking willingness to give information or be examined, over insufficiently standardized data collection forms, up to faulty transfer of data from one source, e.g. paper documentation, into another system, e.g. an electronic medical record [242, 262]. The majority of typical machine learning algorithms are not capable of dealing with missing values. There are classification and recommender system algorithms which can handle incomplete data to a limited extend. However, too extensive sparsity and the absence of particularly characteristic attributes also distorts classification and recommendation results in those methods and weakens the validity of results and generalizability of models.

Based on the relationship between the mechanism underlying the missing data and the actually missing and observed values, three types of missing data are distinguished as listed in table 3.4.

Obtaining unbiased estimates in case of Not Missing At Random (NMAR) requires domain knowledge or modeling of the missing data mechanism [293, 164, 132, 262]. Nevertheless, also the fact that an attribute is missing can be a valuable information as was shown in [202, 313]. In case of Missing Completely At Random (MCAR) and Missing At Random (MAR), depending on the strategy, imputation methods introduce no or only little bias. A variety of methods for dealing with missing or unknown values are proposed and evaluated in the literature [241, 173, 132, 164, 202, 262, 92, 164] which are summarized in appendix F.3.

Table 3.4: Missing data types distinguished according to the relationship between the mechanism underlying the missing data and the actually missing and observed values [293]: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR).

Type	**Description**
MCAR	The missing data mechanism of a considered attribute is unrelated to the values of any other attribute, whether missing or observed.
MAR	The missing data mechanism of a considered attribute is unrelated to the missing values of this attribute but is conditional on observed values of other attributes.
NMAR	The missing data mechanism is related to the missing values and hence is not ignorable.

3.5 Evaluation Metrics

The quality of RS are typically assessed concerning (1) *accuracy metrics*, which evaluate the performance of the preference estimation task, and (2) *decision support metrics*, which evaluate the quality of the derived top-N list of recommendations [153, 140].
For this purpose, it is either common to evaluate quality offline and retrospectively based on a test dataset $\mathbf{R}_{test}$, comprising feedback on previously consumed items, or by conducting live user experiments. Retrospective evaluation suffers from the drawback that appropriateness of a recommendation can only be determined for the often limited number of actually consumed items. In the context of a therapy RS problem this implies that, as the ground truth is only available for actually applied therapy options and is unobserved for all other options, evaluation metrics can only be computed on those applied therapies. However, this approach facilitates, in comparison with live experiments, much quicker and more economical algorithm evaluation and comparison [153].

The metric typically applied to evaluate the accuracy of numerical preference estimates, such as predicted rating, is the Root Mean Square Error (RMSE) [153, 140]. $RMSE_{ui}$ is computed between all estimated $\hat{r}_{ui}$ and observed feedback r_{ui} of user u on items i captured in $\mathbf{R}_{\mathbf{test}}$, yielding the average error for a test user u

$$RMSE_u = \frac{1}{|\mathcal{I}_{test}|} \sqrt{\sum_{i=1}^{|\mathcal{I}_{test}|} (\hat{r}_{ui} - r_{ui})^2} \tag{3.38}$$

To obtain the overall test performance, $RMSE_u$ is further averaged over all test users $\mathcal{U}_{test}$. RMSE reflects the estimation error in the same value domain as the user feedback. This allows the user – here the medical practitioner – to be provided with an interpretable support for his decision-making.

Decision support metrics to evaluate ranked lists of items are derived from Information Re-

trieval (IR) research [153, 140]. Such lists usually rank all available items according to there relevance, however, only a top-N list of recommendations is presented to the user. Accordingly, ranked lists are evaluated up to this predefined cutoff N.
A widely used decision support metric is *precision*@N, which measures for a test user u the proportion of overlapping items between actually consumed (TP_u) and all recommended (TP_u and FP_u) items in the top-N list

$$precision@N_u = \frac{TP_u}{TP_u + FP_u} = \frac{1}{N} TP_u \tag{3.39}$$

Note that the denominator, i.e. the number of recommended items usually is N but becomes the number of actually consumed items if fewer then N items are consumed in order to make *precision*@N possible to become 1. To obtain the overall test performance, also *Precision*@N_u is typically averaged over the test dataset, i.e. all test users u.
Furthermore, Average Precision (AP)@N for a recommendation list of length N can be derived. AP@N additionally takes the position in the top-N list into account. To do so, *precision*@n at each positions $n = 1...N$ are computed and averaged over the number of recommended items N. As above, the denominator becomes the number of actually consumed items if fewer then N items are consumed. However, only *precision*@n are included for which a TP_u is observed which is controlled by the indicator $\delta(n)$

$$AP@N_u = \frac{1}{N} \sum_{n=1}^{N} \frac{TP_u \cdot \delta(n)}{n} \tag{3.40}$$

MAP@N, finally, averages those AP@N over the test dataset, i.e. all test user u. [153, 140]

The *Cohen's Kappa* coefficient [191] is a statistic to measure inter-rater agreement between two raters. The statistic is applicable to categorical attributes and takes the possibility of the agreement occurring by chance into account. *Cohen's Kappa* is defined as

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{3.41}$$

with the observed agreement between raters p_o and the expected agreement by chance p_e. Assuming N elements to be rated, i a distinct category, f_{ii} the true positives, i.e. the diagonal elements of the confusion matrix, f_{i+} the sum of row i and f_{+i} the sum of column i of the confusion matrix, p_o and p_e are defined as

$$p_o = \frac{1}{N} \sum_{i=1}^{g} f_{ii} \tag{3.42}$$

and

$$p_e = \frac{1}{N^2} \sum_{i=1}^{g} f_{i+} f_{+i} \tag{3.43}$$

Cohen's Kappa is defined on the interval $[-1, 1]$ and can be categorized as defined in [191].

4 Clinical Application

The data available for the development and evaluation of an exemplary therapy recommendation system have been provided by and were worked up with the *Clinic and Polyclinic for Dermatology, University Hospital Dresden*. It represents the routine care of patients suffering from different types of the chronic autoimmune skin disease *Psoriasis*. Psoriasis is particularly well suited for data-driven treatment recommendation, since there is only a moderate number of systemic therapy options and treatment effects are readily measurable, occur within short time intervals and can be relatively reliable associated with the treatment applied. The following section 4.1 gives a background on Psoriasis. Sections 4.2, 4.3, and 4.4 detail the extraction of the data from health records, the content and representation of the data, and details on the applied cleaning, transformation and preprocessing strategies, respectively. Finally, section 4.5 provides some descriptive statistics to summarize the provided dataset.

4.1 Psoriasis

4.1.1 Epidemiology

Psoriasis can be regarded as one of the most common dermatological diseases. Due to its chronic and relapsing course, typically the prevalence is reported in epidemiological studies. Even though the occurrence of Psoriasis differs among ethnic groups and geographic regions, the estimated prevalence within Europe is distributed rather homogeneously between 2 % and 3 %. In Germany, the prevalence is estimated to be 2.53 %, which amounts to approximately 2 million regularly treated patients with men (2.71 %) being slightly more often affected than women (2.31 %) [302, 13]. Adults (age 50 - 79, prevalence 3.99 % - 4.18 %) are more affected than children and adolescent (age < 20, prevalence 0.73 %) [302]. Psoriasis is incurable and requires lifelong treatment and rehabilitation. Considering direct medical costs for statutory health insurance and the patient himself as well as indirect costs caused by absenteeism and reduced productivity, overall mean cost-of-illness are estimated to amount to 5000 € per year and patient according to [162, 240]. However, treatment of moderate or severe cases typically exceeds the costs for mild cases by far and can require hospital admission. In 2014, about 20.000 patients suffering from Psoriasis were treated stationary [162, 240].

4.1.2 Pathogenesis

Psoriasis is assumed to be an immune-mediated disorder which is caused by genetic dispositions [138, 192, 170] and can be supported by associated risk factors as e.g. alcohol consumption,

smoking, stress, overweight, climate, infectious diseases and certain medications such as beta-blockers or psychotropic drugs. However, it is still controversial whether the association with especially alcohol consumption and stress is not based on a reversed causality caused by the psychological burden of the disease [357].
Psoriasis is a chronic disorder. Symptoms occur with varying duration and can be interrupted by symptom free intervals. In spite of great advances in the understanding of Psoriasis pathogenesis within the recent years, only little is known about natural history, determinants of spontaneous remission and the role of age and comorbidities [138].

4.1.3 Symptoms, Diagnosis and Comorbidities

Morphology, distribution, and severity of Psoriasis can be highly variable [192]. According to [357], three main forms of Psoriasis can be distinguished, however, often occurring in parallel, which are introduced briefly in the following.

(1) *Psoriasis vulgaris* (figure 4.1 (top)), also denoted as *Plaque psoriasis*, is the by far most common Psoriasis form and is characterized by a pathologically increased formation of epidermis cells, which leads to a scaling of necrotic cells. Typical symptoms are acute exacerbations of erythematous skin lesions, so called *plaques*, that are covered with dead tissue. Patients usually suffer from elevated sensibility as well as pruritus, burning and pain on the affected areas [170, 192, 138, 357]. Psoratic lesions are typically located at the knees, elbows, and scalp. In severe forms of *Psoriasis vulgaris* also breast, back, arms, and legs are extensively affected [357]. *Psoriasis inversa* is a side-specific variant of *Psoriasis vulgaris* occurring at intertriginous sites and is characterized by shiny and red lesions which are typically free of scales [138, 192, 170]. An acute form of *Psoriasis vulgaris*, known as *Psoriasis guttate* and characterized by small papules erupting on the trunk, is typically developed by children and adolescents only (figure 4.1 (bottom right)). *Psoriasis guttate* either diminishes or is transformed into a classic *Psoriasis vulgaris* in one third of the cases [138]. Furthermore, especially patients suffering from *Psoriasis vulgaris* develop in 50 % of the cases additional disease related nail changes [138, 357].

(2) *Psoriasis postulosa*, a rare Psoriasis form, is characterized by reddening and small, non-infectious pustules. Here, depending on the affected body region, the *Psoriasis postulosa palmopantaris* form, occurring at the palms and the sole of the feet, is distinguished (figure 4.1 (bottom left)). *Psoriasis postulosa* is frequently associated with fever and fatigue besides the skin symptoms [138].

(3) *Psoriatic arthritis* (PsA), a rheumatic form of Psoriasis, is characterized by swelling and pain in the joints of fingers, toes or vertebrae. The symptoms are very painful, hinder mobility and can lead to irreversible destruction of joints. *Psoriatic arthritis* is typically developed as comorbidity accompanying other Psoriasis forms, however, in rare cases also occurs without additional skin symptoms [138, 357, 170].
The various manifestations of the Psoriasis disorder and clinical phenotypes and localizations are summarized in table 4.1.

Table 4.1: Forms of Psoriasis manifestations along with International Classification of Diseases (ICD)-10 codes based on [302] and [357]. The given frequency is computed from the numbers given in [302] and is the proportion of occurrences relative to all Psoriasis cases.

Name	**ICD**	**Main Symptoms**	**Localization**	**Freq.**
Psoriasis vulgaris	L40.0	Erythema, scaling, pruritus	Head, elbow, knee, but also chest back, arms and legs	80.10 %
Psoriasis pustulosa	L40.1	Reddening and noninfectious pustules, frequently associated with fever and fatigue	Whole body	2.52 %
Psoriasis pustulosa palmoplantaris	L40.3	Reddening and noninfectious pustules, frequently associated with fever and fatigue	Palm and sole of the foot	5.0 %
Psoriasis guttata	L40.4	Sudden appearance of round lesions	Face, chest and back	2.28 %
Psoriasis arthritis	L40.5	Reddening, swelling and pain in the joints, stiffness of the joints and constraint movements	Joints of fingers, hand, ankle, knee, elbow and spine	10.10 %
Psoriasis inversa	L40.8	Shiny and red lesions, typically free of scales	Intertriginous sites as armpit, inguinal region, navel	n. r.

Diagnosis and classification of Psoriasis is usually performed by visual examination and sense of touch with special focus on the most relevant body regions. At the presence of more atypical presentations, skin biopsies may be helpful for detection and classification of the Psoriasis form [170]. On the one hand, the main task during diagnosis is to rule out other skin diseases such as dermatitis, mycosis fungoides, tinea corporis, and pityriasis rosea [192, 170]. On the other hand, by combining the subjective reports of patient and dermatologist, the severity of the disease and the effect of previous treatments on the course of the disease is is taken into account. [319].

Patients suffering from Psoriasis often feel stigmatized and impaired in their everyday life due to the visual symptoms. This impacts quality of life and can be a severe psychological burden for patients [192]. Even thought the disorder is rarely life-threatening, it is considered as life-ruining and the psychosocial difficulties can result in depression and anxiety [192, 138, 170].
Furthermore, Psoriasis can be associated with chronic-inflammatory comorbidities as rheumatoid arthritis, chronic-inflammatory bowel disorders and metabolic disorders, adipositas and hypertension [138, 13] and has also been linked to increased risk of cardiovascular disease and diabetes [170].

Figure 4.1: *Psoriasis vulgaris* (top) [219], *Psoriasis postulosa palmoplantaris* (bottom left)[110] and *Psoriasis guttata* (bottom left) [109].

4.1.4 Measurement of Severity

For evaluating clinical signs and treatment outcome it is essential to assess the severity of the disease. Various classification instruments, i.e. clinical scores, are available to facilitate objective measurement. To assess severity of Psoriasis symptoms, the more general Psoriasis Area and Severity Index (PASI), but also Psoriasis form specific scores exist. Also to assess live quality and overall health perceptions, a variety of clinical scores are available with differing focus, such as the Dermatology Life Quality Index (DLQI). In the following, PASI and DLQI are briefly introduces as both are the most commonly applied measures in the context of Psoriasis and as both are part of the available data records.

4.1.4.1 Psoriasis Area and Severity Index (PASI)

The most frequently applied measure is the PASI. It combines the severity of lesions and the area affected into a single score and is intended to standardize the subjective, visual assessment of disease severity. Therefore, the attending physician rates the extend to which the four individual body regions head, torso, upper limb and lower limb are affected (from 0 % to 100 %). These observations are transfered into an area score ranging from 0 to 6 and combined with the severity of the clinical signs erythema (reddening), scaling and induration, each rated from 0 to 4. Overall, the PASI ranges from 0 (no disease) to 72 (maximal disease severity) [113, 319].
Based on this score, Psoriasis is often subdivided into the three severity categories mild, moderate and severe form. Table 4.2 summarizes the classification rules along with a severity distribution of Psoriasis patients in Germany according to [12]. Assessment of severity using the PASI is especially possible for moderate and severe *Psoriasis vulgaris* cases. At the presence of mild forms with a small proportion of affected body area (<5 % - 10 %), a reliable overall assessment cannot always be provided using the PASI. In addition to the PASI, various specialized scores exist which evaluate the specific symptoms of specific Psoriasis forms.

Table 4.2: Severity distribution of Psoriasis patients in Germany according to [12]

Severity	**PASI**	**Frequency**
mild	$PASI \leq 10$	60 %
moderate	$10 < PASI \leq 20$	28 %
severe	$PASI > 20$	12 %

The reduction of the PASI is considered as the primary outcome indicator concerning treatment effectiveness. Hence, a dynamical parameter measuring the proportion of study subjects reaching a defined relative PASI improvement at a specific point in time is the primary endpoint in most controlled clinical trials. According to [233] and the current S3-Guidelines [240], the goal of Psoriasis treatment is reaching a PASI reduction of ≥ 75 % after the treatment induction phase, which is maintained after that phase. The treatment induction phase lasts 10 to 24 weeks, depending on the drug, whereas the maintenance phase is defined as the period after the induction phase. Additionally, a lower border is defined which needs to be achieved as shown in 4.2. A systemic therapy is considered as successful and to be continued if PASI reduction is ≥ 75 % and modified if improvement is < 50 %. Modification of a treatment regimen involves dose adjustment, addition of another therapy, i.e. a combination of treatments, or transition to another drug [233].
For those cases where the PASI improvement is ≥ 50 % but < 75 %, the DLQI described in the following section should also be considered.

4.1.4.2 Dermatology Life Quality Index (DLQI)

Due to the considerable impact on life quality, improvement of quality of life is another crucial goal when treating Psoriasis and other chronic skin diseases. One of the most widely used

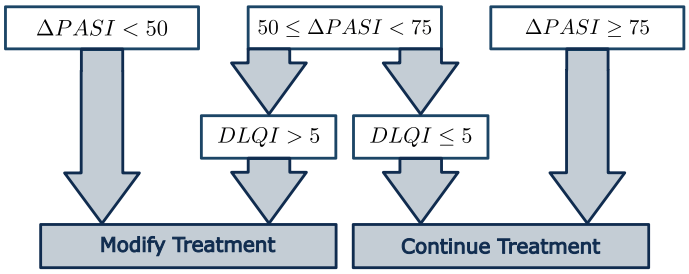


Figure 4.2: Definition of treatment goals, i.e. successful and non-successful therapies of moderate to severe Psoriasis according to [240].

instruments for measuring quality of life in patients with Psoriasis is the DLQI. The DLQI integrates ten questions regarding impairment induced by the respective disease. Overall, the DLQI score ranges from 0 (no impairment) to 30 (maximum impairment) [103]. In case of Psoriasis a $DLQI \leq 10\,\%$ is regarded as mild form, whereas a $DLQI > 10$ is associated with severe impairment [233].

Considering treatment effectiveness, improvement of health related quality of life and patient reported outcomes are gaining importance especially for chronic diseases. Therefore, the evaluation of quality of life has become a secondary endpoint parameter when treating Psoriasis. According to [240] the aspired goal is a DLQI of 0 or 1, indicating no life quality impairment caused by the disease. As shown in figure 4.2, therapy options resulting in DLQI scores higher than 5 are not to be considered as successful. Hence, in those cases where a therapeutic response results in PASI improvements $\geq 50\,\%$ but $< 75\,\%$ and the DLQI is > 5, the therapy should be modified but can be continued otherwise.

4.1.5 Treatment Objectives and Options

Psoriasis is incurable and no complete remission is possible with currently available therapy options. Therefore, the therapeutic objective is to control and reduce clinical signs and symptoms and to minimize ADEs to consequently reduce the disease's impact on the patients life.
The main criterion determining the decision concerning treatment options is the disease severity and possible absolute contraindications [240]. However, also additional factors as the patient's age, occupation and family planning need to be considered when making treatment decisions. Finally, a compromise between the possibly severe ADEs and expected benefits need to be found.

Independent of the disease severity, the current S3 guideline on the treatment of Psoriasis vulgaris [240, 233] recommends a *basic therapy* consisting of ointments with and without active pharmaceutical ingredients for hydration and care of affected skin areas.

For mild Psoriasis vulgaris cases ($PASI < 10$), the symptoms typically can be controlled with *topical therapies* only. Here, the guideline recommends, among others, application of *Vitamine-D* analogs reducing skin cell growth or *glucocorticosteroids*, reducing inflammation. [240, 233]

In case of moderate to severe Psoriasis forms, the S3 guideline recommends the application of a

systemic therapy. Here, a number of immunosuppressive or immunomodulating therapy options exist, administered orally or via subcutaneous (s.c.) or intravenous (i.v.) injection. The four conventional pharmaceutical drugs *Cyclosporine* (CSA), *Methotrexate* (MTX), *Fumaric ester acid* and the retinoid *Acitretin* are recommended as first line treatments. However, especially during long term treatment those therapies are often related to severe ADEs, ranging from nausea, emesis to organ dysfunction or damage. Furthermore, numerous contraindications are described [240] associated with such treatments.

Table 4.3: Systemic therapy options targeting the treatment of Psoriasis and categorized into conventional pharmaceutical drugs and biopharmaceuticals along with type of route of administration (oral, s.c. - subcutaneous, i.v. - intravenous), classification into first- or second-line treatment and absolute contraindications based on [240].

Drug	**Adm.**	**Classific.**	**Absolute Contraindication**
Conventionals			
Acitretin	oral	first-line	- severe renal or hepatic dysfunction - women: pregnancy, breastfeeding, planned child
Apremilast	oral	second-line	- women: pregnancy, breastfeeding
Cyclosporine	oral	first-line	- renal dysfunction - uncontrolled arterial hypertension - active tuberculosis or other severe infections - malignancies present or in medical history
Fumaric acid esters	oral	first-line	- severe renal or hepatic dysfunction - severe gastrointestinal diseases
Methotrexate	oral	first-line	- women: pregnancy, breastfeeding, planned child - severe hepatic diseases - renal insufficiency - active tuberculosis or other severe infections
Biopharmaceuticals			
Adalimumab	s.c.	first-line	- heart failure - active tuberculosis or other severe infections
Etanercept	s.c.	second-line	- heart failure - active tuberculosis or other severe infections
Infliximab	i.v.	second-line	- heart failure - active tuberculosis or other severe infections
Secukinumab	s.c.	first-line	- women: pregnancy, breastfeeding - active tuberculosis or other severe infections
Ustekinumab	s.c.	second-line	- active tuberculosis or other severe infections

Additionally to conventional pharmaceutical drugs, various biopharmaceuticals have been approved within the recent years which have proven to be more effective while causing less severe ADEs than conventional therapies [306]. First-line biopharmaceuticals are *Adalimumab* and *Secukinumab*. Further second-line biopharmaceuticals are *Apremilast*, *Etanercept*, *Infliximab* and *Ustekinumab*. Another second-line biopharmaceutical drug, solely approved for the treatment of *Psoriasis arthritis*, is *Golimumab*. The biopharmaceuticals drugs can be further grouped

according to their mechanism of action, namely into the TNF-α (Tumor Necrosis Factor) antagonists (*Infliximab*, *Etanercept*, *Golimumab*, *Adalimumab* and *Certolizumab*), the interleukin-12 and interleukin-23 (IL-12/13) antibody *Ustekinumab*, and the interleukin-17 (IL-17) antibody *Secukinumab*. However, due to their considerable higher costs, the guideline recommends the application of biopharmaceuticals in cases only where conventional treatment options show insufficient drug response, are contraindicated or ADEs exceed the benefits of the treatment. Since the beginning of the work on this thesis, a number of additional biopharmaceutical drugs have been approved, namely Ixekizumab, Brodalumab, Efalizumab, Tildrakizumab, Guselkumab and Risankizumab. However, as they are neither contained in the available data, nor described in the current Psoriasis treatment guideline [240], they are not included in this study.
Table 4.3 lists the pharmaceutical systemic therapy options mentioned in the current S3 guideline and considered in this work. All drugs are listed along with application type, their classification into first- or second-line treatment and absolute contraindications [240].

Besides conventional pharmaceutical drugs and biopharmaceuticals, phototherapies, as narrow band ultraviolet B light or combinations of psoralen with exposure to Psoralen and Ultraviolet A Light (PUVA), are considered as systemic treatment and are recommended for moderate to severe Psoriasis cases.

Especially due to the burdensome ADEs and unsatisfactorily effective therapies, patients are often unsatisfied with their treatment. This only moderate patient satisfaction results in low adherence to treatments. Studies have shown that only 25 % of patients are entirely satisfied with their treatment [329] and only 40 % of patients adhere to their prescribed medications [289].

4.2 Data Acquisition

As stated initially, the available dataset for development and evaluation of a therapy recommender system was extracted and prepared in collaboration with the *Clinic and Polyclinic for Dermatology, University Hospital Dresden* and reflects routine care of Psoriasis treatment. In total, information on $N = 1424$ consultations from $P = 239$ patients was manually extracted from health records by hospital employees. For the most part, the information stored in the health record is unstructured and consist of written text and notes. In order to derive structured consultation representations, attributes, which are assumed to determine the therapy decisions, were defined together with corresponding categories and value ranges in collaboration with medical experts. Using Microsoft® Access®, input forms were provided for the clinic to facilitate standardized data entry and to structure the data in a relational database. Within an initial revision process, corrupted and invalid data was corrected where obviously possible and missing values, which are assumed to be constant over consultations for a specific patient, were padded until the next valid value (Last Observation Carried Forward (LOCF)). Finally, the Microsoft® Access® database is converted to MariaDB® and stored on a local server. An Entity Relationship Diagram (ERD) of the database structure is shown in figure B.9. The entire processing pipeline, including data acquisition and the mentioned initial preprocessing step, is presented schematically in figure 4.3.

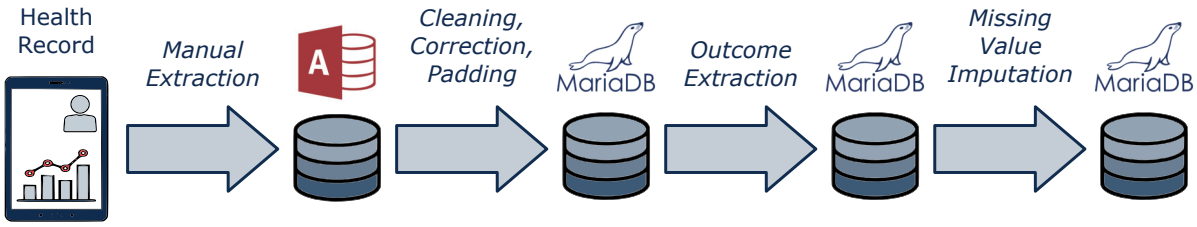


Figure 4.3: Data acquisition and preprocessing pipeline. To facilitate structured data extraction from unstructured health records by hospital employees, Microsoft® Access® input forms are provided. The data is preprocessed, transformed to structured consultation representations, and stored on a MariaDB® server.

4.3 Data Description

4.3.1 Consultation Sequence

The number N_p of medical consultations for each patient $p \in P$ varies. However, all consultations for an individual patient are from consecutive time steps with no in-between missing instances, resulting in a gapless sequence of consultations for each patient as schematically pictured in figure 4.4. For each consultation n and patient p, patient describing attributes $\mathbf{x}_n^p$ (section 4.3.2), i.e. *patient data*, and treatment describing attributes $\mathbf{y}_n^p$ (section 4.3.3), i.e. *therapy decisions* and *therapy outcome*, are distinguished. Additionally, treatment history attributes $\mathbf{a}_n^p$ (section 4.3.4), i.e. *therapy history*, of a patient p and regarding a consultation n is recorded, which collects the outcome of all therapies ever applied in a patients' consultation sequence preceding consultation n.

4.3.2 Patient Describing Attributes

Patient describing attributes, i.e. *patient data*, includes demographic attributes and diagnosed comorbidities as well as attributes concerning diagnosed Psoriasis types and health status. Overall, the dataset contains 23 patient describing attributes for each consultation which are summarized in table 4.4. Here, all attributes are listed along with level of measurement, range of values and availability relative to all consultations. The availability of attributes specifying comorbidities is given relative to the number of documented comorbidities for a consultation. The scale of measurement is highly inhomogeneous and varies among the various attributes. Besides interval scaled numeric (quantitative) attributes the available categorical (qualitative) attributes range from dichotomous and multi-categorical to attributes with ordinal properties. Despite of initial data padding, patient data is partially just intermittently available. Depending on the attribute, this leads to large proportions of missing values and low availability as can be seen in table 4.4.

For both, Psoriasis form and comorbidities, several conditions can co-occur simultaneously. That is why for each of the 6 Psoriasis forms listed in table 4.1 and attributes associated with the 34 comorbidities listed in table B.2 one individual attribute is defined. Each patient describing attribute vector $\mathbf{x}_n^p$ therefore sums up to $D = 126$ attributes.

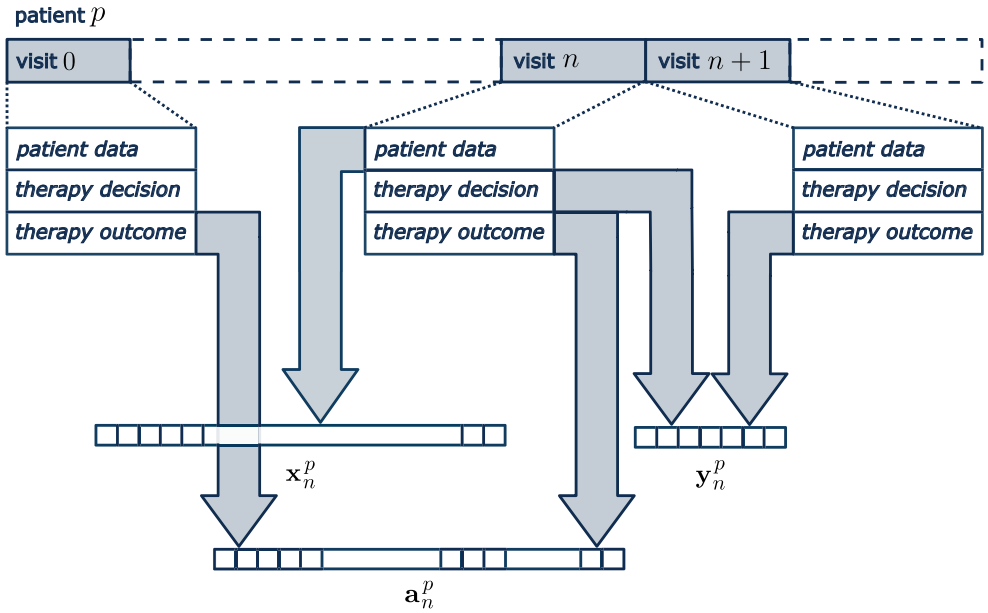


Figure 4.4: Consultation sequence of a patient p. The data captured for each consultation n out of N_p consultations comprises patient describing attributes $\mathbf{x}_n^p$, i.e. *patient data*, treatment describing attributes $\mathbf{y}_n^p$, i.e. *therapy decision* from consultation n and *therapy outcome* derived from the subsequent consultation $n+1$, and treatment history attributes $\mathbf{a}_n^p$, i.e. *therapy history*.

Furthermore, in order to facilitate mathematical operations on the data at hand, categorical, i.e. dichotomous, nominal, and ordinal attributes must be transformed to numeric values. Dichotomous attributes (e.g. gender) are simply encoded into binary attributes. In case of nominal attributes (e.g. child planned), one distinct numeric value is assigned to each of the available categories of a specific attribute (*one-hot encoding*). Ordinal attributes (e.g. comorbidity status) are mapped to discrete numeric values with respect to the ordering of the categories.

Figure 4.5 schematically details the matrix $\mathbf{X}^p$ which holds patient describing attributes for all N_p consultations of an exemplary patient p.

4.3.3 Treatment Describing Attributes

Treatment describing attributes comprise (i) topical and systemic therapies prescribed or recommended by the attending physician, i.e. *therapy decisions*, and (ii) associated *therapy outcome*. *Therapy outcome*, and therewith also the information whether the physician's therapy decision was actually applied, is derived from the subsequent consultation as illustrated in figure 4.4 and described in section 4.4.1. Consequently, the final consultation in a patient's consultation sequence is lacking those application and outcome attributes. Furthermore, the information whether the prescribed or recommended treatment was changed between two consecutive consultations can be derived from a patient's consultation sequence. This information is missing for each first consultation in a consultation sequence as the preceding treatment is unknown.

As defined in chapter 1, the focus of this work is on recommending a systemic therapy for

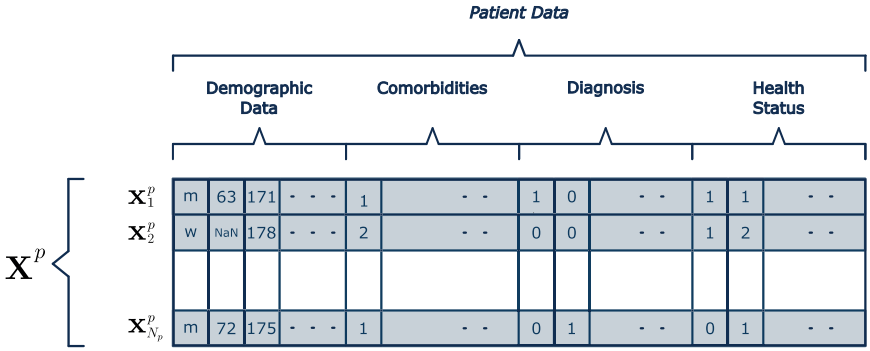


Figure 4.5: Patient describing attributes $\mathbf{X}^p$, i.e. *patient data*, for all N_p consultations n of an exemplary patient p.

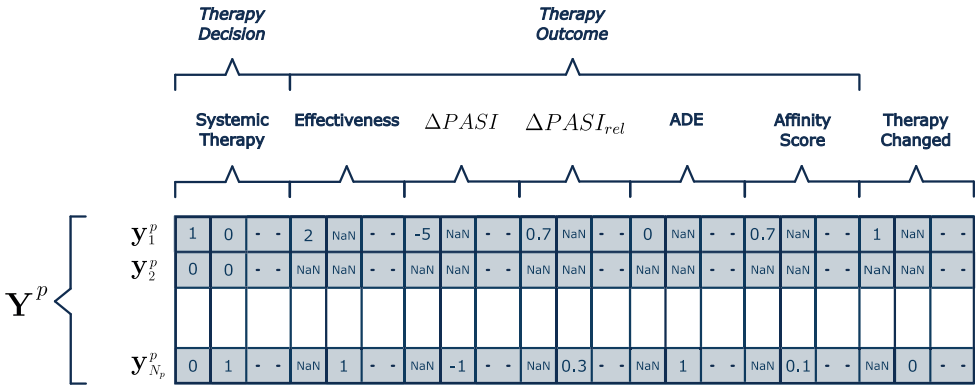


Figure 4.6: Treatment describing attributes $\mathbf{Y}^p$, i.e. *therapy decisions* and *therapy outcome*, for all N_p consultations n of an exemplary patient p.

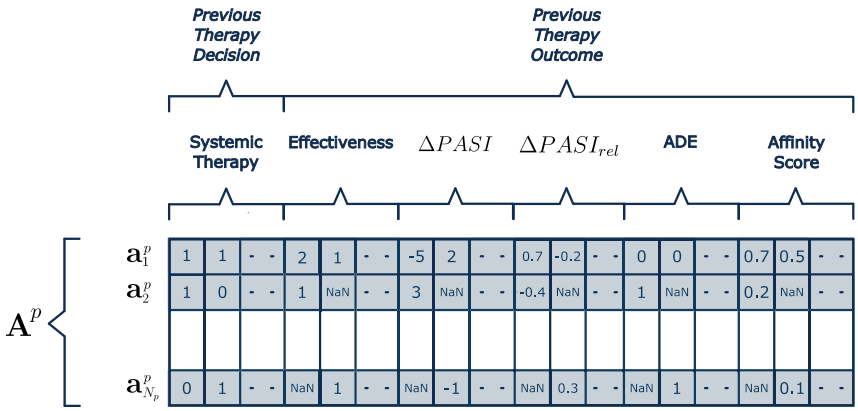


Figure 4.7: Treatment history describing attributes $\mathbf{A}^p$, i.e. *therapy history*, for all N_p consultations n of an exemplary patient p.

Table 4.4: Patient describing attributes, i.e. *patient data*, stored in the $N \times D$ *Data Matrix* $\mathbf{X}$ ($\mathbf{X}'$, $\mathbf{X}''$, see section 4.4.2).

Attribute	**Scale**	**Range**	$\mathbf{X}$	$\mathbf{X}'$	$\mathbf{X}''$
Demographic Data and Comorbidities					
Gender	nominal	male, female	100	100	100
Weight/kg	interval	50 ... 154	54.35	54.75	100
Size/cm	interval	152 ... 204	39.61	39.29	100
Year of birth	interval	1931 ... 1998	100	100	100
Living in partnership	dichotomous	yes, no	53.06	100	100
Education	ordinal	university degree, high school, secondary school, lower secondary school, no graduation	0.24	-	-
Profession	nominal	employed, housewive/-man, in education, not employed	2.98	-	-
Planned child	nominal	child planned, no child planned within the next 12 month and reliable contraception, postmenopausal	13.53	77.29	100
Comorbidities	-	see table B.2 (multiple)	-	-	-
Status	ordinal	not available, unclear, available	100	100	100
Treatment	dichotomous	yes, no	100	100	100
Disease-free	dichotomous	yes, no	100	100	100
Diagnoses and Health Status					
Type of Psoriasis	nominal	see table 4.1 (multiple)	98.71	98.71	100
Year of first diagnosis	interval	1950 ... 2015	90.98	90.98	100
Family anamnesis	ordinal	positive (reliable diagnosis), positive (no reliable diagnosis), negative	54.91	54.91	100
Skin changes: Face	nominal	currently, previously, never	7.00	100	100
Skin changes: Feet	nominal	currently, previously, never	9.42	100	100
Skin changes: Hands	nominal	currently, previously, never	12.24	100	100
Skin changes: Nails	nominal	currently, previously, never	20.45	100	100
Skin changes: Genital	nominal	currently, previously, never	2.82	100	100
PASI	interval	0 ... 72	71.98	100	100
Severity rated by patient	ordinal	healed, mild, moderate, severe, very severe	16.34	-	-
Patient satisfaction with treatment	interval	0 ... 10	9.50	-	-

a given patient and medical consultation. Hence, those consultations associated with topical therapies only are neglected in the following. In case of consultations, for which a combination of systemic treatment along with a supplementary topical treatment are recommended, the observed outcome is assumed to be attributed to the systemic treatment only.
In total, there are 13 distinct systemic therapy options applied in the dataset as listed table B.1. $M^{conv} = 5$ conventional pharmaceutical drugs, $M^{bio} = 6$ biopharmaceutical drugs and including Golumimumab, the generalized group of phototherapies $M^{UV} = 1$, and other not specified systemic treatments $M^{others} = 1$. Additionally, combinations of conventional treatments as Methothrexate and Acitretin with biopharmaceutical drugs or phototherapies are common therapy options. Therefore, those combinations are considered as $M^{combi} = 9$ additional individual therapies resulting in overall $M = 22$ systemic therapy options. However, as later detailed in figure 4.14, those $M = 22$ systemic therapy options are partly just represented by very few occurrences in the data at hand.

Four indicators to quantify therapy outcome can be derived from the collected data. Firstly, *effectiveness*, which represents the patient's perception in the ordinal values *poor*, *moderate* and *good* (1). Secondly, two more objective indicators are extracted from the change of the PASI ($\Delta PASI$) between two consecutive consultations and which are associated with the applied treatment. The discrete $\Delta PASI \in [-72, 72]$ as it is, which does not take the underlaying absolute PASI into account (2a) and the improvement or deterioration of the PASI relative to the absolute value, i.e. $\Delta PASI_{rel} \in [-1, 1]$ (2b). The latter indicator is inspired by the definition of a relative PASI reduction as primary endpoint in clinical studies and yields a continuous ratio scaled value ranging from $\Delta PASI_{rel} < 0$ (deterioration) over $\Delta PASI_{rel} = 0$ (no change) to $\Delta PASI_{rel} > 0$ (improvement). To avoid too much impact from small PASI fluctuations at low absolute values, PASI changes which maintain the absolute value within a range $PASI < 5$ are considered as good outcome, i.e. the disease is successfully controlled and $\Delta PASI_{rel} = 1$. Finally, the occurrence of ADEs are reported as additional negative therapy response (3). Analogously to the patient attributes, also outcome indicators are not always completely given but have missing values as can be seen in table 4.5. Only therapy prescriptions or recommendations for which at least $\Delta PASI_{rel}$ or the subjective *effectiveness* are given or ADE have been reported are denoted as having known outcome in the following.

Additionally, a summarizing outcome parameter is defined merging three of the aforementioned indicators into one score. This *affinity* score $a_{n,m}$ is intended to express the overall effect of a treatment m in a consultation n. *Affinity* is modeled as weighted sum of *effectiveness* ($f_{1,n,m}$) and $\Delta PASI_{rel}$ ($f_{2,n,m}$), and is additionally penalized for occurring ADEs ($f_{3,n,m}$). The weights $w_i \in [0, 1]$ for each component $i \in [1, 2, 3]$ allow to vary the impact of the three described components and δ_i controls the components inclusion. If neither of the two parameters $f_{1,n,m}$ or $f_{2,n,m}$ is given, the *affinity* score is undefined. If only ADEs have been reported, the mean of the *affinity* score value range (0.5) is penalized. Otherwise, the *affinity* score is computed from the available indicators according to

$$a_{n,m} = \frac{\delta_1 \cdot w_1 \cdot f_{1,n,m} + \delta_2 \cdot w_2 \cdot f_{2,n,m}}{\delta_1 \cdot w_1 + \delta_2 \cdot w_2} - \delta_3 \cdot w_3 \cdot f_{3,n,m}$$

with δ_i being 1 for the available and included components. All weights w_i are set to 1 in this work. For *affinity* score computation, all three components are mapped to numeric values in the domain $0 \leq f_{i,n,m} \leq 1$ according to the definitions 4.1, 4.2, and 4.3. In case of $f_{2,n,m}$, a sigmoid function is applied to $\Delta PASI_{rel}$ with the intention to facilitate a linear relation in case of small relative PASI variations with disproportionate impact of increasing values.

$$f_{1,n,m} = \begin{cases} 0.1 & \textit{poor} \text{ effectiveness} \\ 0.5 & \textit{moderate} \text{ effectiveness} \\ 0.9 & \textit{good} \text{ effectiveness} \end{cases} \tag{4.1}$$

$$f_{2,n,m} = \frac{1}{1 + e^{-5 \cdot \Delta PASI_{rel}}} \tag{4.2}$$

$$f_{3,n,m} = \begin{cases} 0.25 & \text{if ADE was reported} \\ 0 & \text{if no ADE was reported} \end{cases} \tag{4.3}$$

All treatment describing attributes, namely *therapy decision*, the indicators specifying *therapy outcome*, and the information whether treatments were changed, are listed in table 4.5 along with level of measurement, range of values and the availability of outcome indicators relative to the number of applied systemic therapies. Similar to the patient describing attributes, the scale of measurement ranges from interval and ratio scaled quantitative attributes to qualitative attributes with dichotomous and multi-categorical nominal and ordinal properties. Analogously to Psoriasis forms and comorbidities in the patient data, one individual attribute is defined for each therapy option and outcome indicator. Therefore, each outcome indicator is represented by a M dimensional vector which holds for each therapy $m \in M$ the observed attribute value and is undefined otherwise. Hence, the resulting sparse therapy attribute vector $\mathbf{y}_n^p$ representing one consultation n of patient p comprises 7 attributes for each therapy option, i.e. $7 \cdot M$ attributes.

Figure 4.6 schematically details the matrix $\mathbf{Y}^p$ which holds treatment describing attributes for all N_p consultations of an exemplary patient p.

Table 4.5: Treatment describing attributes comprising *therapy decisions* and associated *therapy outcome* stored in the $N \times 7 \cdot M$ *Outcome Matrix* $\mathbf{Y}$ ($\mathbf{Y}'$, $\mathbf{Y}''$, see section 4.4.2). The outcome indicators' availability is given relative to the number of applied systemic therapies.

Attribute	**Scale**	**Range**	$\mathbf{Y}$	$\mathbf{Y}'$	$\mathbf{Y}''$
Systemic therapy	nominal	see table 4.5	-	-	-
Effectiveness	ordinal	good, medium, bad	98.72	100	100
$\Delta PASI$	interval	-27 ... 18	57.99	100	100
$\Delta PASI_{rel}$	ratio	-1 ... 1	57.99	100	100
ADE	dichotomous	yes, no	100	100	100
Affinity score	ratio	0 ... 1	99.53	100	100
Therapy changed	dichotomous	yes, no	81.56	81.56	100

4.3.4 Treatment History Attributes

Additional to $\mathbf{y}_n^p$, the treatment history attributes, i.e. the Psoriasis *therapy history* of a patient p and regarding a consultation n is recorded. Therefore, the outcome of all therapies ever applied in a patient p's consultation sequence preceding consultation n are collected in $\mathbf{a}_n^p$. Besides those therapies which can be derived from a patient's consultation sequence, the data additionally provides for each patient all known therapies applied previously to the first consultation in the sequence. Outcome of a therapy option is always updated with the most recently observed outcome for this respective treatment. Analogously to the treatment describing attributes described in 4.3.3, the vector $\mathbf{a}_n^p$ comprises 6 attributes for each therapy option $m \in M$, i.e. $6 \cdot M$ attributes. In table 4.6, all attributes are listed along with level of measurement, range of values and the availability of attributes relative to the number of previously applied therapies. Figure 4.7 schematically details the matrix $\mathbf{A}^p$ which holds treatment history describing attributes for all N_p consultations n of an exemplary patient p.

Table 4.6: Previous treatment describing attributes comprising *therapy decisions* and associated *therapy outcome* which were ever applied previously to the target consultation n stored in the $N \times 6 \cdot M$ *Previous Outcome Matrix* $\mathbf{A}$ ($\mathbf{A}'$, $\mathbf{A}''$, see section 4.4.2). The outcome indicators' availability is given relative to the number of previously applied systemic therapy options.

Attribute	**Scale**	**Range**	$\mathbf{A}$	$\mathbf{A}'$	$\mathbf{A}''$
Systemic therapy	nominal	see table B.1	-	-	-
Effectiveness	ordinal	good, medium, bad	49.14	49.56	100
$\Delta PASI$	interval	-27 ... 18	12.19	23.71	100
$\Delta PASI_{rel}$	ratio	-1 ... 1	12.16	23.68	100
ADE	dichotomous	yes, no	100	100	100
Affinity score	ratio	0 ... 1	63.80	64.04	100

4.3.5 Data Representation

Independent of individual patients and chronological ordering of the N_p consultations of each patient $p \in P$, $\mathbf{X}^p$, $\mathbf{Y}^p$ and $\mathbf{A}^p$ can be concatenated to yield the three matrices $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{A}$ which hold the entire set of all N instances as pictured in figure 4.8. The $N \times D$ *Data Matrix* $\mathbf{X}$ comprises the D dimensional patient data summarized in table 4.4 for each instance n. In the $N \times 7 \cdot M$ *Outcome Matrix* $\mathbf{Y}$ the applied systemic treatment, the five outcome indicators and the information whether the treatment was changed, as summarized in table 4.5, are given. Finally, the $N \times 6 \cdot M$ *Previous Outcome Matrix* $\mathbf{A}$ stores for each of the M systemic treatment options the information whether it was previously applied and which outcome was observed according to table 4.6.

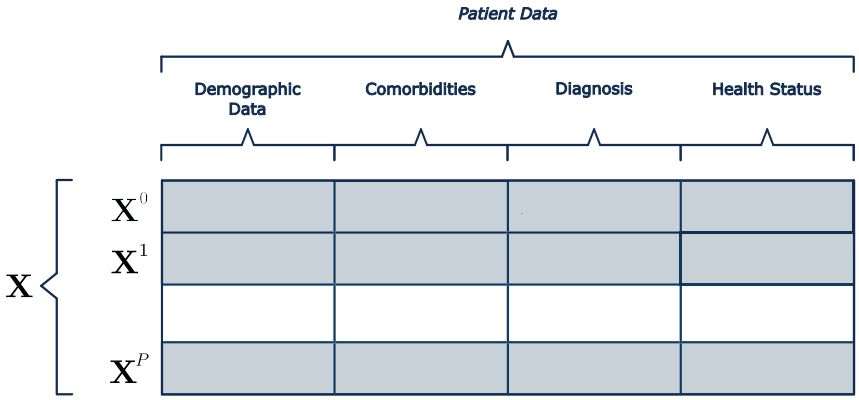


Figure 4.8: Concatenation of the P individual patient matrices $\mathbf{X}^p$ yielding the *Data Matrix* $\mathbf{X}$ ($\mathbf{X}'$, $\mathbf{X}''$, see section 4.4.2) which holds the entire set of all N instances. Concatenation of $\mathbf{Y}$ ($\mathbf{Y}'$, $\mathbf{Y}''$, see section 4.4.2) and $\mathbf{A}$ ($\mathbf{A}'$, $\mathbf{A}''$, see section 4.4.2) is done equivalently.

4.4 Data Preprocessing

As prescribed above and pictured in figure 4.3, initial data preprocessing was already conducted immediately after data collection before transferring the data to the MariaDB® database. This initial step mainly involved data cleaning, namely erasing empty instances as well as detection and correction of non-plausible or invalid entries, data padding where assumed to be valid, and data transformation into numerical values. In the following, further preprocessing steps are outlined, namely extracting outcome information as described in section 4.4.1 and imputing missing values as described in section 4.4.2.

4.4.1 Outcome Extraction

Firstly, in accordance with figure 4.4 and to yield the data as described in 4.3, for each consultation n the information whether the physician's therapy decisions were actually applied is extracted from the sequential data. Furthermore, the recorded outcome indicators are associated with the treatment applied in consultation n and the *affinity* scores are calculated according to section 4.3.3. A table storing this outcome information is added to the MariaDB® database and linked to the respective consultations.
Secondly, for each individual patient, the accumulation of outcomes of all treatments ever applied previously to the first recorded consultation of this patient is extended by therapies applied in the recorded consultation sequence. That means, outcome indicators observed in the consultation sequence preceding consultation n are added as described in section 4.3.4. Analogously to the consultation treatment outcome, each consultation's therapy history is stored in a table which is added to the database and linked to the respective consultations.

4.4.2 Missing Value Imputation

As stated in section 4.3.2, especially patient describing attributes contain numerous missing values. As further data processing can be limited by data sparsity, strategies need to be developed to cope with this incompleteness of attributes. In this work, the following assumptions are made regarding the patient describing attributes stored in the *Data Matrix* $\mathbf{X}$ which contain missing values.

- The dichotomous attribute *Living in partnership* and the nominal attributes *Skin changes* are considered as NMAR. Missing values are assumed to be equivalent with the negation of the respective attributes.
- The remaining attributes with missing values are considered as MCAR or MAR and are filled according to further assumptions detailed in the following.

Overall, a two stage imputation strategy is realized yielding a dataset with a reduced number of missing values $\mathbf{X}'$ and a complete dataset $\mathbf{X}''$. Imputation methods as introduced in section 3.4.2 are employed. Stage one ($\mathbf{X}'$) relies on domain knowledge and more reliable assumptions:

1. Impute NMAR attributes according assumption stated above (*Single value imputation*)
2. Discard attributes dropping below a minimum number of entries ($< 10\,\%$ availability)
3. Perform a sequential filling approach where appropriate (*LOCF*)
4. Impute attributes according to domain knowledge (*Single value imputation*)

Stage two applies more uncertain assumptions:

1. Apply statistical imputations (*Single value imputation*)
2. Apply statistical imputations conditional to other attributes (*Single value imputation*)
3. Replace the remaining missing values by a new category (*Missing indicator*)

The realized rules are described in table 4.7 and the resulting attributes' availability relative to all consultations are summarized in table 4.4, respectively.

Also outcome indicators for therapies, which are known to have been applied, are not always completely given but have missing values. As can be seen in table 4.5, especially $\Delta PASI$ and $\Delta PASI_{rel}$ contain many data gaps as a consequence of missing data in the PASI. The dichotomous ADE indicator was recorded using a checkbox and has no missing values. Padding of missing therapy attributes aims at yielding denser versions $\mathbf{Y}'$ and $\mathbf{Y}''$ of the *Outcome Matrix*. The number of missing $\Delta PASI$ and $\Delta PASI_{rel}$ values is directly affected by the described PASI imputation procedure and the missing values are already entirely eliminated in the first imputation stage. In consultation sequences in which not a single PASI value is present, which is not the case in the given data, missing $\Delta PASI = 0$ is imputed in the second stage. This directly effects $\Delta PASI_{rel}$ and corresponds to no PASI change.
For imputing missing *effectiveness* indicators, a LOCF strategy, analogously to PASI imputation, is utilized in stage one. In cases were the *effectiveness* indicator cannot already be completely

Table 4.7: Two stage imputation strategy to cope with missing values in $\mathbf{X}$, yielding the data matrices $\mathbf{X}'$ and $\mathbf{X}''$, respectively. The resulting attribute availability is summarized in table 4.4.

Attribute	**Imputation Strategy**
Stage 1:	
Living in partnership	Only the affirmative value is assumed to be actively entered, *not* living in partnership is imputed for missing values.
Education and profession	Dropped due to availability $< 10\,\%$.
Planned child	Women age $age > 50$: impute *postmenopausal* Men $age > 50$: Impute *no child planned*
PASI	Missing values are assumed to stem from omitted inputs due to unchanged values. 1. Fill forward: missing values in subsequent consultations are imputed with the last valid value. 2. fill backwards: missing values in the preceding consultations are imputed with the next valid value.
Skin changes	Missing values are assumed to correspond to *never occurred*.
Severity rated by patient	Dropped due to low availability and high correlation with PASI ($r = 0.65$).
Patient satisfaction with treatment	Dropped due to availability $< 10\,\%$.
Stage 2:	
Weight and size	For women and men impute median of group.
Planned child	An additional category *unknown* is imputed.
Year of first diagnosis	Imputed value is derived from the median age of the first diagnosis.
Psoriasis type	Fill with most common diagnosis, i.e. *psoriasis vulgaris*.
Family anamnesis	An additional category *unknown* is imputed.

padded in $\mathbf{Y}'$, the remaining missing values are replaced with the mean of the value range, i.e. *moderate* effectiveness, in the second imputation stage. The summarizing *affinity* parameter directly depends on $\Delta PASI_{rel}$ and *effectiveness* and does not require individual processing. The resulting availability of all indicators is summarized in table 4.5.
The realized rules are described in table 4.8 and the resulting availability of all indicators is summarized in table 4.5, respectively.

Finally, also in case of the accumulation of treatments applied previously to a patient and consultation, outcome indicators are frequently missing. Partly this data overlaps with the *Outcome Matrix* $\mathbf{Y}$. However, there are ~5000 additional applied treatments captured in the *Previous Outcome Matrix* $\mathbf{A}$ which were applied previously to the first consultation recorded for a patient p and for which no sequential information is available. Hence, the first stage imputation strategy proposed for the $\mathbf{Y}$ reduces missing values in $\mathbf{A}'$ just to a limited extend. To obtain a complete version $\mathbf{A}''$ of the *Previous Outcome Matrix*, a more generous second imputation level is applied analogously to $\mathbf{Y}''$. The resulting availability relative to the number of previously

applied treatments is summarized in table 4.6.

Table 4.8: Two stage imputation strategy to cope with missing values in $\mathbf{Y}$ and $\mathbf{A}$, yielding the data matrices $\mathbf{Y}'$ and $\mathbf{Y}''$ and $\mathbf{A}'$ and $\mathbf{A}''$, respectively. The resulting outcome indicator availabilities are summarized in table 4.5 and table 4.6.

Indicator	**Imputation Strategy**
Stage 1:	
Effectiveness	Missing values are assumed to stem from omitted inputs due to unchanged values. 1. Fill forward: missing values in subsequent consultations are imputed with the last valid value if the same treatment was applied. 2. fill backwards: missing values in the preceding consultations with same applied treatment are imputed with the next valid value.
$\Delta PASI$	Impute 0, which corresponds to no PASI change evoked by the applied treatment.
$\Delta PASI_{rel}$	Directly affected by $\Delta PASI$.
Stage 2:	
Effectiveness	Impute middle *effectiveness* value, i.e. moderate effectiveness.
$\Delta PASI$,	Impute 0, which corresponds to no PASI change.
$\Delta PASI_{rel}$	Directly affected by $\Delta PASI$. The remaining are imputed with 0.

4.5 Data Summary

In this section descriptive statistics, performed on all $N = 1242$ consultations and $P = 239$ patients, aim at giving insight into the data distributions of patient describing attributes, i.e. *patient data*, treatment describing attributes, i.e. *therapy decisions* and associated *therapy outcome*, and attributes describing previously applied therapies, i.e. *therapy history*. For attributes affected by data imputation, the effect of the applied data padding on the statistics is discussed. The number of consultations per patients varies strongly as can be seen in figure 4.9, ranging from 13 (5.43 %) patients with just a single consultation to a single patient with 16 consultations. 75 (31.38 %) patients are represented by 5 consultations in the provided data, which is the most frequent number of consultations for an individual patient.

4.5.1 Patient Describing Attributes

As depicted in figure 4.10 (a), the overall data comprises 135 (56.49 %) male and 104 (43.51 %) female patients with an age distribution as shown in figure 4.10 (b) with mean 56.27 years and a standard deviation of 15.10 years. The distribution of those demographic characteristics is in accordance with the prevalence reported in [302] and summarized in 4.1.

As can be seen in figure 4.11a, the clearly most common comorbidity present in the recorded data is *arterial hypertension*, which affects patients in 52.58 % of the recorded consultations. Furthermore, metabolic diseases, especially *hyperlipidemia* with 26.73 % and *diabetes mellitus*

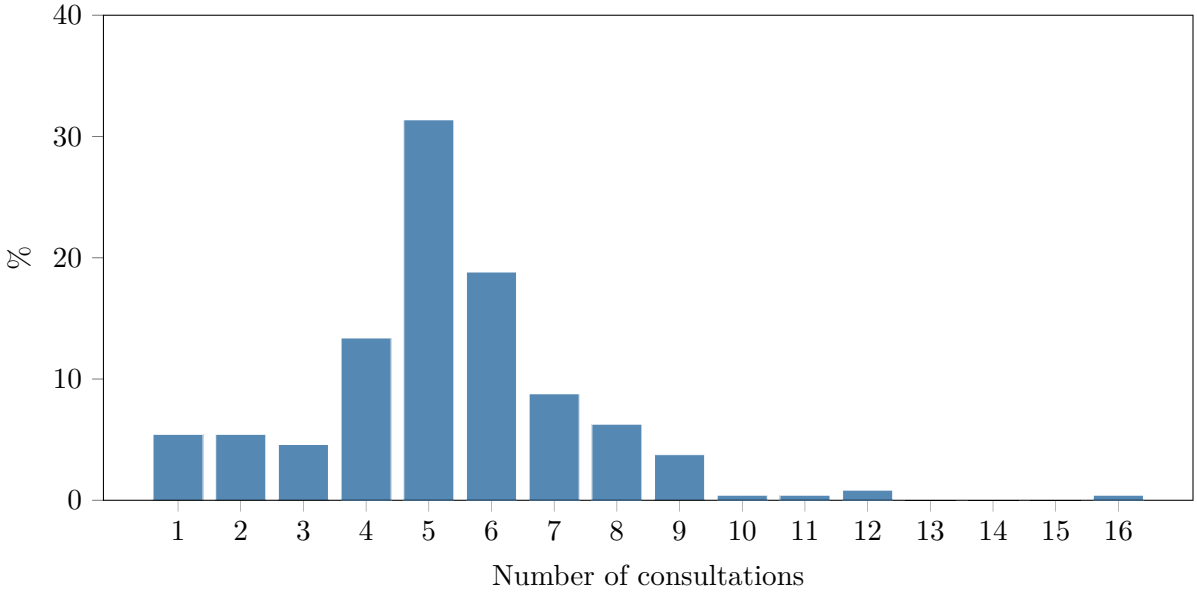


Figure 4.9: Relative number (%) of consultations per patient.

type 2 with 15.94 % of consultations, are an apparently widespread group of comorbidities occurring in patients suffering from Psoriasis. Also hepatic diseases, namely *hepatopathy* (14.81 %) belong to the most prevalent comorbidities in the dataset. Bowel or gastrointestinal diseases, on the other hand, even though stated to be associated with Psoriasis (4.1), are only rarely present in the data.
Furthermore, in 11.67 % of consultations psychological disease diagnoses are recorded with 6.52 % cases of depression. Additionally, figure 4.11a shows that also addictive diseases stand out. In 20.69 % of the consultations patient have stated to be smoker (21.65 %), ex-smoker (4.51 %) or are assumed to abuse alcohol (6.76 %). However, especially in cases of alcohol abusus and unspecified psychological diseases, diagnoses are often unclear as also show in figure 4.11a.

Figure 4.12 (a) shows the overall occurrence of *Psoriasis vulgaris* and its side-specific and acute forms *Psoriasis inversa* and *Psoriasis guttata*, *Psoriasis pustulosa* and *Psoriasis arthritis* relative to all 239 patients in the database. Besides *Psoriasis postulosa palmoplantaris*, no other *Psoriasis postulosa* cases are contained in the data. The shown distribution demonstrates the occurrence of diagnosed types relative to all patients in the dataset.
Only for 136 (56.90 %) from all 239 patients just a single Psoriasis type is diagnosed whereas 101 (42.26 %) patients suffer from more than one type (two types: 93 (38.91 %), three types: 8 (3.35 %). Additionally, there are also two patients for which the Psoriasis type is unknown.
The Psoriasis type distribution illustrates *Psoriasis vulgaris* clearly being the most prevalent type occurring in 209 (87.4 %) of all 239 patients followed by *Psoriasis arthritis* with 80 cases (33.47 %). However, whereas from the 209 *Psoriasis vulgaris* cases 113, which is 47.28 % of the entire dataset, are single *Psoriasis vulgaris* diagnoses, *Psoriasis arthritis* mainly occurs in combination with other Psoriasis types. Only three (1.26 %) cases with *Psoriasis arthritis* only are present in the data. Additionally, 90 patients, i.e. 37.66 % of all cases in the database, suffer

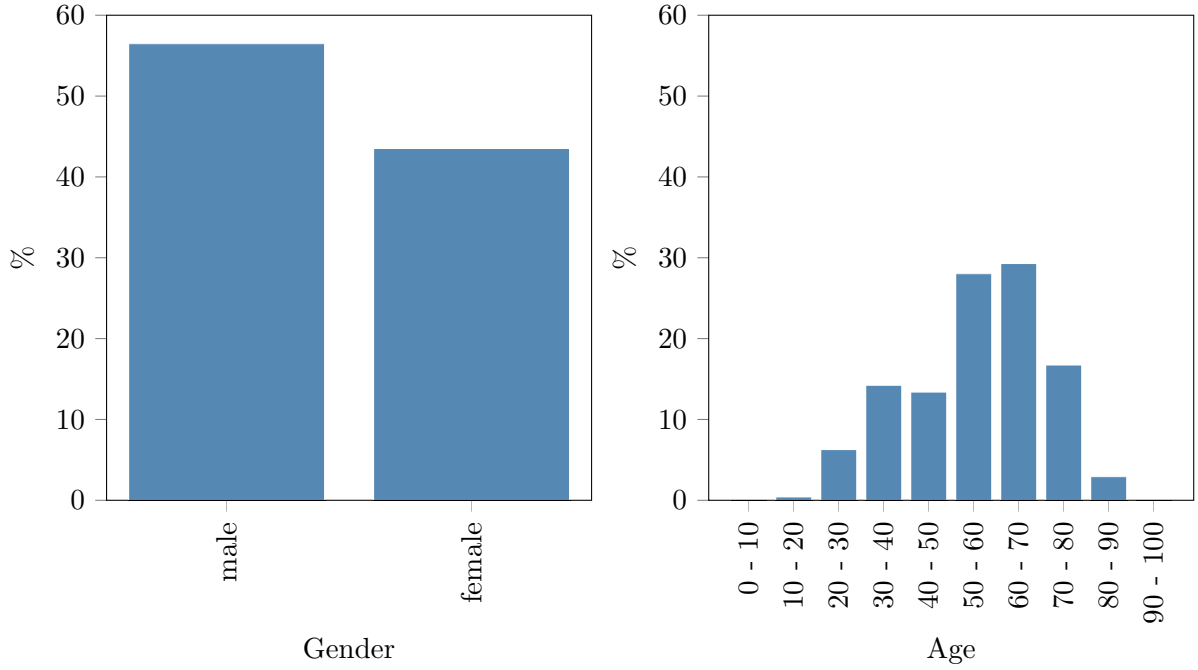
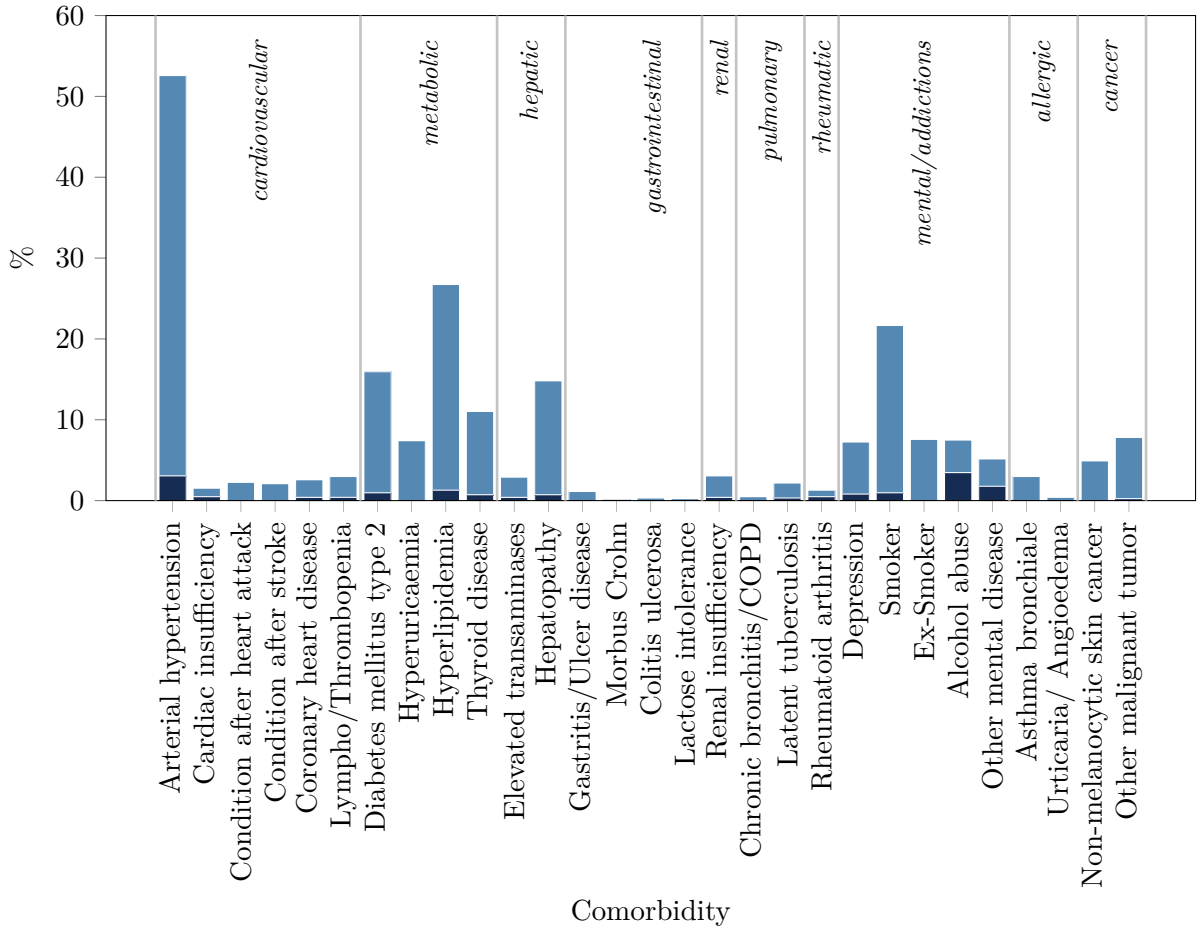


Figure 4.10: Relative gender (a) and age (b) distribution (%) over all patients.



(a) Relative comorbidity distribution (%) over all consultations with suspected (—) and confirmed diagnosis (—).

from additional disease related nail changes.
Figure 4.12 (b) shows the combinations of diagnoses. Most cases for which more than one Psoriasis type is diagnosed are combinations of *Psoriasis vulgaris* with other types. As can be seen, especially *Psoriasis arthritis* and *Psoriasis guttata* often occurs in combination with *Psoriasis vulgaris* in the datasets. In contrast, *Psoriasis postulosa palmoplantaris* is in most cases diagnoses without combinations with other types. Note that in figure 4.12 (b) single diagnoses and all combinations with a supplementary diagnosis are added cumulatively for each Psoriasis type. As a result, each of the 8 cases with combinations of three diagnosed Psoriasis types appears twice on the associated bars.
As described in section 4.4, the imputation strategy for missing Psoriasis type values is to fill data gaps with the most prevalent diagnosis, namely *Psoriasis vulgaris*. This approach has only minor impact on the overall diagnosis statistics. The number of patients in the dataset with a single Psoriasis type only are increased to 138 (57.74 %). And, consequently, also the overall number of *Psoriasis vulgaris* cases and the single *Psoriasis vulgaris* diagnoses are slightly increased to 211 (88.28 %) and 115 (48.12 %), respectively.

Figure 4.13a demonstrates the distribution of PASI over all consultations without and with missing data imputation. As can be seen, most consultations are associated with small PASI resulting in a long-tailed distribution with mode 2, mean PASI of 6.14 and standard deviation 6.47. Classified into severity categories according to [12], the distribution is comparable to the Psoriasis severity distribution indicated in [12] and listed in table 4.2 with a slight surplus of mild cases (80.87 %) compared to moderate (15.10 %) and severe cases (4.03 %). 28.02 % of PASI values are missing and are padded according to the imputation scheme described in section 4.4. The applied PASI imputation strategy has only minor impact on the PASI distribution. Especially the severity categories remain basically unchanged as also shown in figure 4.13a. Only the mode of the distribution is slightly shifted towards a PASI of 4.

4.5.2 Treatment Describing Attributes

In contrast to patient data, the sequence of therapy decisions, i.e. prescriptions or recommendations, and therapy outcome is often interrupted by missing entries. From all 1242 instances in the dataset 1108 consultations (89.21 %) are provided along with a systemic treatment recommendation or prescription, 857 consultations (69.00 %) are associated with an applied systemic treatment and 853 consultations (68.68 %) are associated with a known treatment outcome.

Concerning missing therapy decisions, two sets are distinguished. Those consultations which are only associated with a recommended or prescribed topical treatment (94) and those cases in which no treatment at all is prescribed or recommended (40), which account for 7.57 % and 3.22 % of all 1242 consultations, respectively.

Concerning consultations which are not associated with any systemic treatment application, four sets are distinguished. As a consequence of missing therapy decision described above, 94 consultations are only associated with an applied topical treatment (i) and 40 consultations are associated with no applied treatment at all (ii), accounting for 7.57 % and 3.22 % cases, respectively. Furthermore, lacking patient adherence, i.e. therapies prescribed or recommended by the

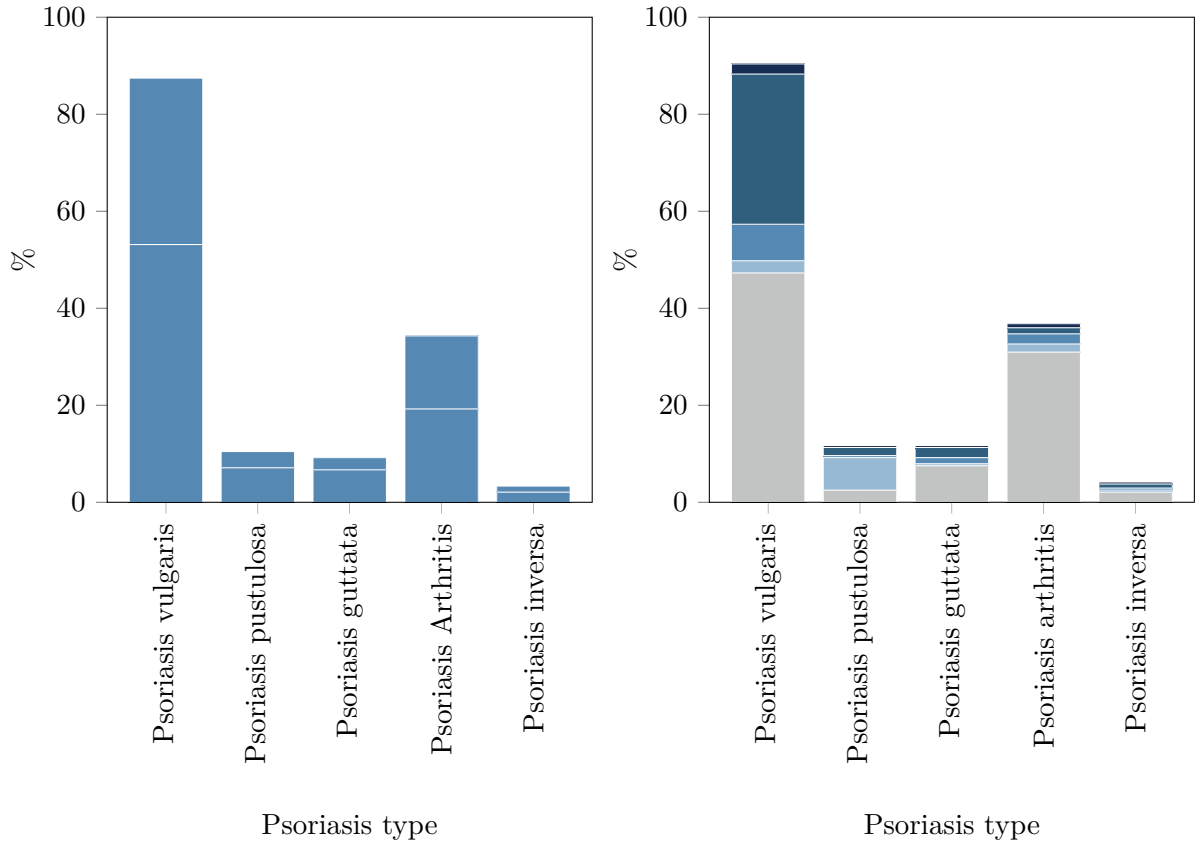
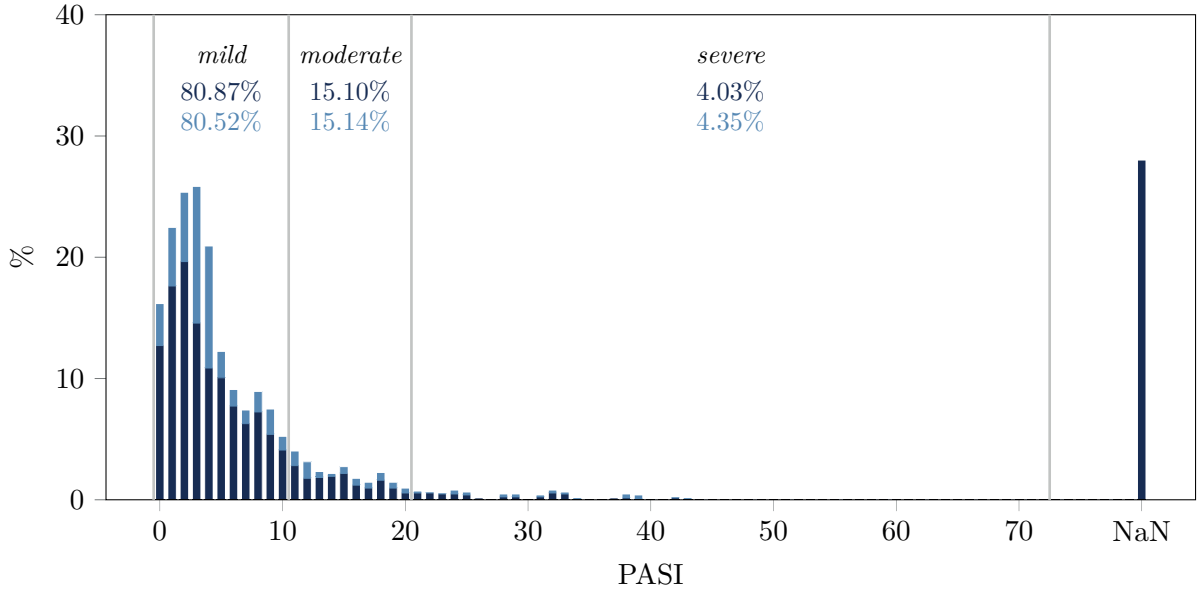


Figure 4.12: (a) Relative occurrence (%) of diagnosed Psoriasis types. (b) Relative occurrence of single diagnoses and combinations with supplementary Psoriasis types, i.e. Psoriasis vulgaris (—), Psoriasis postulosa palmoplantaris (—), Psoriasis guttata (—), Psoriasis arthritis (—), and Psoriasis inversa (—).



(a) PASI distribution (%) over consultations and categorization into severity levels [12] without (—) and with (—) missing value imputation.

attending physician but not applied by the patient account for 12 (0.97 %) additional instances with missing applied systemic treatment in the consultation sequence (iii). Finally, the last consultation of each of the 239 patients is associated with a systemic therapy prescription or recommendation but outcome is unknown yet, resulting in 239 (19.24 %) additional instances with missing outcome (iv).
Concerning consultations for which information on the outcome of systemic treatments is missing, five sets can be distinguished. For all 4 sets with missing systemic treatment application described above, accounting for 385 (31.00 %) cases, obviously no outcome information is available. However, in case of 4 additional consultations (0.32 %), recommended or prescribed systemic treatments are labeled as applied but none of the outcome indicators $\Delta PASI_{rel}$ or subjective *effectiveness* are given, and no ADE is reported resulting in the 853 consultations with known treatment outcome.

In figure 4.14 both therapy decision and application are shown as well as the 134 (10.79 %) consultations with no or topical treatment only. Therapy decisions comprise those therapies which were actually applied but also those 12 (0.97 %), which are not applied due to lacking patient adherence and the 239 (19.24 %) new consultations for which information on application or outcome is yet unknown. As can be seen, recommendation and prescription of the various systemic therapy options are distributed unequally. Frequency of prescribed or recommended therapies ranges from therapy combinations occurring very rarely, e.g. the Acitretin/Ustekimumab combination occurs only once in the available data, to treatment with Ustekimumab only, which is prescribed or recommended in 219, i.e. 25.55 % of the consultations. In general, the group of biopharmaceutical drugs clearly are recommended or prescribed more frequently than conventional pharmaceuticals or combinations of both. 538, i.e. 63.44 % of all recommended or prescribed drugs are biopharmaceuticals, whereas only in 209 (24.65 %) and 101 (11.91 %) cases conventional drugs or combinations of both were recommended or prescribed, respectively.

Figure B.4 (a) and figure B.1 show the patient's subjective assessment of effectiveness for applied treatments. *Effectiveness* is the most prevalent outcome indicator. Only for 11 consultations this indicator is missing. According to the overall distribution pictured in figure B.4, the majority of systemic therapies are considered to have good outcome and only declining quantities are considered moderately or badly effective, respectively. Even though the ratio of *good*, *moderate* and *bad* effectiveness values varies among the different therapies, as shown in figure B.1, it is hardly possible to identify a generally applicable tendency towards better or worse individual therapy options. However, the ratio of *good* therapies seems to be comparably higher for the group of biopharmaceuticals or drug combinations in comparison with conventional treatments.
Replacing missing *effectiveness* indicators by doing the forward and backward filling as proposed in 4.4.2, only negligible effect on the overall *effectiveness* distributions are observed due to the small number of affected instances.

Absolute and relative PASI change between two consecutive consultations, i.e. $\Delta PASI$ and $\Delta PASI_{rel}$, are the most absent outcome indicators due to irregularly recorded PASI. This indicator is missing for 360, i.e. 42 % of all consultations with applied therapies. As shown in

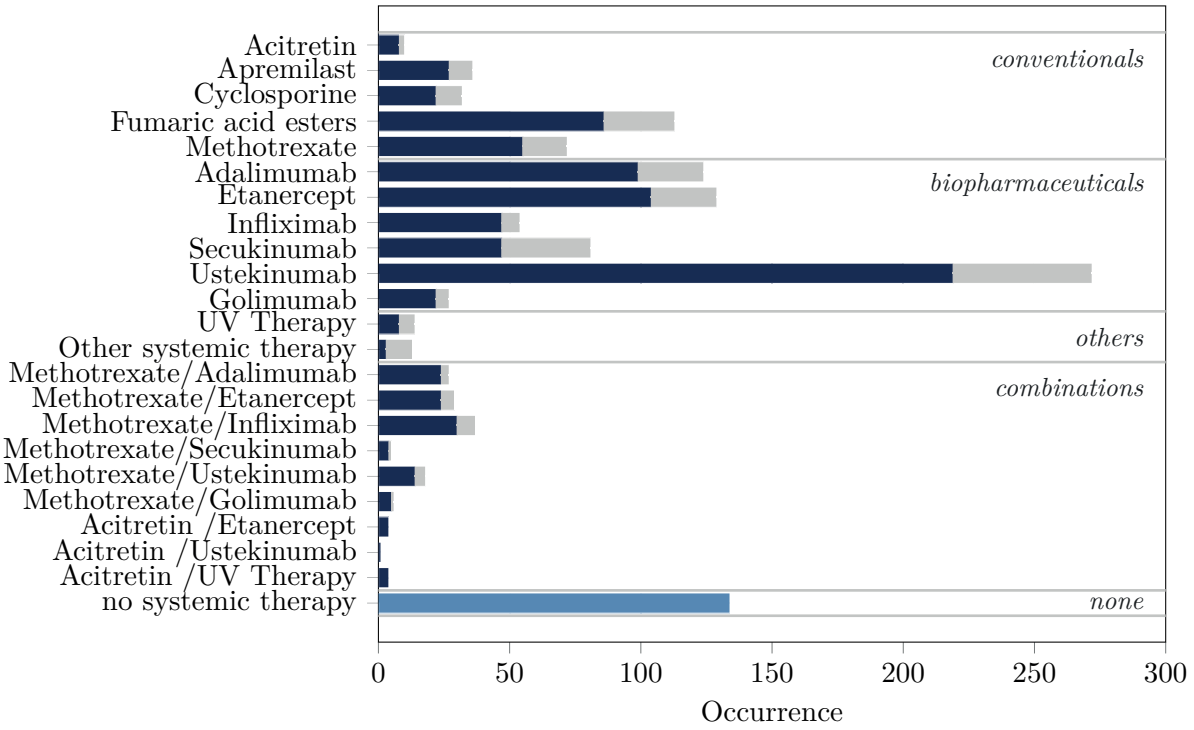


Figure 4.14: Therapy decision (—), i.e. recommended and prescribed therapies, application of therapy recommendations (—), and consultations with no or topical treatment only (—).

figure B.4 (left bars), overall $\Delta PASI_{rel}$ appears to follow a normal distribution which is slightly right-skewed. Whereas mode and median of the $\Delta PASI_{rel}$ distribution are $\Delta PASI_{rel} = 0$, the mean and standard deviation are $\Delta PASI_{rel} = 0.14$ and $\Delta PASI_{rel} = 0.47$, respectively. Consequently, there is a large portion of therapies which appear not to evoke an explicit improvement but also potentially prevent deterioration of the health status and control the disease. However, the overall number of cases in which the treatment appears to improve the PASI exceeds the number of those cases in which no effect leads to an decreasing PASI. This is not effecting the median of the distribution due to the large portion of cases with $\Delta PASI_{rel} = 0$ but causes a higher average value. The straightforward interpretation is that, comparable to the *effectiveness* indicator, the majority of the applied therapies can be associated with positive outcome. Considering small PASI fluctuations at low absolute values as controlled cases with good outcome, as described above, significantly reinforces this characteristic as demonstrated in figure B.4 (right bars). The distribution of $\Delta PASI_{rel}$ values over all systemic therapy options after this modification is shown in figure B.2. Comparable to the subjective *effectiveness* indicator, there are no outstanding therapy options concerning $\Delta PASI_{rel}$ improvement or deterioration.
By applying the forward and backward filling approach described in 4.4.2, the number of cases with $\Delta PASI = 0$ is significantly increased. This, in turn, increases the number of applied treatments which control the disease if small fluctuations are considered as controlled cases.

Computing *Spearman's rank correlation* coefficient r_s for therapy *effectiveness* and the $\Delta PASI_{rel}$

score, a only a weak association between the two outcome measures of $r_s = 0.32$ can be found. If considering small PASI fluctuations at low absolute values as good outcome, the monotonic relation between *effectiveness* and the $\Delta PASI_{rel}$ is significantly increased to $r_s = 0.67$, which is interpretable as a strong association. However, computing r_s for therapy *effectiveness* and the $\Delta PASI_{rel}$ score after missing PASI imputation ($\mathbf{Y}'$), this monotonic relation is decreased to $r_s = 0.46$. This reduction indicates that the chosen *single value imputation* does not necessarily represent the underlying data sufficiently but potentially introduces noise and the resulting data should be used with caution.

As can be seen in both, figure B.5 (a), the vast majority of therapies in the database don't provoke ADEs. Only for 53, i.e. 6.18 % of the applied pharmaceuticals ADEs were noted in the respective consultation's record. Analyzing the ADE distribution over the different therapy options shown in figure B.3, especially for some biopharmaceuticals, such as Infliximab or the most often recommended or prescribed treatment Ustekinumab, not a single case of ADE is reported. The conventional treatments, such as with fumaric acid ester, Apremilast, or Cyclosporine, seems to provoke comparably many ADEs.

As prescribed in 4.3, the *affinity* score summarizes the three aforementioned indicators in one single parameter and is intended to express the overall effect of a therapy. The overall distribution of *affinity* scores pictured in figure B.5 shows that most of the therapies applied in the captured consultations lead to high *affinity* scores, and thus have positive outcome. As mentioned above, in case of 5 consultations none of the required outcome indicators are given for the applied treatment and, according to the definition from 4.3.3, no *affinity* score can be computed for those treatments. The distribution of *affinity* scores over all systemic therapy options shown in figure 4.15 facilitates a more detailed insight into the overall effect of treatment options. As can be observed for the *effectiveness* indicator, the ratio of consultations with *affinity* ≥ 60, in the following considered as treatments with good outcome, is comparably lower for conventional treatments than for biopharmaceuticals. Especially, Cyclosporine, but also Ultra Violet (UV) therapy have only few occurrence with outcome in this *affinity* range. Furthermore, the combinations of Methotrexate with biopharmaceuticals shows a comparably large ratio of high *affinity* scores. In contrast, the very few Acitretin combinations show only moderate outcome. Overall, 607 (70.83 %) of the consultations with known applied treatment show good response according to the definition stated above.
Applying the forward and backward filling approach described in 4.4.2, missing *effectiveness* and $\Delta PASI_{rel}$ values can be eliminated and consequently *affinity* scores also computed for the 4 instances with missing *affinity* value. The imputation approach does not maintain the distribution of *affinity* scores but especially the number of cases with good outcome, i.e. *affinity* ≥ 60, is increased to 667 (77.83 %) consultations. Computing *Spearman's rank correlation* coefficient r_s for the *affinity* score and both indicators, *effectiveness* and $\Delta PASI_{rel}$, strong monotonic relations $r_s = 0.82$ and $r_s = 0.90$ can be found. As already observed for the correlation between *effectiveness* and $\Delta PASI_{rel}$, also correlation between the individual indicators and the *affinity* score is negatively affected by the missing value imputation strategies proposed in 4.4.2. After imputation, the $r_s = 0.81$ and $r_s = 0.77$ are yielded, respectively. As those values can

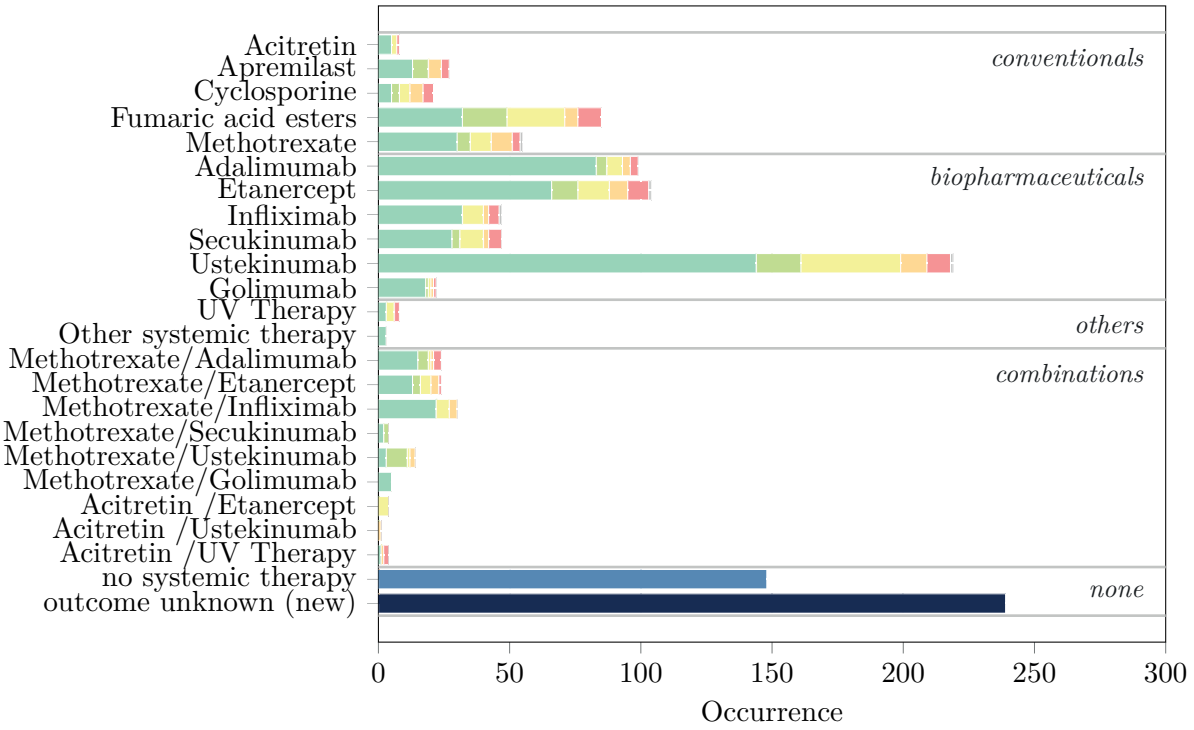


Figure 4.15: *Affinity* scores associated with applied therapies. *Affinity* scores range from good outcome (—) to bad outcome (—). Five consultations have missing *affinity* values (—). Additionally, consultations without systemic treatment (—), recommendations which were not applied, and new consultations without known outcome (—) are shown.

still be interpreted as strong correlation, the applied *single value imputation* strategy can be considered a valid approach regarding the summarizing *affinity* score.

In 122 cases, i.e. 14.24 % out of all 857 consultations which are associated with an applied systemic therapy, this treatment was changed compared to the preceding consultation. The information whether the therapy has been changed is only available for patients with more than one consultation in the database as this information is missing for the first consultation of each patient. Applied therapies are labeled as changed either if the previous systemic treatment is different (61) or no systemic treatment was applied (61). However, if interruptions between the application of the same systemic treatment for a patient are ignored, the overall number of treatment changes is further reduced to 113 cases (13.19 %). From those, 29 are initial systemic treatments, i.e. changes from topical to systemic treatments, and 84 are changes of the prescribed or recommended systemic treatment option. The *Sankey* diagrams pictured in figure 4.16 illustrates those 113 treatment changes. Here, some patterns can be observed which capture the prescription or recommendation schemes of the attending physician. Treatment with Methotrexate is always replaced by the second-line treatment Apremilast or supplemented with or replaced by a biopharmaceutical drug. Also Cyclosporine is in the majority of cases replaced by a biopharmaceutical drug. Treatments with Fumaric acid esters, on the other hand, is in the majority of cases replaced by other conventional treatment option instead of changing

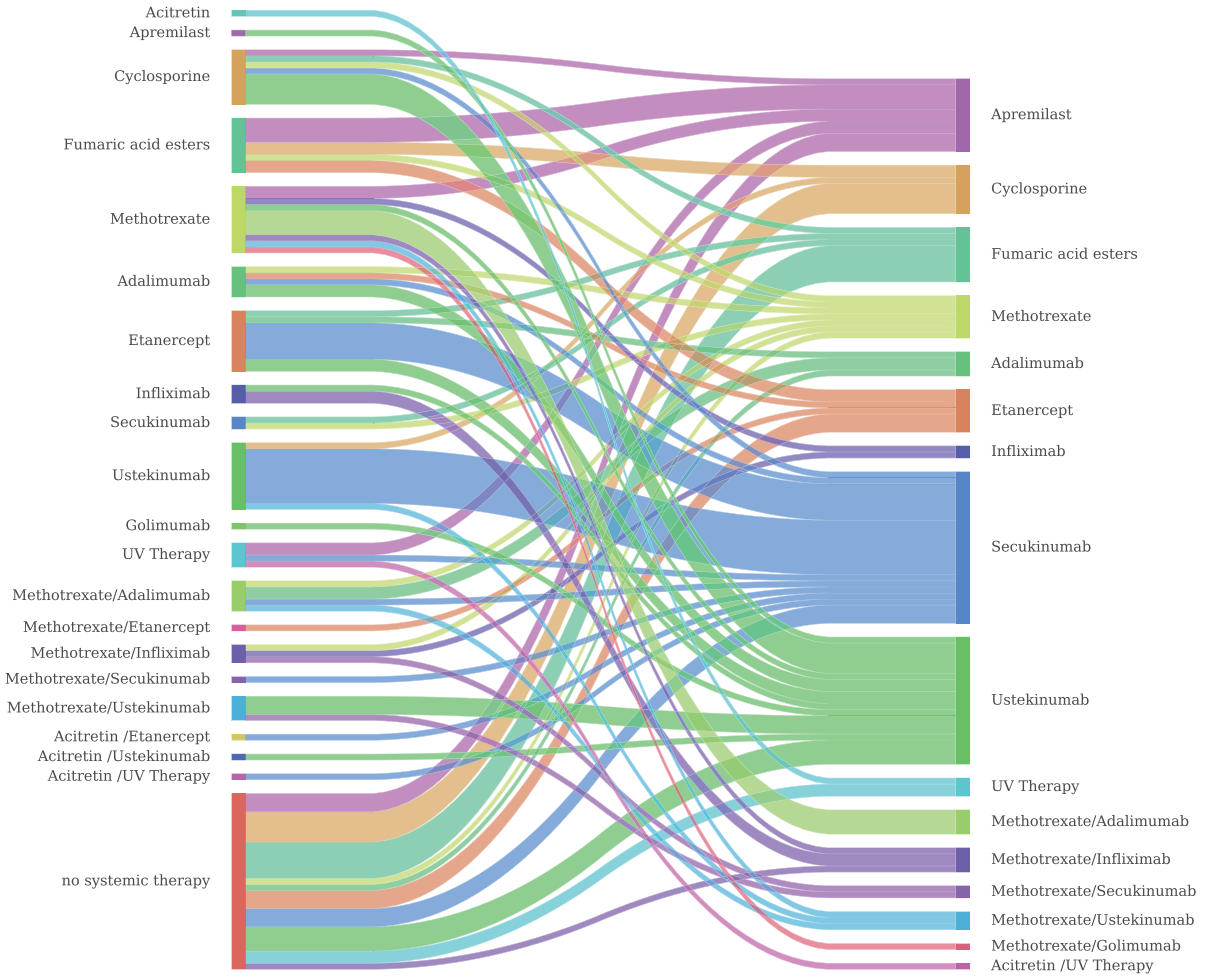


Figure 4.16: *Sankey* diagramm of treatment changes. Therapies are labeled as changed if the previous systemic treatment is different or no systemic treatment was applied. Interruptions between the application of the same systemic treatment for a patient are ignored.

to biopharmaceuticals. In most cases, Fumaric acid esters are replaced by the the second-line conventional treatment Apremilast. Acitretin and Apremilast are rarely changed, however, also only little represented in the data. Furthermore, all conventional treatments seem to have a preferred biopharmaceutical follow-up therapy as Ustekinumab for Cyclosporine, Etanercept for Fumaric ester acid and supplementation of Methothrexate with Adalimumab. There are only few cases in which a biopharmaceutical drug is replaced by a conventional drug. Regarding treatment with biopharmaceutical drugs only, Secukinumab clearly is a favorites when it comes to changing biopharmaceutical treatment. And, as also clearly can be seen, only very few cases exist in which treatment with Secukinumab was changed again. Interestingly, if it was changed, it was replaced by treatments with conventional pharmaceuticals. Finally, in case of drug combinations, the typical follow-up treatment is dropping the conventional supplement and treating with the biopharmaceutical only or changing to other combinations or biopharmaceutical drugs. The comparably large portion of cases where treatment was changed from no systemic treatment to the application of a systemic drug divides approximately evenly into cases with initial conventional first-line treatments and follow-up biopharmaceutical treatments prescriptions or recommendations.

4.5.3 Treatment History Attributes

As was already shown in table 4.6, the ratio of known previously applied therapies relative to all possible options for all consultations amounts to 5773 (21.13 %). Figure 4.17 shows the distribution of the number of previously applied treatments over all consultations in the dataset. As can be seen, the mode of the distribution is five previous treatments per consultation with a clearly positive skew (right skewed). In case of seven consultations no previous systemic treatment is known.

The number of known outcome indicators associated with those previously applied treatments relative to the number of applied treatments is comparable to the therapies associated with the recorded consultations. The *effectiveness* indicator is given for 2837 (49.14 %) of the previously applied treatments and $\Delta PASI$ and $\Delta PASI_{rel}$ indicators are given in only 812 (12.19 %) and 702 (12.16 %) cases. The low $\Delta PASI$ and $\Delta PASI_{rel}$ availability stem from the lacking PASI and temporal information for all treatments applied previously to the first consultation for a patient consultation sequence. $\Delta PASI$ and $\Delta PASI_{rel}$ are only available for treatments applied during one of the recorded consultations. The density of the *affinity* score is, with 3683 (63.80 %) cases, significantly higher, as *affinity* is also computed if ADEs were reported, as described in 4.3.3. The ratio of reported ADEs is comparably high in the previously applied treatments with 1269 (21.98 %) occurrences.
Figure 4.18 shows the distribution of the 5773 accumulated treatments applied previously to each consultation n along with the distribution of the available *affinity* scores.

When comparing the frequency and observed outcome (*affinity*) of treatment options applied previously and within consultations shown in figure 4.18 and figure 4.15, different distributions can be observed. Within the accumulation of previously applied therapies the number of conventional treatments clearly exceeds the number of biopharmaceutical drugs or combina-

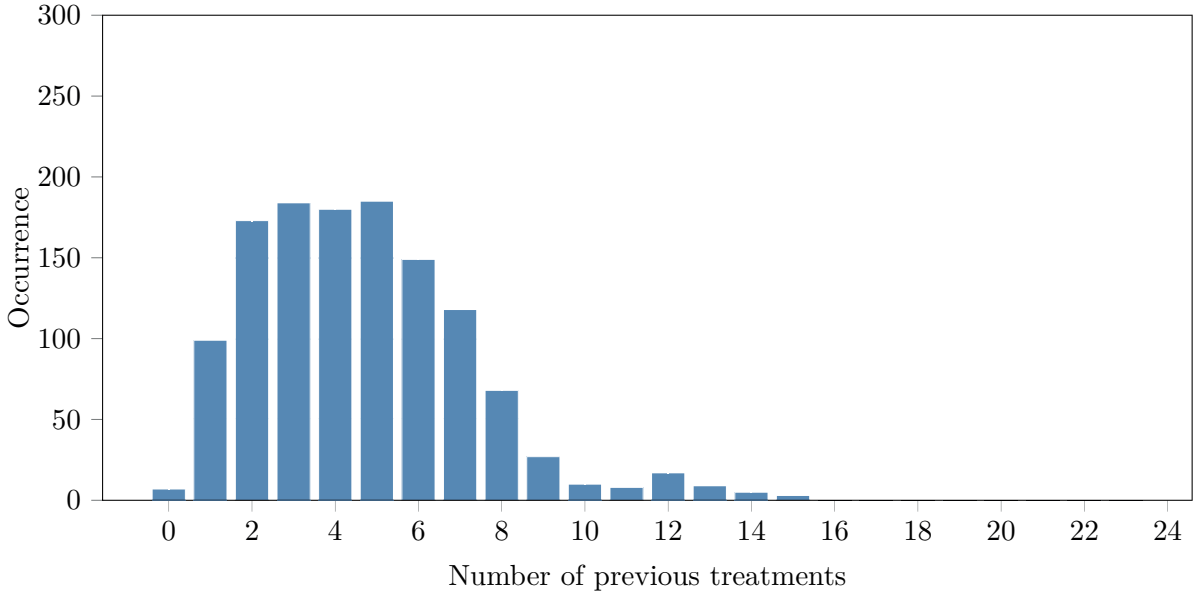


Figure 4.17: Distribution of number of known previously applied treatments over consultations in the dataset.

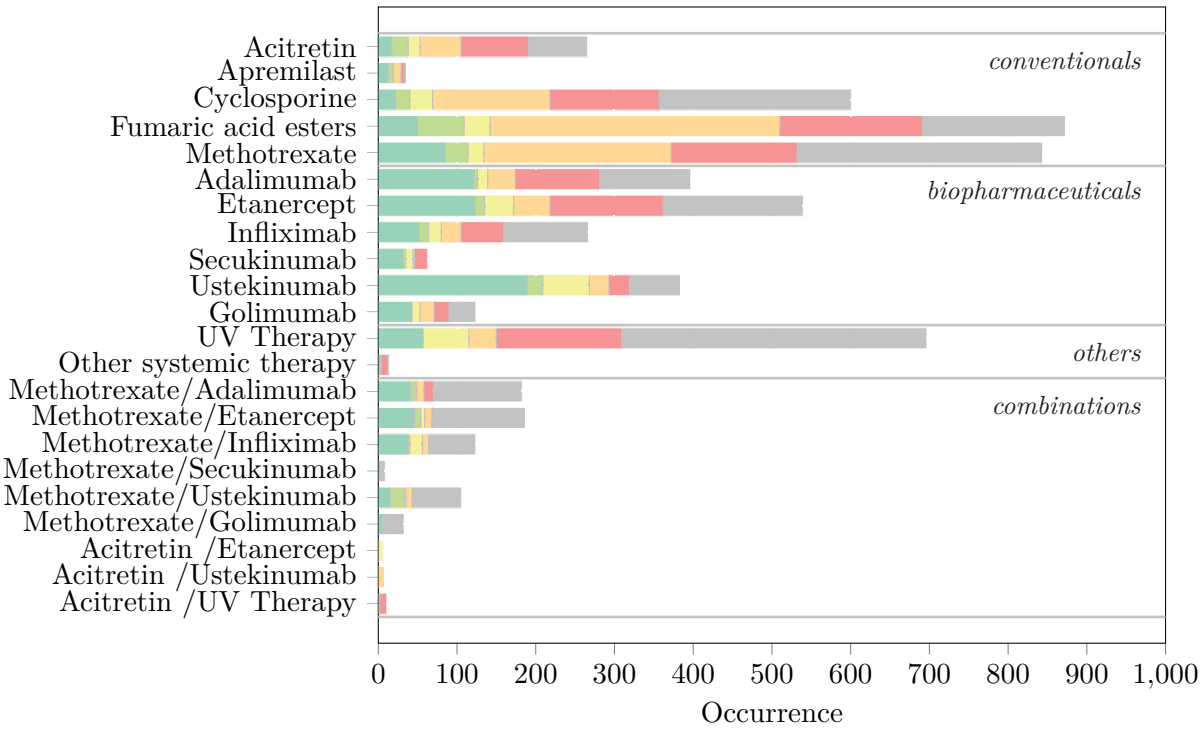


Figure 4.18: Affinity score distribution of previously applied therapies over consultations. *Affinity* scores range from good outcome (—) to bad outcome (—). For 24 consultations no previously applied treatments are known.

tions of both. Especially, Apremilast, the only second-line conventional drug, is rarely applied which indicates that from the group of second-line treatments, biopharmaceuticals are the pre-

ferred choice of the attending physicians. The different distributions additionally indicate that the largest portion of accumulated cases stem from treatment applications preceding the first consultation collected in the present data. Regarding observed outcome, conventional pharmaceutical drugs show in the majority of cases of previously applied treatments only moderate or bad outcome. Biopharmaceuticals and especially combinations of the conventional drug Methotrexate with biopharmaceuticals show comparably better response, however, come by the majority from recorded consultations. Nevertheless, for the largest portion of accumulated cases only moderate or bad outcome is observed which reflects the patients' disease history with inappropriate treatment. When comparing treatment frequency and response, it must be kept in mind that for patients which are represented in the data by multiple consultations, also the treatment history partly occurs multiple times in the data. Thus, those patients have larger impact on the shown distribution. Additionally, the distribution is influenced by the differing times when treatments were approved and hence entered routine patient care.

4.6 Study on Inter-Rater Reliability

For extension and validation of the provided ground truth, therapy recommendations are collected from six experts (dermatologists from different clinics in Germany) for a subset of 100 consultations from different patients. Additionally, the experts are asked to assess the retrospective therapy decisions if therapy recommendations disagree. The randomly selected subset comprises 74 consultations in which therapy was actually changed and 26 without change. A web-based survey tool was developed which presents consultations, represented by *patient data* and *therapy history*, to the expert and requests a treatment decision out of the in table 4.3 listed options. Therefore, up to three options are requested to be selected and prioritized. If none of the selected options meet the ground truth, the expert surveyed is asked to confirm whether the actually applied treatment is an acceptable alternative for him/her or not.

Out of the six experts, only four of them assessed all 100 test consultations, one assessed 50 and one 26 cases. These overall numbers are reflected by the number of at least one therapy recommendation, i.e. priority 1, for the respective consultation. The number of recommendations of the various priorities are summarized in figure 4.19 (a). The declining number of recommendations with decreasing priority can be interpreted such that each expert obviously has his/her favored therapy option for which he/she considers few or no alternatives. Figure 4.19 (b) shows the ratio of consultations for which the assessing expert's recommendations agree with the ground truth or not. Additionally, the ratio of cases in which the actually applied treatment, even though not selected by the survey expert, is still regarded as an acceptable alternative.

In the following, only expert 2, 4, 5, and 6 are considered who have assessed all 100 test consultations. To quantify inter-rater reliability, *Cohen's* [191] and *Fleiss'* [104] *Kappa* scores are computed. *Cohen's Kappa* between ground truth and the highest priority treatment option of each expert yields and average score of $\kappa_1 = 0.33$ with standard deviation of $\kappa_1 = 0.2$, whereas the largest agreement is shown for expert 6 with $\kappa_1 = 0.35$. Among the four experts having rated all 100 sample patients, the agreement between expert 4 and 5 is the largest with $\kappa = 0.39$. All

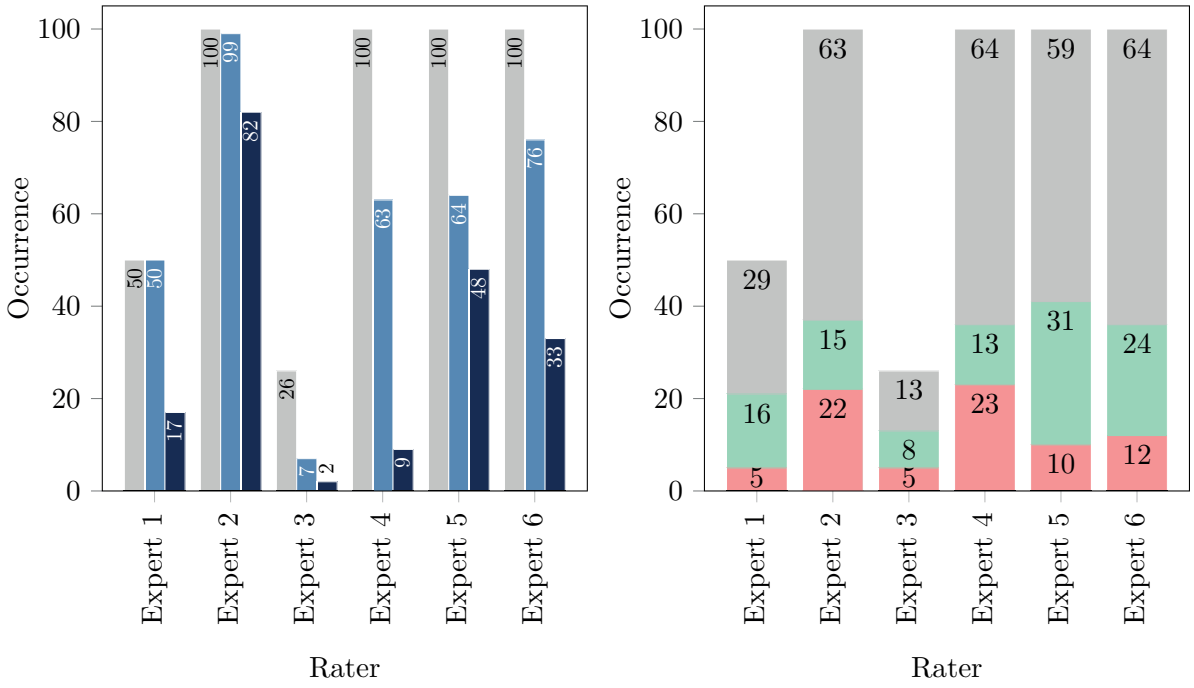


Figure 4.19: Number of recommendations of the various priorities 1 (—), 2 (—) and 3 (—) (a) and number of consultations for which the expert's recommendations agree with the ground truth (—), are considered an acceptable alternative (—) or not (—) (b).

those scores are considered *fair* agreement according to [191]. *Cohen's Kappa* between ground truth and either of the selections of each expert yields larger values and *moderate* agreement with $\kappa_{all} = 0.49(0.05)$ and the largest score $\kappa_{all} = 0.53$ for expert 6. *Fleiss' Kappa*, which assesses the overall agreement among all four experts and the actually applied treatment, attains *fair* agreement with $\kappa = 0.34$ if only the highest priority recommendations are included, but *moderate* agreement with $\kappa = 0.47$ if computed from all selections. It is further observed, that, with decreasing number of commonly assessed consultations but increasing number of raters, *Cohen's Kappa* scores also decrease whereas, however, *Fleiss' Kappa* increases.

According to the computed Kappa scores, the variety of therapies considered to be optimal is rather large and the agreement at most *moderate*. This finding reflects clear preference differences among the surveyed experts and the existence of many therapy options which are considered to be appropriate. This observation is additionally confirmed by the fact that the surveyed experts consider the actually applied therapy in at least 36 % (expert 4) of the presented cases an acceptable alternative even though not selected among the three most suitable treatments. The expert with the largest deviation from ground truth (41 %), expert 5, considers the actually applied therapy in 75.6 % of those cases still as acceptable alternative. In general, the largest deviations among therapy recommendations can be observed in cases where therapy history of the respective patient does not contain any previous systemic therapies or therapy was changed in this consultation. In contrast, large agreement can be observed in cases were a well functioning

therapy is continued.

On the one hand, this overall heterogeneity demonstrates the difficulty the decision maker is confronted with to select the optimal option and how subjective the selection is. On the other hand, the low inter-rater reliability also makes clear that the available ground truth should be considered with caution as there obviously exist alternative *hidden* ground truths which encode possibly more optimal treatment options.

5 Therapy Recommendation Algorithms

Initially, in section 5.1, the system to be developed is brought into context with the CDSS definitions from the literature described in section 2.1. Furthermore, the transfer of RS concepts into CDSSs is motivated and essential differences to their typical application are named. In section 5.2, the evaluation strategy is defined. Detailed descriptions of the applied algorithms are given in section 5.3, 5.4, and 5.5, respectively. In section 5.6, finally, the integration of evidence-based exclusion rules is specified. The described algorithms and results are published in [134, 136, 133].

5.1 Introduction

As already introduced in chapter 1 and with respect to the taxonomies described in section 2.1, the overall objective of this work is developing a CDSS which supports decision-making regarding the prescription of drugs [121] or ordering of treatments [386]. The clinical practitioner is intended to be provided with individualized and patient-specific therapy recommendations based on a *patient-data model* [318]. As the intention is to take detailed patient characteristics, such as diagnosed disease and health status, comorbidities and the patients' previous response to drugs into account, the system to be developed can be, according to [366], characterized as *medication-related* CDSS with *advanced* decision support. Furthermore, in accordance with [386], such a system can be categorized into the group of *expert system* CDSSs. However, whereas expert systems typically are defined to derive recommendations or suggestions using knowledge stored in rule sets (if-then rules), here a *non-knowledge-based* or *intelligent computing system* [29, 255] is to be developed, which is capable of extracting knowledge automatically from the available data.

As they are very successfully applied in other product recommendation settings, methods borrowed from the RSs domain are considered to be particularly appropriate. RSs typically aim at providing a target user with personalized recommendations by predicting his or her preference in order to derive a ranked list of items. Personalized therapy recommendations can be regarded as a comparable task considering patients as target users and the therapy options as items. However, with two essential differences:

(a) *Source of feedback*: RSs typically leverage implicit or explicit feedback to derive personal suggestions. In the therapy recommendation setting, generating such feedback is ideally not a subjective process but it is derived from (multifactorial) objective measures which quantify treatment outcome but can also encompass additional aspects, such as costs or applicability of a treatment options [49].

(b) *Stakeholder making decisions*: In contrast to traditional RS applications, in the therapy recommender system setting, the final choice of the product is not made by the user, i.e. the patient, but by the attending physician, however, ideally incorporating the patient's preferences.

According to section 3.2, five basic RS concepts are typically distinguished. As no attributes describing the actual items, i.e. drugs, are given, content-based approaches cannot be applied. Furthermore, as the objective is to develop a system which automatically extracts information from the available data and adapts to evolving datasets, knowledge-based approaches, depending on static domain knowledge, are also discarded. In this work, the concept of deriving recommendations by exploiting correlations among users (i.e. patients) is employed. Hence, in order to predict the adequacy and derive personalized recommendations, the transfer of CFs to CDSSs is studied. In this work, basically two neighborhood-based, i.e. memory-based methods (section 5.3), differing in the data used to represent a consultation, and two model-based CF approaches (section 5.4 and 5.5) are compared.
In the therapy recommendation setting, all those algorithms have in common to (1) predict outcome of the included therapy options and (2) rank the treatments according to this prediction. Suchlike, it is intended to recommend treatments regardless of general popularity or average efficiency, but rather a selection that is tailored to a target patient and consultation at hand. In order to take the multifactorial outcome aspects into account, the summarizing *affinity* score is employed to quantify treatment response. Furthermore, to incorporate evidence and reduce the risk of inappropriate recommendations, (3) exclusion rules, such as contraindications and recommendations regarding the sequence of treatments as described in the current S3-Guidelines [240] and specified by the *Clinic and Polyclinic for Dermatology, University Hospital Dresden*, are implemented in a post-filtering layer (section 5.6). As described in section 5.2, accuracy of the predicted outcome is evaluated by computing the RMSE between predicted and actually observed outcome (*affinity* score) and the quality of the ranked list of recommendations is assessed by computing the Mean Average Precision (MAP)@3. For model selection and system evaluation a *nested cross-validation* as described in section 5.2 is employed. Figure 5.1 shows the processing and evaluation chain for a recommendation query together with all inputs and associated outputs.

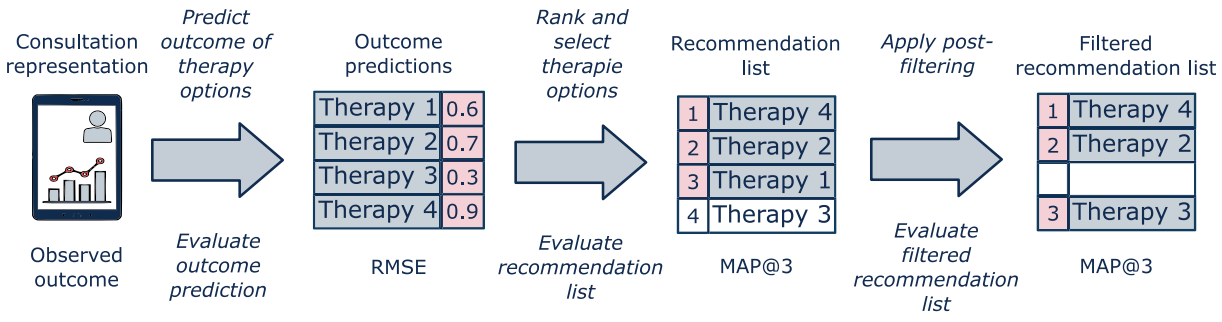


Figure 5.1: Therapy recommendation processing and evaluation pipeline.

As described in section 4.3, the number of available pharmaceutical drugs along with combina-

tions add up to $M = 22$ different treatment options. Some options are, however, just represented by very few occurrences in the available data. Neighborhood-based CF methods can be capable of coping with the so called *long tail* problem [258], i.e. a large proportion of items for which only little user feedback is available. Reliability, especially of the model-based CF algorithm, however, renders difficult for underrepresented therapy options. In order to facilitate comparability between the investigated algorithms, consultations associated with therapies with occurrences below a defined threshold (30 consultations) are neglected in the following. Here, a trade-off must be found between the amount of data which is discarded and a lower boundary of data representing a treatment option for appropriate modeling. Assuming a 5-fold cross-validation and stratified data partitioning, the chosen number guarantees on average a minimum number of 6 consultations for all included treatments and folds. Out of the $M = 22$ treatment options, the included therapies are the two conventional drugs Fumaric acid esters (86) and Methotrexate (54), the five biopharmaceutical drugs Adalimumab (99), Etanercept (103), Infliximab (46), Secukinumab (47), Ustekinumab (218), and the combination of Methotrexate and Infliximab (30). Figure 5.2 pictures the distribution of the included therapy options. In total, 687 (80.54 %) out of the 853 consultations and 181 (75.73 %) out of the 239 patients with known systemic treatment outcome are included into system development and evaluation.

As is shown in figure 4.3.5, the data of the overall N available consultations is organized in related matrices *data matrix* $\mathbf{X}$, *previous outcome matrix* $\mathbf{A}$, and *outcome matrix* $\mathbf{Y}$. Individual patients and chronological ordering of consultations are ignored and each consultation is considered as an independent instance. Based on these matrices, the *consultation representation matrix* $\tilde{\mathbf{X}}$, the *historic consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{hist}$, the *complete consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{all}$, and the *consultation outcome matrix* $\tilde{\mathbf{Y}}$ are derived as follows.
The *consultation representation matrix* $\tilde{\mathbf{X}}$ (figure 5.3) is the concatenation of the $N \times D$ *data matrix* $\mathbf{X}$ and the $N \times M$ submatrices which represent the attributes *effectiveness*, $\Delta PASI$, ADE, *affinity* score, and *therapy decisions* stored in the *previous outcome matrix* $\mathbf{A}$ for each of the M therapy options. The idea of extending the *data matrix* $\mathbf{X}$ is to additionally capture information on treatment history in the consultation representations. The selection of attributes, namely which previous outcome indicators to incorporate, is based on preliminary studies and are those which showed the highest increase in performance.
In the *historic consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{hist}$, the $N \times M$ submatrix of $\mathbf{A}$ is stored which represents the *affinity* scores of all previously applied treatments. The *complete consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{all}$ holds the $N \times M$ submatrix of *affinity* scores of $\mathbf{A}$, however, including the treatment and outcome applied in the current consultation n of a patient p which is stored in $\mathbf{Y}$. Thus, a vector $\tilde{\mathbf{a}}^{all}$ associated with a consultation n 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 (figure 4.4).
Finally, the *outcome matrix* $\tilde{\mathbf{Y}}$ holds the $N \times M$ *affinity* score submatrix of $\mathbf{Y}$.

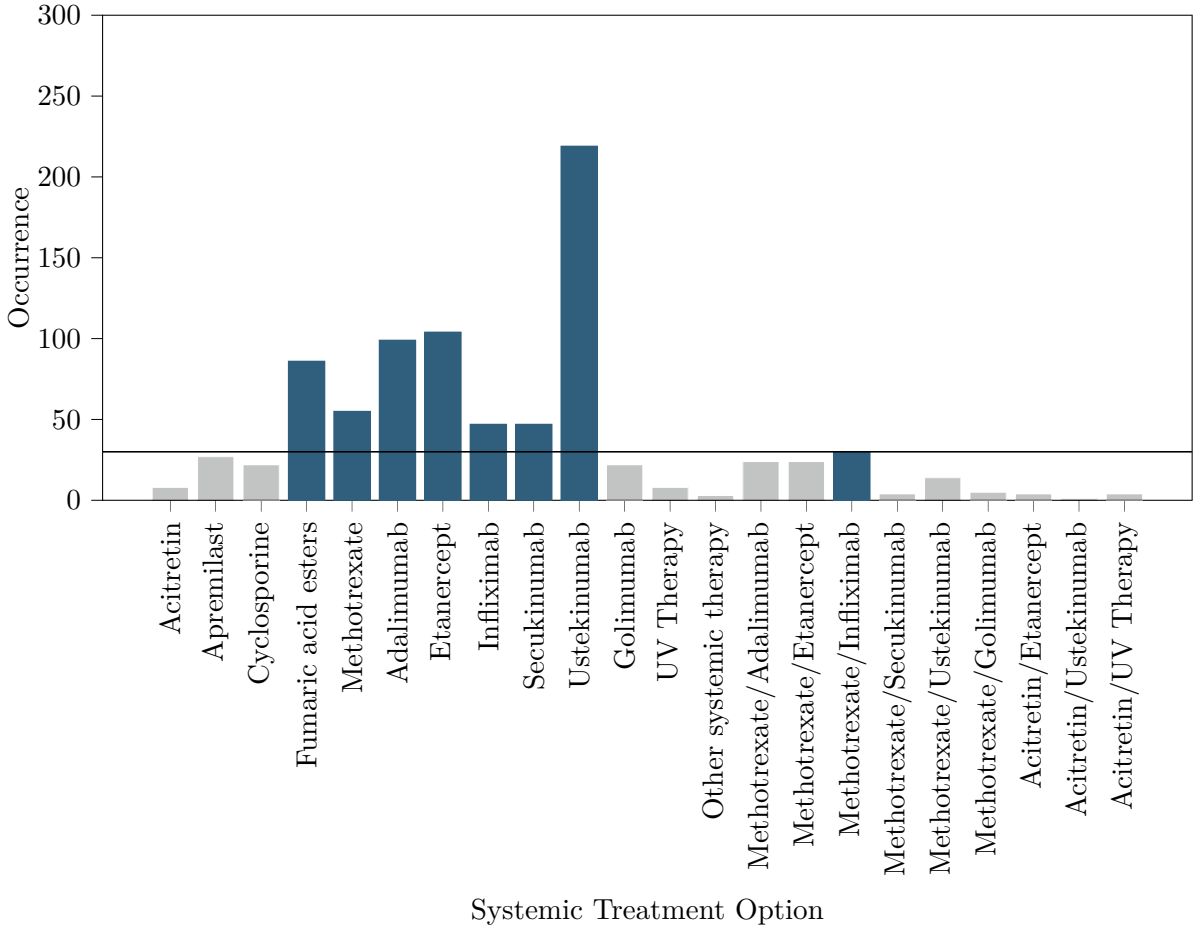


Figure 5.2: Overall occurrence of therapy options, applied threshold (30 consultations), and selected subset of options.

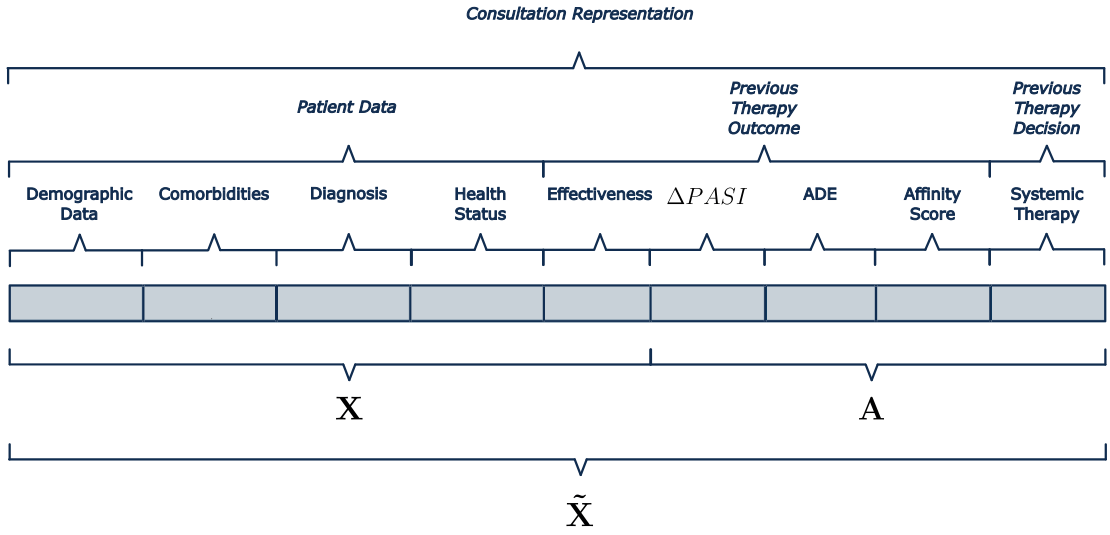


Figure 5.3: *Consultation representation matrix* $\tilde{\mathbf{X}}$, i.e. concatenation of *data matrix* $\mathbf{X}$ and selected outcome indicators from the *previous outcome matrix* $\mathbf{A}$.

Figure 5.4 visualizes the available input to the therapy recommendation system. The information actually used varies among the algorithms applied.

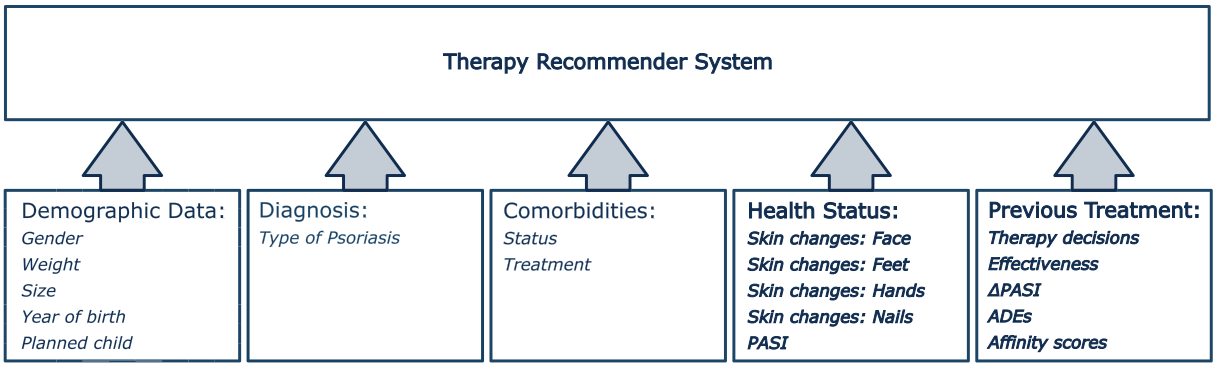


Figure 5.4: Therapy recommender system input. The information actually used varies among the applied algorithms.

5.2 Evaluation Strategy

As the temporal consultations of the individual patients cannot be regarded to be independent and identically distributed (i.i.d.), a patient-wise evaluation scheme is applied in this work. However, the application of a holdout method, which divides the limited number of 181 patients into a training and test partition, would potentially result in a very biased estimate of the generalization performance. The results would be highly dependent on the chosen partitions. An issue which is additionally enforced when holding out a third validation partition for model selection, i.e. hyperparameter tuning. Hence, 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 in his work for model selection and generalization performance estimation which was found to provide almost unbiased performance estimates [280, 364]. The realized approach is a nesting of two patient-wise cross-validation loops as pictured in figure 5.5 exemplarily for the *consultation representation matrix* $\tilde{\mathbf{X}}$. $\tilde{\mathbf{A}}^{hist}$, $\tilde{\mathbf{A}}^{all}$ and $\tilde{\mathbf{Y}}$ are partitioned in analogously.

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 the accuracy of the predicted outcome is evaluated by computing the RMSE between predicted and actually observed *affinity* score. Additionally, the MAP@3 assesses the generated ranked list of recommendations up to position 3. 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 5-fold 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 suchlike that in no iteration i the same

patient enters different folds. 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 5-fold cross-validation performance is calculated for all considered model variants (grid search) and the best performing model is selected.

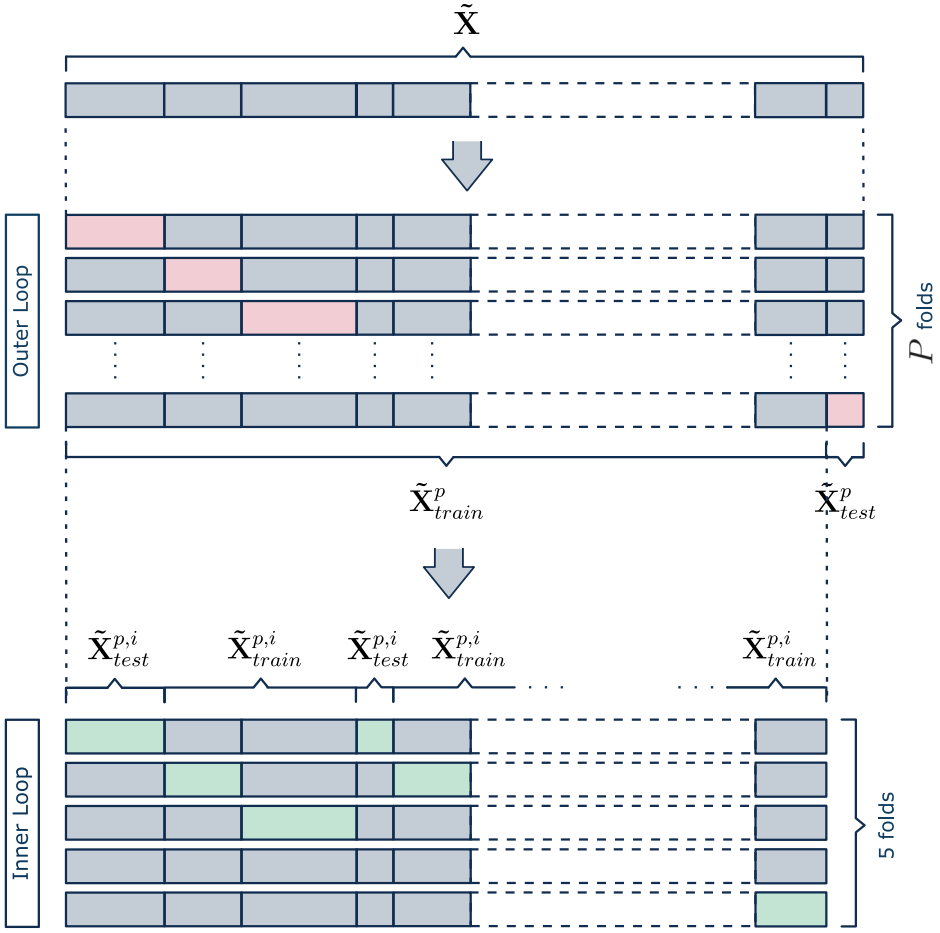


Figure 5.5: *Nested cross-validation* approach [280] for model selection and evaluation. The outer loop implements a patient-wise cross-validation over all $p \in P$ patients, the inner loop implements a 5-fold cross-validation, however, without mixing consultations of a patient p into test and training 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 in the identical way.

Finally, the performance of the most promising algorithms and system variants is compared with the recommendations of human experts. Therefore, the subset of 100 test consultations and dermatologists' recommendations described in section 4.6 are used. Comparable to the *nested cross-validation* from above, a patient-wise cross-validation scheme is employed. For each of the test consultations, an individual model, based on the remaining patients' consultations, is utilized within an outer validation loop. The respective model is optimized and selected on the basis of the inner 5-fold cross-validation results.

5.3 Collaborative Filtering

As stated above, the neighborhood-based CF approach is transferred to the application of therapy recommendation. In case of the user-based approach as applied in his work, the prediction of a rating $\hat{r}_{ui}$ of user u on an item i is based on the feedback on i of the subset of most similar users to u. In the therapy recommendation setting, consultations n 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. Suchlike, patterns in both, patient characteristics and response to previous treatments are intended to be revealed.

Each consultation k of a training partition is represented by a respective vector $\tilde{\mathbf{a}}^k$ from the *complete consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{all}$ and holds the *affinity* scores for all treatments ever applied to the respective patient up to and including this training consultation, as described above. In contrast, each test consultation n is represented by a test vector $\tilde{\mathbf{a}}^n_{test}$ which stems from the *historic consultation-therapy outcome matrix* $\tilde{\mathbf{A}}^{hist}$ and hence only holds the *affinity* scores of therapies applied up to the consultation under consideration. All $\tilde{\mathbf{a}}^k$ and $\tilde{\mathbf{a}}^n_{test}$ are aggregated in $\tilde{\mathbf{A}}_{train}$ and $\tilde{\mathbf{A}}_{test}$, respectively. The outcome matrices $\tilde{\mathbf{Y}}_{train}$ and $\tilde{\mathbf{Y}}_{test}$ hold the ground truths $\tilde{\mathbf{y}}^k$ and $\tilde{\mathbf{y}}^n_{test}$, which are the *affinity* scores of the systemic treatment actually applied in the respective consultations k and n. The objective is to predict the outcome of treatments in test consultation n. The subsets of training and test data $\tilde{\mathbf{A}}_{train}$, $\tilde{\mathbf{A}}_{test}$ and $\tilde{\mathbf{Y}}_{test}$ depend on the iterations p and i according to the evaluation strategy described in section 5.2.

When predicting outcome $\hat{y}^n_m$, an *affinity* score for each therapy option m is estimated for all test consultations n as visualized in figure 5.6 for treatment option m_1 and an exemplary n. As described in section 3.2.2.1, in the user-based CF approach this outcome prediction is regarded as a neighborhood-based regression problem and is computed as a linear combination of observed outcomes in the neighborhood of 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 $\hat{y}^n_m$ by computing the weighted average of all observed outcomes for each m according to

$$\hat{y}^n_m = \frac{\sum_{k=1}^{K} \tilde{a}^k_m \cdot s^{n,k}}{\sum_{k=1}^{K} |s^{n,k}|} \tag{5.1}$$

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 systemic treatment options for which an *affinity* prediction is available are ranked according to that prediction. The top-N ranked entries are recommended and evaluated regarding recommendation quality. If ties occur, i.e. the *affinity* score prediction of two therapy options equal, they are broken by recommending the more effective treatment according to the entire training partition.

To evaluate the accuracy of the predicted outcome, RMSE between predicted and actually observed outcome is computed as described in section 5.2. For each test consultation n, only one

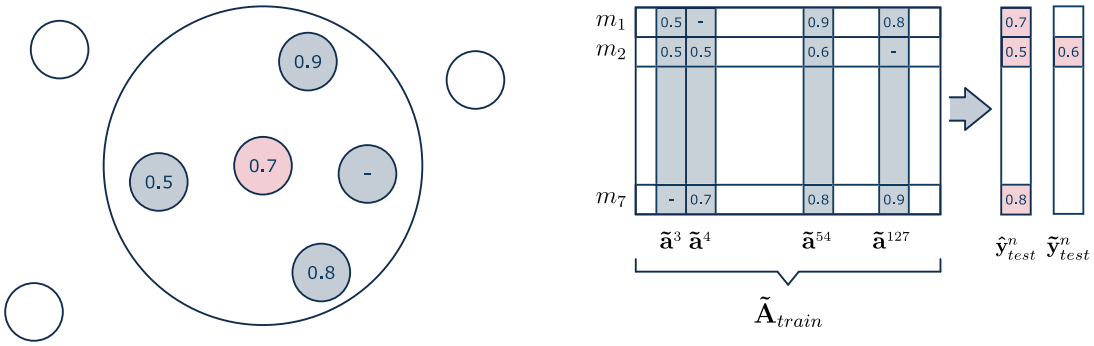


Figure 5.6: Outcome $\hat{y}^n_m$ of treatment option m_1 is estimated for a test consultation n by aggregating all outcomes observed for m_1 in the treatment history of the K most similar training data consultations. Therefore, the weighted average of all outcomes for m_1 observed in that neighborhood is computed. [133]

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. 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 and this overlap directly affects the MAP@3. 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 option. On the other hand, for small neighborhood sizes K, the *coverage* [152] of available treatment options can become very low which reduces the probability of recommendations overlapping with the actually applied treatment. Therefore, the ratio of neighbors from which RMSE can be computed (*overlap*) and ratio of overall recommended treatment options (*coverage*) are also monitored during evaluation. 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 $overlap > 95\,\%$ and (2) prediction accuracy (RMSE) is minimal.

The applied similarity measure $s^{n,k}$ to compare consultation representations, has crucial impact on the prediction results. In the following, various approaches are introduced and studied which differ in data and similarity measure utilized to define $s^{n,k}$.

5.3.1 Conventional Collaborative Recommender (CF)

Firstly, a *conventional* neighborhood-based CF approach is implemented. Consultations are compared based on the outcome of commonly applied previous therapies as pictured in figure 5.7. 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 which have not yet been applied to the respective patient are transferred to this consultation. This approach is comparable to recommending items based on users' ratings on previously purchased products. The underlying assumption 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.

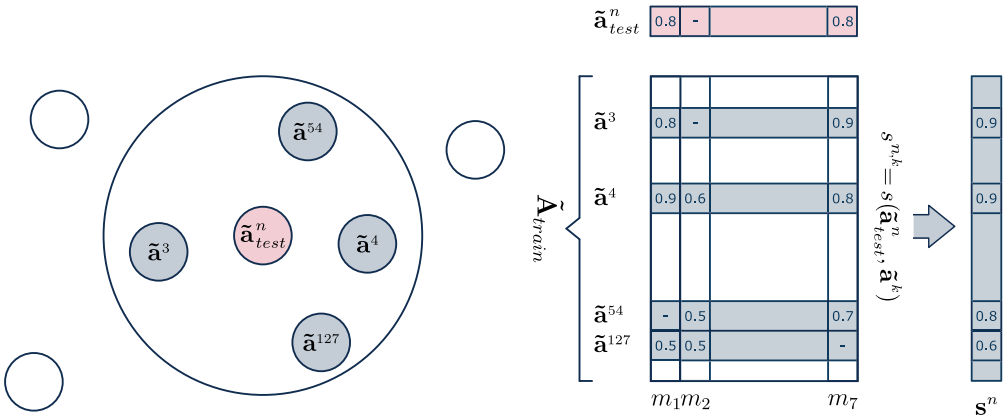


Figure 5.7: In the conventional CF approach, consultations are compared concerning treatment history stored in $\tilde{\mathbf{a}}^k$ and $\tilde{\mathbf{a}}^n_{test}$, respectively.

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 $\tilde{\mathbf{a}}^n_{test}$ and all training consultation representations $\tilde{\mathbf{a}}^k$. 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. However, consultation representations are rather sparse, meaning that outcome for treatment options is only intermittently available. On this account, similarity between two consultation representations is only computed between commonly available entries which does not impact the choice of similarity functions but can affect the reliability of the computed value.

As detailed in section 3.1, there are numerous functions suitable for computing pairwise similarity $s^{n,k}$ of such consultation representations. In the context of RS, especially *Cosine similarity* and the *Pearson correlation coefficient* are widely used, however, also the *Euclidean* or *Manhatten distance*, i.e. the *Minkowski metrics* with $p = 1$ and $p = 2$ are appropriate for similarity computation. All four metrics are considered in the model selection process. As the proposed CF algorithm is based on similarity measures $s^{n,k}$, the distance metrics *Euclidean* and *Manhatten distance* need to be converted to similarity measures. Here, a RBF, as introduced in section 3.1.2 is employed for that purpose.

As was shown in own preliminary studies, best performance could be achieved with the RBF spread parameter $\sigma = 0.25$. However, as was also found, σ has only minor influence on prediction and recommendation performance, which only becomes slightly apparent when k increases. This can be explained with the overall small distance for a wide range of k which in turn results in overall high similarity coefficients for a wide range of σ. Consequently, as similarity coefficients are very homogeneous, weighting of values contributing to the averaged score plays just

a minor role. The number K of neighboring consultations which are included into the computation, however, is crucial and needs to be chosen cautiously. It can be assumed that small K means low bias but high variance, while with increasing K, variance is decreased at the expense of increased bias. Thus, K is considered as a hyperparameter that is optimized in the model selection process.

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 vectors. This, in turn, can impact accuracy of recommendations as was shown in other CF applications [152, 27]. To overcome such reliability issues, *significance weighting* [152] or *shrinkage* [27] were proposed, which penalize the similarity measure by taking the number of mutually rated items into account. In this therapy recommendation setting, however, applying such methods did not increase performance according to own preliminary studies.

5.3.2 Patient-data Collaborative Recommender (DR)

To overcome the *cold start* problem, the stated reliability issues and to make use of the additional, presumably meaningful information in the patient data, the described conventional CF is extended to a hybrid *patient-data* approach (see section 3.2.3). Information on previous treatment outcome is combined with available patient data for similarity computation.

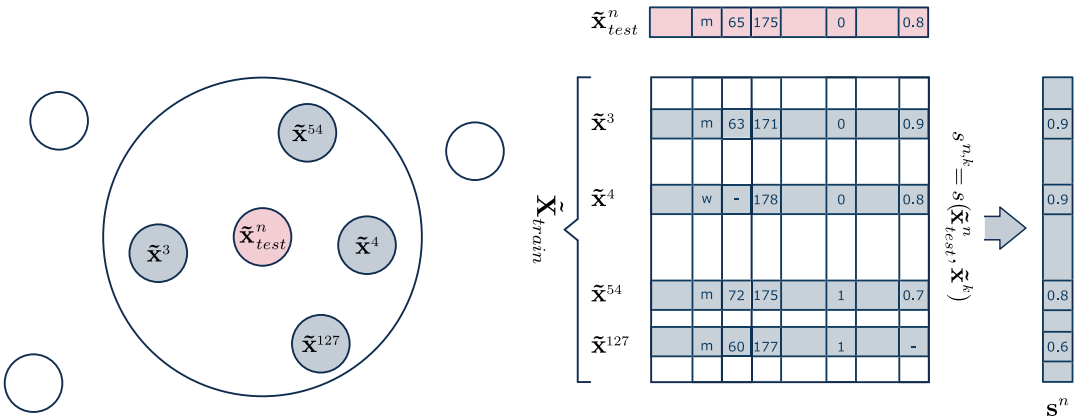


Figure 5.8: In the patient-data CF approach, consultations are compared concerning patient data and treatment history stored in $\tilde{\mathbf{x}}^k$ and $\tilde{\mathbf{x}}^n_{test}$, respectively.

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. As described in section 5.1 and visualized in figure 5.3, $\tilde{\mathbf{X}}$ 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 5.8 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 different levels of measurement. Thus, the similarity function applied in section 5.3.1 to determine $s^{n,k}$ are not appropriate

for the concatenated data. Furthermore, as described in chapter 4 and listed in table 4.4, also the patient data vectors $\tilde{\mathbf{x}}$ comprise, dependent on the applied imputation strategy, frequently missing values. As mentioned in section 3.1.4, the *Gower similarity coefficient* measures similarity at the presence of mixed data types and 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

$$s^{n,k} = \frac{\sum_{d=1}^{D} \delta_d^{n,k} \cdot w_d \cdot \rho_d^{n,k}}{\sum_{d=1}^{D} \delta_d^{n,k} \cdot w_d}. \tag{5.2}$$

$\rho^{n,k}_d$ 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. Hence, with increasing number of missing values, the overall similarity measure $s^{n,k}$ is based on comparison between fewer attributes and the reliability of $s^{n,k}$ suffers. Therefore, when evaluating the proposed patient-data CF approach, the impact of the missing values in the patient data is investigated. For this purpose, the originally provided data and the two proposed levels of imputation are compared.

Furthermore, also the *Euclidean distance*, i.e. the *Minkowski metric* with $p = 1$ can be employed do derive a similarity function $s_{Euc}(\tilde{\mathbf{x}}^n_{test}, \tilde{\mathbf{x}}^k)$ using a RBF as introduced in section 3.1.2. The spread parameter has shown to achieve good results if set to $\sigma = 0.25$, but has only minor influence on prediction and recommendation performance. Prerequisite for computing the *Minkowski metric* are all attributes in the attribute space having a quantitative data type which allows for pairwise attribute subtraction. Hence, qualitative attributes must be converted to a quantitative 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 are transformed to the interval scale analogously to the *Gower similarity coefficient*. As a result of this attribute preprocessing strategy, the dimension of the consultation representation $\tilde{\mathbf{X}}$ is further expanded to $D = 165$ 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

$$s^{n,k} = K_\sigma \left(\sqrt{\frac{\sum_{d=1}^{D} \delta_d^{n,k} \cdot (\tilde{\mathbf{x}}^k_d - \tilde{\mathbf{x}}^n_{test,d})^2}{\sum_{d=1}^{D} \delta_d^{n,k}}} \right) \tag{5.3}$$

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

The number of comparable attributes is significantly increased in the patient-data CF approach which is prone to reinforce the problems associated with the *curse of dimensionality*. According to own preliminary experiments, however, dimensionality reduction methods such as MF (see section F.2), did not prove to be beneficial for the problem at hand. The performance regarding prediction accuracy or recommendation quality could not be improved by applying such methods.

5.3.2.1 Attribute Selection and Weighting (DR-RBA)

As described in section 3.1.5, individual attributes typically are of varying importance concerning the similarity $s^{n,k}$ between two user or patient representations. Attributes can even be entirely irrelevant or redundant, introduce noise and worsen the informative value of the similarity coefficient. Moreover, the *curse of dimensionality* 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. Furthermore, low data dimensions bear the additional potential to reveal insights into the factors determining therapy outcome and to lower computational complexity and required storage.

Consequentially, it is an obvious strategy to modify the above proposed patient-data CF approach in order weight the individual attributes according to their relevance (attribute weighting) and to remove irrelevant ones (attribute selection) before computing similarity. As described in section 3.1.4 and applied in section 5.3.2, the *Gower similarity coefficient* facilitates to assign weights w_d to each attribute when computing similarity. Accordingly, weights for attributes to be discarded are set to 0 and scaled in accordance with the estimated attribute relevance otherwise. Subsequently, the local neighborhood of a target consultation can be computed and outcome can be estimated for available treatment options as described in section 5.3. The unweighted version of the *Gower similarity coefficient* is considered as baseline in the following.

Selection or adjustment of attribute weights can either be based on *a priori* knowledge about attributes or attribute importance is extracted automatically from the given data. Wrapper methods are powerful tools regarding automatic attribute selection by determining the performance of an attribute subset (section 3.1.5). For assigning numeric weights to individual attributes which appropriately reflect their importance regarding a given classification or regression task, however, filter methods are more suitable. As wrapper methods, many filter-based attribute weighting and selection algorithms rely on supervised information such as known class labels or dependent variables of some training data. Based on such information, univariate or multivariate criteria can be defined which measure importance directly, e.g. from the correlation between attributes and the given class label or dependent variable. In the proposed patient-data CF approach, however, heuristic similarity measures are used to (1) identify relevant samples, i.e. consultation representations from the data basis and to (2) define the impact of those individual samples on the outcome prediction. It is therefore an obvious approach to incorporate *a priori*

assumptions concerning similarity and dissimilarity between consultation representations and determine attribute weights suchlike 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 by [171] and, among others, extended by [177]. In contrast to most other filter approaches, especially RBAs are supposed to have the potential to even take interactions and dependencies among attributes into account [363]. In this work, a generalization of the mentioned algorithms is adapted to the given patient-data CF approach (algorithm 1). The attribute weights are determined for each outer cross-validation loop using the training sets $\tilde{\mathbf{X}}^{p}_{train}$. Within an iterative process a random sample, i.e. the target consultation $\tilde{\mathbf{x}}^{j}$ is drawn from $\tilde{\mathbf{X}}^{p}_{train}$ and, based on this sample, each dimension w_d of an attribute weight vector $\mathbf{w}$ is updated according to

$$w_d = w_d + (\bar{\rho}_d^{hits} - \bar{\rho}_d^{misses})/J \tag{5.4}$$

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* ($\tilde{\mathbf{X}}^{j,hits}_{train}$) and the K_{RBA} nearest neighbors with different class, i.e. *nearest misses* ($\tilde{\mathbf{X}}^{j,misses}_{train}$). 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, yielding weights w_d in the interval $[-1, 1]$ when w_d is initialized with zeros. The underlying assumption is that attributes with large average similarities $\bar{\rho}_d^{hits}$ between target $\tilde{\mathbf{x}}^{j}$ and nearest hits $\tilde{\mathbf{X}}^{j,hits}_{train}$ bear meaningful information regarding the class label. Thus, w_d is increased depending on that similarity. Conversely, attributes with large average similarities between target $\tilde{\mathbf{x}}^{j}$ and nearest misses $\tilde{\mathbf{X}}^{j,misses}_{train}$, i.e. large $\bar{\rho}_d^{misses}$, are assumed to be not informative regarding the class label and thus w_d is decreased depending on that similarity. Here, 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. The unweighted *Gower similarity coefficient* is also used to initially find the nearest neighbors, i.e. *nearest hits* and *nearest misses*. In contrast to the algorithms proposed in [171] and [177], where the attribute weight vector w_d is initialized with zeros, here a variable initial value in the interval $[0, 1]$ can be chosen. After iterating over all J randomly selected observations the resulting weight vector w_d is normalized by the sum of all weights. As proposed in [171], all attribute weights dropping below a predefined *relevance threshold* thr_w, are discarded. That means, only attributes are selected for which applies $w_d \geq thr_w$. 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 total number of iterations J is parameterized suchlike that each observation is selected once.

The typical RBA assumes a supervised classification problem where each sample is associated with a distinct class. In the context of the present therapy RS problem, each sample, i.e. consultation, is characterized by a quantitative 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* $\tilde{\mathbf{X}}_{train}^{j,hits}$ to a target consultation $\tilde{\mathbf{x}}^j$ are the K_{RBA} closest consultations associated with equal therapy and similar response, whereas *nearest misses* $\tilde{\mathbf{X}}_{train}^{j,misses}$ 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 5.9 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 $\tilde{\mathbf{a}}^k$ 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.

As described in section 3.1, with growing dimensionality the meaning of distance or similarity looses validity and hence the determination of neighbors becomes increasingly random. Consequently, also the performance of RBAs has been shown to deteriorate with a growing number of attributes and learned weights become increasingly unreliable. Furthermore, RBAs define the nearest neighbors in the original unweighted attribute space which is highly unlikely to be the same in weighted space [339]. To address both stated issues, several works suggest iterative

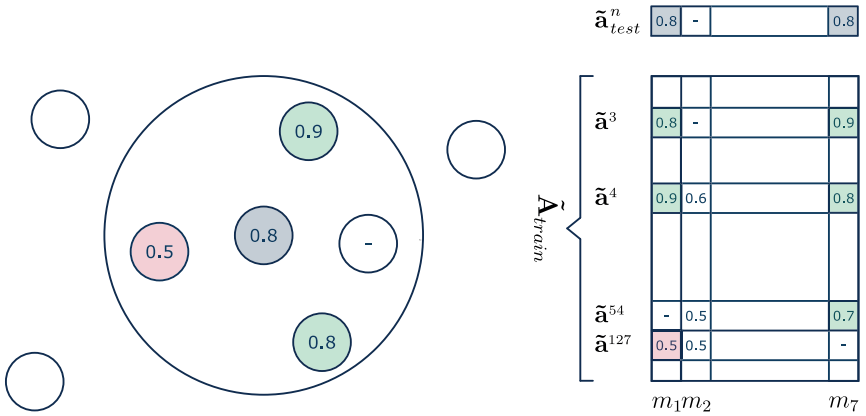


Figure 5.9: The RBA algorithm assumes a supervised classification problem. Consultations in the neighborhood of a target consultation j are labeled as similar or dissimilar if the same treatment was applied, here treatment option m_1, and the outcome polarity is equal (green) or different (red), respectively. No information regarding the similarity label is available for consultations were the treatment from consultation j was never applied (white).

RBA approaches, which run the attribute weighting algorithms multiple times. In each iteration the attribute weights $\mathbf{w}$ from the previous iteration are used to compute the neighborhoods of target observations $\tilde{\mathbf{x}}^j_{train}$. Suchlike, low scoring attributes from the previous iteration have less influence on the similarity computation in the subsequent iteration or are completely discarded [93]. However, the iterative RBA approach did not show any performance improvements in the present problem in preliminary studies.

The general advantage of attribute weighting and selection approaches over dimensionality reduction approaches is their capability to maintain the physical meaning of the original attributes in favor of explainability and interpretability. As stated beforehand, the numeric weights assigned to individual attributes which reflect attribute relevance allow to extended the explanation of recommendations by another dimension. In addition to information derived from the included neighborhood, the influence of the individual factors can be shown and, if necessary, even manually adjusted. As attribute weights are learned automatically from the data, no domain knowledge concerning the attributes is required.

5.3.2.2 Metric Learning (DR-LMNN)

Dropping irrelevant attributes and scaling the remaining dimensions in the attribute space according to the estimated relevance of an attribute is a straightforward approach and has shown to improve the performance of neighborhood-based classification algorithms in other applications [363]. Especially applying linear transformation to the data are widely and successfully used preprocessing strategies in the context of classification and data analysis. In comparison with attribute weighting, 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. Suchlike, in contrast to the proposed RBAs, correlations among attributes and the distribution of the data

Algorithm 1: CF-RBA

Input : D-dimensional patient data $\tilde{\mathbf{X}}^p_{train}$ of size N^p_{train}
M-dimensional outcome $\tilde{\mathbf{Y}}^p_{train}$ of size N^p_{train}
Number of iterations J
Number of nearest neighbors K_{RBA}
Initial attribute weights vector $\mathbf{w}^{init}$
Relevance threshold thr_w

Output : D-dimensional attribute weight vector $\mathbf{w}$

Initialize : Initial $\mathbf{w}$ with $\mathbf{w}^{init}$

for $j = 1...J$ **do**
  Randomly select target consultation $\mathbf{x}^j$;
  find K_{RBA} *nearest hits* $\tilde{\mathbf{X}}^{j,hits}_{train}$ and *nearest misses* $\tilde{\mathbf{X}}^{j,misses}_{train}$;
  for $d = 1...D$ **do**
    Compute average similarity of *nearest hits* $\bar{\rho}^{hits}_d$;
    Compute average similarity of *nearest misses* $\bar{\rho}^{misses}_d$;
    Update attribute weight vector $w_d = w_d + (\bar{\rho}^{hits}_d - \bar{\rho}^{misses}_d)/J$;
  end
  Normalize attribute weight vector $w_d = w_d/sum(\mathbf{w})$;
  Discard attribute dropping below relevance threshold $w_d < thr_w$;
end

in the attributes space are taken into account. This bears the potential to yield more meaningful neighborhoods, however, at the expense of explainability as the resulting latent features will be more difficult to interpret.

Also in the context of the proposed patient-data CF approach 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 of the patient-data CF, linear transformation of the data before computing similarity may be a beneficial preprocessing approach.

Mahalanobis distance [211, 254] incorporates linear transformation $\mathbf{x}' = \mathbf{L}\mathbf{x}$ of the present data $\mathbf{x}$ before computing *Euclidean distance* between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$ in the transformed attribute space according to

$$d_L(\mathbf{x}_i, \mathbf{x}_j) = ||\mathbf{L}(\mathbf{x}_i - \mathbf{x}_j)||_2^2 \tag{5.5}$$

Expressing the metric formulated above in terms of the square matrix $\mathbf{M} = \mathbf{L}^T\mathbf{L}$ yields the squared distance

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{M}(\mathbf{x}_i - \mathbf{x}_j) \tag{5.6}$$

$\mathbf{M}$ is allowed to denote any positive semidefinite matrix 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.

Additionally, in contrast to only rotating and scaling the data in order to adapt to its distribution assuming Gaussian distributions, generalized *Mahalanobis metrics* can exploit additional *a priori* information regarding similarity and dissimilarity between individual samples before computing distance (section 3.1.6). The objective of those 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 [377] learns such a generalized *Mahalanobis metric* as described in the following and is especially intended for neighborhood-based classification algorithms. Due to its intuitive approach and successful application in other domains, the LMNN algorithm is employed in this work.

In the following, the *Euclidean distance* is considered as baseline metric for computing similarity. As initially discussed in section 5.3.2, prerequisite to allow for applying *Minkowski metric* and also for applying linear transformation to the data are all attributes in the attribute space having equal quantitative data type and being normalization to the closed unit interval $[0, 1]$. Additionally, linear transformation cannot be applied to vectors with missing values. Consequently, the metric learning strategy can only be applied to the version of $\tilde{\mathbf{X}}^{p}_{train}$ in which all missing patient attributes were imputed as described in section 4.4.2 and missing therapy attributes were complemented with zeros.

Comparable to the proposed RBA algorithm, a transformation matrix $\mathbf{L}$ 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 $\mathbf{L}$ 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 objective function to be minimized, i.e. the loss function (equation 5.9), is composed of two competing objectives ϵ_{pull} and ϵ_{push}. Its relative impact is controlled using a meta parameter $\nu \in [0, 1]$ 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 consultations $\tilde{\mathbf{x}}^{k}$ labeled as similar are penalized. Here, the binary matrix η indicates whether $\tilde{\mathbf{x}}^{k}$ is a target neighbor of $\tilde{\mathbf{x}}^{j}$

$$\epsilon_{pull}(\mathbf{L}) = \sum_{j,k} \eta_{jk} ||\mathbf{L}(\tilde{\mathbf{x}}^{j} - \tilde{\mathbf{x}}^{k})||^2. \tag{5.7}$$

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 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. The binary matrix $\mathbf{y}^0$ indicates whether labels in $\tilde{\mathbf{y}}^{j}$ and $\tilde{\mathbf{y}}^{k}$ match.

$$\epsilon_{push}(\mathbf{L}) = \sum_{j,k,l} \eta_{jk}(1 - y^{0}_{jl})[1 + ||\mathbf{L}(\tilde{\mathbf{x}}^{j} - \tilde{\mathbf{x}}^{k})||^2 - ||\mathbf{L}(\tilde{\mathbf{x}}^{j} - \tilde{\mathbf{x}}^{l})||^2]_+ \tag{5.8}$$

By minimizing the combined loss function

$$\epsilon(\mathbf{L}) = (1-\nu)\epsilon_{pull}(\mathbf{L}) + \nu\epsilon_{push}(\mathbf{L}) \tag{5.9}$$

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

The required optimization of equation 5.9 can be formulated as a Semidefinite Program [377], i.e. a linear program with positive semidefinite matrix. As Semidefinite Programs are convex, the global minimum can be efficiently computed. The transformed *Euclidean distances* in equation 5.9 is replaced by equation 5.10 with the additional constraint of $\mathbf{M}$ to only have non-negative eigenvalues. Suchlike, a well-defined pseudometric can be learned. For each pair of differently labeled inputs the non-linear *hinge loss* is replaced by a slack variable ϵ_{jkl}. Finally, the resulting semidefinite program is defined by the conditions

$$\min_{\mathbf{M}} \sum_{j,k} \eta_{jk}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^k)^T\mathbf{M}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^k) + \nu \sum_{j,k} \eta_{jk}(1 - y^0_{jl})\epsilon_{jkl} \tag{5.10}$$

Subject to:

1. $(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^l)^T\mathbf{M}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^l) - (\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^k)^T\mathbf{M}(\tilde{\mathbf{x}}^j - \tilde{\mathbf{x}}^k) \geq 1 - \epsilon_{jkl}$
2. $\epsilon_{jkl} \geq 0$
3. $\mathbf{M} \succcurlyeq 0$

Comparable to the above described RBA, the LMNN algorithm proposed by [377] assumes a supervised classification problem where each sample is associated with one class label which corresponds to a distinct ground truth. However, as also described above, in the context of the therapy RS problem, each consultation is characterized by observed ground truth, which is quantified as numeric outcome indicators for applied therapies, and unobserved ground truth, which are all therapy options which have not been applied. Assumptions concerning similarity or dissimilarity between pairs of consultations can only be derived from samples with commonly applied therapies and with known outcomes, which yields the three groups of relationship described above: similar consultations (1), dissimilar consultations (2), and consultations for which no information on similarity is given (3).

As stated above, in the context of the RBA algorithm, *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 above, *target neighbors* are the K_{LMNN} closest consultations associated with equal therapy and similar 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 5.10 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 towards $\tilde{\mathbf{x}}^j$. 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.

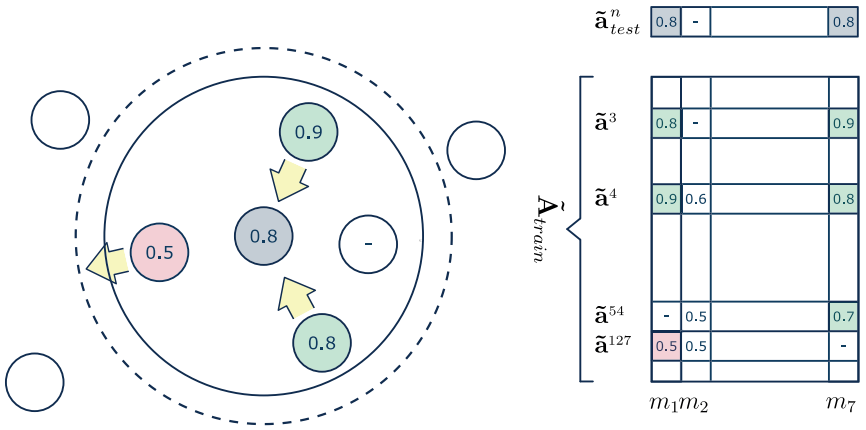


Figure 5.10: The LMNN algorithm assumes a supervised classification problem. Consultations are labeled with respect to a target consultation j and *a priori* similarity and dissimilarity assumptions as introduced in section 5.3.2.1. The LMNN algorithm intents to cause the target consultation $\tilde{\mathbf{x}}^j$ to be surrounded by samples of the same class while being separated from samples of different classes. [133]

5.4 Sparse Linear Model (SLIM)

In both, the conventional (section 5.3.1) and the patient-data CF approach (section 5.3.2), one target patient specific linear model is applied to predict outcome for all treatment options. Outcome observed in neighboring consultations are considered as independent variables and the similarities between consultations are model coefficients. Consultations are either represented by outcome observed for previous treatments only or by additional incorporation of patient data, respectively. However, as already discussed, these models are subject to some limitations. Firstly, the reliability of the model coefficients derived from heuristic similarity measures highly depend on the included attributes and the available data. Secondly, interdependencies among patients and therapies remain untapped and only sub-optimal solutions regarding the known outcome are applied. Finally, the computational efficiency during runtime deteriorates with an increasing number of consultations included in the database.

To overcome those limitations, the application of a linear regression model inspired by the SLIM [246] outlined in section 3.2.2.2 is modified for therapy recommendation. In contrast to

all previously studied methods which use patient-specific models, an item-based approach is implemented in the following. For the present therapy recommendation application this means that an individual model is learned for each treatment option.
In case of the proposed SLIM implementation, a therapy specific weight vector $\mathbf{w}_m$ is employed to model outcome y_m^n of treatment m and test consultation n as

$$\hat{y}_m^n = \tilde{\mathbf{x}}_{test}^{n\,T} \mathbf{w}_m \tag{5.11}$$

In accordance with section 3.2.2.2, for each treatment option an independent objective function can be formulated

$$L(\mathbf{w}_m) = \frac{1}{2}||\mathbf{y}_m - \tilde{\mathbf{X}}_{train}\mathbf{w}_m||_2^2 + \frac{\beta}{2}||\mathbf{w}_m||_2^2 + \lambda||\mathbf{w}_m||_1 \tag{5.12}$$

which is to be minimized. By utilizing all columns in $\tilde{\mathbf{X}}_{train}$ as independent variables, outcome information and additional side information are included into the linear model. However, in contrast to the methods proposed in [247], this side information is associated with users instead of items. Furthermore, compared to the original approach, each row of $\tilde{\mathbf{Y}}_{train}$ only contains observed outcome regarding one therapy option. Like this, dependencies concerning the temporal sequence of applied therapies are maintained in contrast to the SLIM algorithm, however, on the expense of only little training data.
The model hyperparameters, namely the L_1 and L_2 regularization weights β and λ, need to be optimized in the inner cross-validation loop. Furthermore, as it is the case in the previously described metric learning approach, the various scales of the patient describing attributes require normalization and the application of complete consultation representation matrices $\tilde{\mathbf{X}}_{train}$ and $\tilde{\mathbf{X}}_{test}$ for training and testing. Accordingly, all missing patient attributes were imputed as described in section 4.4.2, missing therapy attributes were complemented with zeros, and all attributes are min-max-normalized.

Finally, as for the neighborhood-based CF approaches, for each test consultation in $\tilde{\mathbf{X}}_{test}$ the accuracy of the predicted outcome is evaluated by calculating the RMSE between predicted and actually observed outcome in $\tilde{\mathbf{Y}}_{test}$. In the subsequent recommendation step, the *affinity* scores predicted by each of the regression models are ranked. The top-3 ranked entries are recommended and evaluated by computing the MAP at position 3.

Compared to the neighborhood-based approaches, the proposed algorithm is capable of adapting to the underlying data by finding an optimal solution to model the observed outcome. This approach has the potential to reveal patterns in the data and to achieve superior results. Moreover, the linear coefficients for each model can reveal insights into the importance of respective attributes. Especially, attribute selection, inherent to the *elastic net* regularization, and the additional requirement for solutions to satisfy $w_m \geq 0$ facilitates interpretability. As the range of all attributes are rescaled to the same range, the obtained coefficients can be directly associated with the average importance of the respective attributes. Moreover, by multiplying the weights with the actual attribute values, more detailed analyses concerning the individual attributes' contribution to the prediction can be provided. However, SLIM highly depends on sufficient

training data for each therapy option. When incorporating too few training consultations, the learned model is hardly capable of generalizing unseen data.

5.5 Gradient-boosted Regression Trees (GBM)

The proposed SLIM algorithm intends to find optimal coefficients to model the observed feedback on an item. Therefore, a linear relationship is assumed between the feedback observed on other items and additional side information included into the model. This assumption, however, is not always valid.
As introduced in section 3.3, there exist a variety of further supervised regression and classification algorithms which learn functions to map input vectors $\mathbf{x}$ to quantitative or qualitative target values such as the observed feedback. One of the most popular and successfully applied machine learning algorithms, especially for classification tasks, are DTs. However, DTs can be also capable of learning regression models as introduced in section 3.3. As they are non-parametric, no assumptions on the relation between dependent and independent variables are made and good classification and regression accuracies can be achieved with comparably little training data by revealing even non-linear relationships.

Beyond that, DT induction and classification algorithms such as C4.5 and Classification and Regression Tree (CART) are capable of handling quantitative and qualitative data types as they are present in the data at hand. Both algorithms can also cope with missing values to a certain extent. Additionally, embedded attribute selection makes them comparably robust to large attribute spaces containing irrelevant attributes. However, as pointed out in section 3.2.2.3, one main issue related to DTs is their lacking generalization capability. To overcome this problem and yield an universally powerful model, in this work a Gradient Boosting Machine (GBM) ensemble strategy, namely *Regularized Gradient Boosting*, is employed. This method is supposed to be superior to other ensemble approaches concerning accuracy and generalization capability especially at the presence of a limited data foundation. A popular and very efficient GBM implementation providing state-of-the-art results is the *XGBoost* implementation[1] which is utilized in the following.

As it is the case with the aforementioned SLIM algorithm, an item-based approach is implemented. For each treatment option a *boosted* regression tree ensemble is provided as outcome prediction model. As objective function to be minimized, the squared error (section F.1.6), is applied. Comparable to the above linear model approach, the subsets of training consultation representations in $\tilde{\mathbf{X}}_{train}$, which are associated with the individual therapy options, are used as model input and the respective treatment response in $\tilde{\mathbf{Y}}_{train}$ as ground truth label. Even thought DTs and the applied GBM are capable of coping with missing values, the complete and min-max-normalized version of $\tilde{\mathbf{X}}$ is utilized for the sake of comparability.
In order to prevent overfitting in spite of the small data foundation, the base learner are trained on a subsample of the training data only in each iteration. Here, subsampling 80 % of all training data has shown best results in preliminary experiments. Also learning rate $\mu = 0.001$ and the

[1] https://xgboost.readthedocs.io

L_1 and L_2 regularization weights $\beta = 0.01$ and $\lambda = 0.01$ are defined based on own preliminary studies. Number of trees to fit n_{trees}, maximum base learner tree depth d_{max}, and the minimum sum of instance weights w_{child}, which is needed in a node before stopping further partitioning, are determined by a grid search in the inner cross-validation loop. Especially d_{max} and w_{child} are essential to prevent overfitting problems. The remaining parameters are left on default values.

Also in case of the GBM approach, prediction accuracy (RMSE) is evaluated for each of the models individually. A recommendation list, which is evaluated in terms of MAP@3, is generated by ranking the individual model outputs according to their outcome predictions. Comparable to RBA and SLIM, also the regression tree ensemble model is capable of providing insights into attribute importance. Beyond that, as introduced in section F.1.7, local or global surrogate models can supplement interpretability. Suitable surrogate models are linear models, comparable to SLIM, or DTs which can be easily translated into interpretable rule sets.

5.6 Evidence-based and Expert-based Exclusion Rules

As already pointed out in chapter 1 and section 5.1, reducing the potential of generating inappropriate or even harmful recommendations plays a crucial role regarding trust and acceptance of CDSSs. In contrast to other domains of RSs, e.g. e-commerce applications, in the area of health and medicine, failures in recommendations accompany high risks. Particularly at the presence of small training databases it can be assumed that patterns, which would exclude inappropriate treatments from the top-3 recommendation list, can remain unrevealed. Therefore, to increase confidence and minimize risk that emanates from automatically generated data-based therapy recommendations, a rule-based post-filtering layer is implemented. Post-filtering is one possible strategy to incorporate context, such as time or place, a user interacts with a recommender system, into the recommendation process [2]. In contrast to *pre-filtering*, which only considers users or item according to the current context for the recommendation generation, *post-filtering* modifies the resulting recommendation list based on contextual information. Here, treatment options, which are recommended by the RS algorithm though are ruled out due to an exclusion criterion, are simply removed from the recommendation list. This results in other therapies moving into the top-3 list.

Basically, three groups of rules can be distinguished which are derived from the current S3-Guidelines on the treatment of *Psoriasis vulgaris* [240] and additional specifications provided by advising clinicians from the *Clinic and Polyclinic for Dermatology, University Hospital Dresden*. (1) Absolute contraindications due to comorbidities or the current life situation as described in the S3-Guidelines [240] and listed in section 4.3, (2) Psoriasis type specific exclusion criteria, and (3) recommendations regarding the sequence of applied therapies derived from the S3-Guidelines [240] or specified by the advising clinicians. The grouped rules are summarized in table 5.1. Additionally to the stated rules, (4) the requirement not to recommend treatments which have been applied and discontinued before was requested by the clinical experts. The integration of rule sets derived from the S3-Guidelines can be regarded as the incorporation of external evidence within the framework of EbM.

Table 5.1: Exclusion rules based on external evidence and specification by clinical experts.

Condition	**Exclusion**
(1) Absolute contraindications	
Arterial hypertension	Cyclosporine
Renal dysfunction	Acitretin, Cyclosporine, Fumaric acid esters, Methotrexate
Hepatic dysfunction	Fumaric acid esters, Methotrexate
Gastrointestinal disease	Fumaric acid esters
Malignancies	Cyclosporine, Biopharmaceuticals
Tuberculosis or other severe infections	Cyclosporine, Methotrexate, Biopharmaceuticals,
Planned child	Actitretin, Methotrexate
Pregnancy, breastfeeding	Actitretin, Apremilast, Methotrexate, Sekukinumab
(2) Psoriasis types specific	
Psoriasis arthritis only	Acitretin, Cyclosporine, Fumaric acid esters, UV therapy
No *Psoriasis arthritis*	Golimumab
(3) Sequence of therapies	
Not any first-line conventional pharmaceuticals applied	Second-line conventional pharmaceuticals, First-line biopharmaceuticals, Second-line biopharmaceuticals
Not all first-line conventional pharmaceuticals applied	Second-line biopharmaceuticals

All rule sets are implemented suchlike that they can be applied individually. Three strategies are studied and compared within this work:

a Exclusion of therapies which are contraindicated and due to Psoriasis type, i.e. rule sets (1) and (2).

b Exclusion of therapies as in (a) and due to sequence of therapies, i.e. rule set (3).

c Exclusion of therapies as in (b) and exclusion of therapies already applied in a patients therapy history (4).

6 Results

In the following chapter, the performance of the various therapy recommendation algorithms and variations introduced in chapter 5 are compared. In section 6.1, results from model selection, i.e. the inner cross-validation loop, are contrasted and discussed for each of the proposed methods and the best model for each approach is selected. In section 6.2, generalization performance estimates, yielded in the outer cross-validation loop, are compared and discussed. Finally, in section 6.3, the recommendation performance is compared with human experts.

6.1 Model Selection

6.1.1 Collaborative Filtering

6.1.1.1 Conventional Collaborative Recommender (CF)

In the following, for all proposed approaches, mean values and standard deviations of the inner cross validation results (i.e. average over all 5 folds) for each of the discussed scores are shown. Figure 6.1 (a) and (b) demonstrate outcome prediction accuracy (RMSE) and the agreement between the top-3 ranked recommendations and the attending physician's successful choice (MAP@3) of the conventional CF approach (CF) as described in section 5.3.1. Additionally, figure 6.2 (a) shows the ratio of test consultations for which RMSE could be computed, i.e. for which predictions *overlap* with the actually applied treatment, and figure 6.2 (b) shows the ratio of treatment options appearing in the recommendation list, i.e. *coverage*.

Two baseline approaches are discussed and compared with the algorithms' results. Firstly, the average *affinity* scores for each treatment are computed as outcome prediction baseline (*average efficiency*). Ranking those predictions according to outcome provides one recommendation baseline. Secondly, the individual therapies' frequency of application in the training partitions (*overall popularity*) are employed as second recommendation baseline. As can be seen in figure 6.1 (a), exploiting a selected neighborhood to derive outcome predictions from is clearly superior to utilizing only the average *affinity* scores. Also, the ranked list of recommendations derived from the conventional CF predictions achieves significantly higher MAP@3 scores than those based on averages only. Nevertheless, as demonstrated in figure 6.1 (b), ranking based on the treatment options' average outcome is still superior to only recommending treatments based on the application frequency.

When comparing the group of correlation-based similarity measures *Cosine similarity* and *Pearson correlation coefficient* with the *Minkowski metrics*, clearly distinctive results are yielded. Outcome prediction of the conventional CF using *Cosine similarity* or *Pearson correlation*

coefficient as similarity measure is overall inferior to algorithms using *Manhatten* or *Euclidean distance*. Regarding the ability to rank the actually applied and successful therapy among the top-3 options, however, *Cosine similarity* and *Pearson correlation* outperform the *Minkowski metrics* clearly for a wide range of K. On the one hand, as can be seen in figure 6.2 (b), *Cosine similarity* and *Pearson correlation* are capable of retrieving already at very small neighborhood sizes a large ratio of consultations with high degree of overlap with the actually applied treatment. Simultaneously, coverage (figure 6.2 (a)) 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 already for small neighborhoods which only slowly deteriorates with rising K. Yet, the neighboring consultations only allow for comparably bad outcome prediction. The *Minkowski metrics*, on the other hand, facilitate much better RMSE values which, however, is based on only few overlapping therapies between prediction and ground truth, at least for small K. Also *coverage* is already for small K rather large which overall yields inaccurate recommendations and low quality therapy ranking.
One major difference between the two groups of similarity measures is how missing entries, i.e. treatments which have not been applied in both consultation representations, are treated. *Cosine similarity* and *Pearson correlation* coefficient assesses the orientation of two vector representations without taking the magnitude of the vectors into account. To do so, both similarity measures scale the consultation vectors to unit length before calculating the dot product. Hence, similarity is increased for co-occurring treatments but is penalized by each applied treatment in either of the two consulting vectors which do not overlap. In contrast, the *Minkowski metrics* consider magnitude when computing the distance between two vectors and no normalization is applied. In the present application, only co-occurring treatment applications are included into the distance computation. Thus, all dimensions with missing entries in either of the vectors are simply ignored. Hence, *Cosine similarity* and *Pearson correlation* coefficient incorporate more information when comparing consultation representations which explains the more selective recommendation list. Nevertheless, to derive accurate outcome predictions it is obviously more important to observe very similar outcome on co-occurring treatments than to have overall similar vectors regarding the number of applied therapies.

Within the two groups of similarity measures there are only minor differences regarding prediction and recommendation performance. *Pearson correlation*, which is simply the *Cosine similarity* after deducting the mean of *affinity* scores of a consultation representation vector, performs slightly better for rising K concerning both, outcome prediction and therapy recommendation. The same is true for *Euclidean distance* compared with the *Manhatten distance*. This observation can be explained with the fact that with rising order p of the *Minkowski metric* the impact of larger differences between dimensions on the overall distance rises. Consequently, consultation representations containing larger *affinity* differences for treatments become more distant for *Euclidean distance* ($p = 2$) compared with *Manhatten distance* ($p = 1$).

As specified in 1, the primary evaluation criterion of this work is the accuracy of outcome predictions. However, as additional criterion the ratio of neighbors overlapping the actually

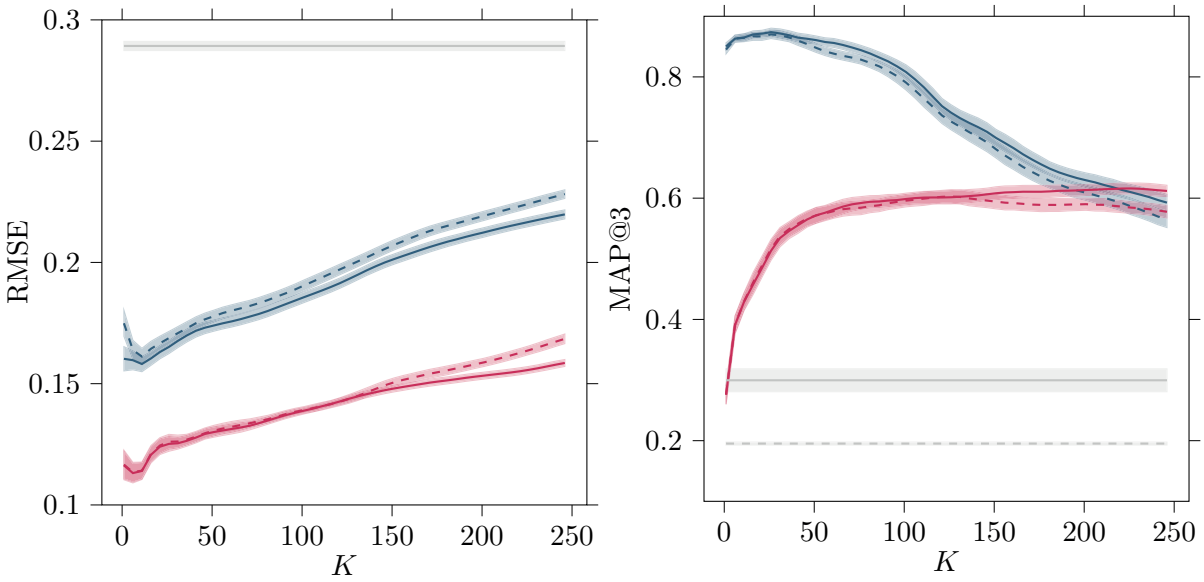


Figure 6.1: CF: (a) RMSE between estimated and observed outcome and (b) MAP@3 evaluating the ranked list of recommended therapies. The similarity and distance measures *Cosine similarity* (- - -), *Pearson correlation* (——), *Manhatten distance* (- - -) and *Euclidean distance* (——) are compared. Additionally, outcome prediction and treatment recommendation based on *average efficiency*, i.e. *affinity* score of each treatment option averaged over all training consultations (——), and *overall popularity*, i.e. frequency of application in the training consultations (- - -), are shown. RMSE and MAP@3 are computed for a neighborhood size range $K \in [1, 250]$.

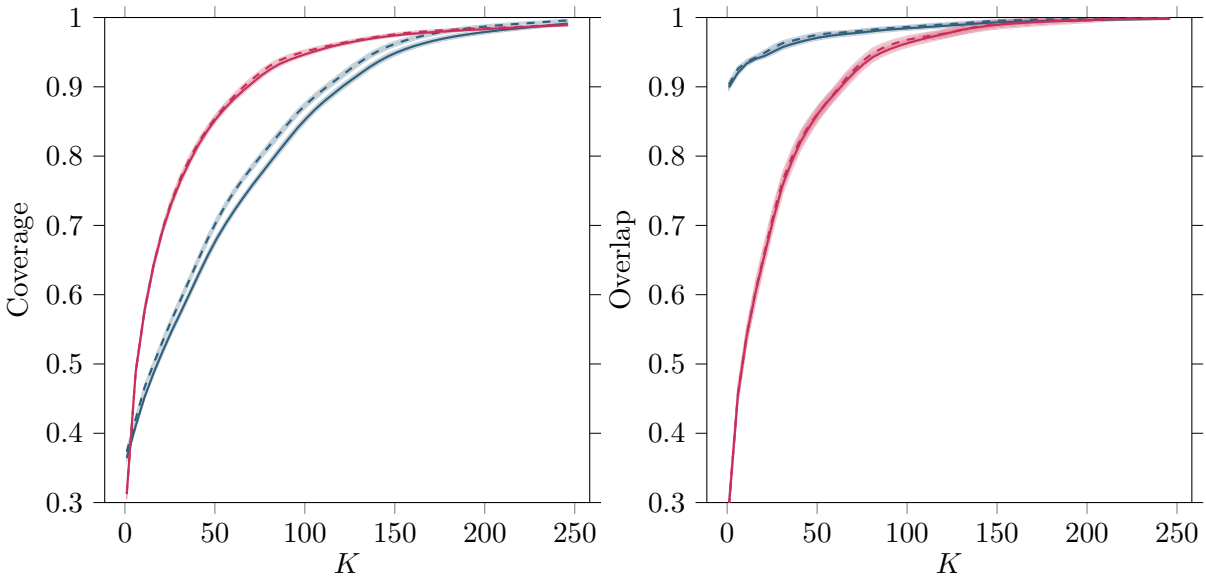


Figure 6.2: CF: (a) Coverage of available treatment options and (b) ratio of neighbors overlapping the actually applied therapy. The similarity and distance measures *Cosine similarity* (- - -), *Pearson correlation* (——), *Manhatten distance* (- - -) and *Euclidean distance* (——) are compared. *Coverage* and *overlap* are computed for a neighborhood size range $K \in [1, 250]$.

applied therapy is defined to exceed $overlap \geq 0.95$ in order to base the selection on reliable values. Table 6.1 summarizes the selected K for all proposed conventional CF approaches together with the corresponding evaluation metrics. Additionally, the baseline results are demonstrated. For the correlation-based similarity measures, the selected neighborhood size K is considerably smaller than those of the *Minkowski metrics* due to a much larger ratio of neighbors overlapping the actually applied therapy (*overlap*) already for small K.

Table 6.1: CF: Inner cross-validation results (5-fold cross-validation). Best K for which RMSE is minimal and the $overlap \geq 0.95$ criterion is met. Additionally, average and standard deviation of the evaluation metrics RMSE, MAP@3, *coverage* and *overlap* are shown.

Metric	K	RMSE	MAP@3	Coverage	Overlap
CF (Cosine)	23.98 (3.20)	0.17 (0.00)	0.87 (0.01)	0.55 (0.02)	0.95 (0.00)
CF (Pearson)	27.91 (3.00)	0.17 (0.00)	0.87 (0.01)	0.56 (0.02)	0.95 (0.00)
CF (Manhatten)	85.50 (5.45)	0.14 (0.00)	0.59 (0.01)	0.94 (0.01)	0.96 (0.00)
CF (Euclidean)	89.23 (6.14)	0.14 (0.00)	0.60 (0.01)	0.94 (0.01)	0.95 (0.00)
Average efficiency	- (-)	0.29 (0.00)	0.30 (0.02)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	0.46 (0.00)	0.20 (0.00)	1.00 (0.00)	1.00 (0.00)

In a preliminary experiment, also the performance of an *implicit* version of the conventional CF algorithm was studied. Instead of utilizing *affinity* scores to represent consultations, the binary information whether treatments have been previously applied to a patient are utilized to compare consultations. Independent of the applied similarity or distance measure, this approach has shown to be inferior to the *explicit* CF. The straightforward interpretation is that the information stored in the *affinity* scores is of essential value for consultation comparison. This observation substantiates the hypothesis that there exist patterns in treatment outcome which can serve as information source for personalized treatment recommendations.

6.1.1.2 Patient-data Collaborative Recommender (DR)

As introduced in section 5.3.2, the patient-data CF approach (DR) compares consultations represented by both, treatment history and available patient data. Hence, a much wider range of attributes with distinct properties are incorporated into the similarity computation. Two measures for computing similarity between consultation representations, *Euclidean distance* and *Gower similarity*, are compared.

As can be seen in table 6.2 and figure 6.3 (a), also outcome prediction accuracy of both patient-data CF versions clearly outperform the RMSE baseline, i.e. the *average efficiency*. The RMSEs of both approaches are small for small K and becomes larger with rising neighborhood size. The quality of the top-3 list of recommendations, shown in figure 6.3 (b), is also capable of achieving better results than by only recommending the overall most popular or most efficient treatments. However, MAP@3 shows a clear maximum for a rather small neighborhood size which quickly deteriorates with rising K and asymptotically approaches the baselines. The characteristic of both evaluation measures can be interpreted suchlike that there obviously is

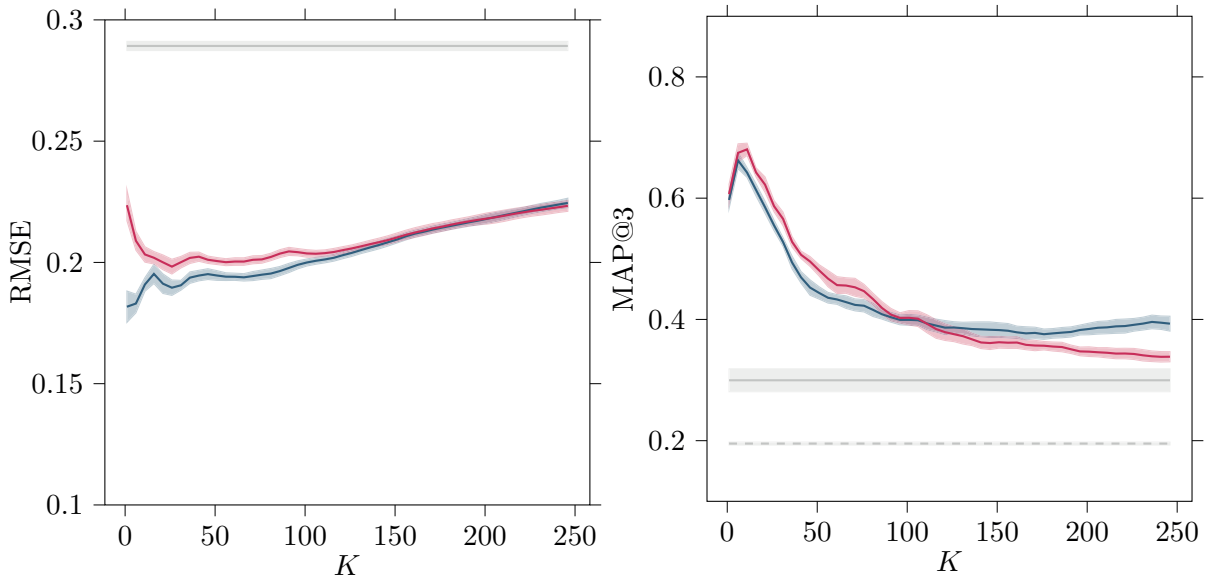


Figure 6.3: DR: (a) RMSE between estimated and observed outcome and (b) MAP@3 evaluating the ranked list of recommended therapies. The similarity and distance measures *Euclidean distance* (—) and *Gower similarity* (—) are compared. Additionally, outcome prediction and treatment recommendation based on *average efficiency*, i.e. *affinity* score of each treatment option averaged over all training consultations (—), and *overall popularity*, i.e. frequency of application in the training consultations (- - -), are shown. RMSE and MAP@3 are computed for a neighborhood size range $K \in [1, 250]$.

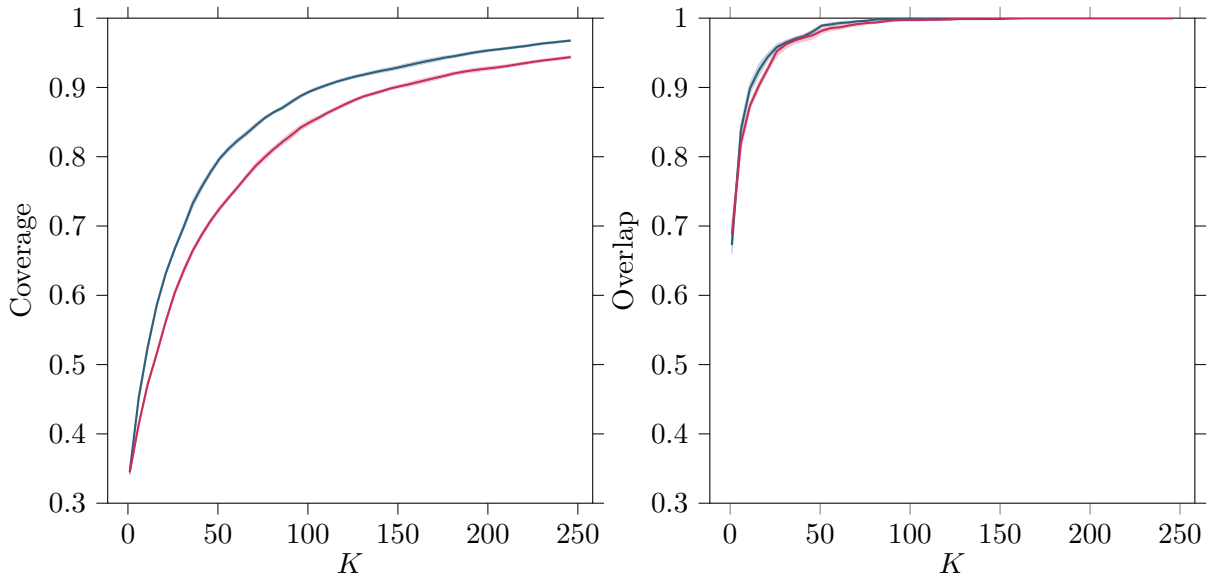


Figure 6.4: DR: (a) Coverage of available treatment options and (b) ratio of neighbors overlapping the actually applied therapy. The similarity and distance measures *Euclidean distance* (—) and *Gower similarity* (—) are compared. *Coverage* and *overlap* are computed for a neighborhood size range $K \in [1, 250]$.

an optimal neighborhood size. For too small K, the included data is insufficient. For rising neighborhood sizes, the range of treatment options rises and hence noise and inappropriate information is increasingly included.

When comparing *Euclidean distance* with *Gower similarity* to compute similarity between consultations, only outcome prediction performance (RMSE) shows noteworthy differences. *Gower similarity* outperforms *Euclidean distance* which indicates that considering scale of measurement, i.e. data type, is obviously beneficial when comparing attributes. As shown in figure 6.4 (a) and (b), the behavior of *coverage* and *overlap* of both patient-data CF approaches is similar to the correlation-based conventional CF. Both patient-data CF approaches are capable of retrieving already at very small neighborhood sizes a large ratio of consultations with high degree of overlap with the actually applied treatment to compute RMSE from. Yet, the neighboring consultations only allow for comparably bad outcome predictions. *Coverage*, on the other hand, is comparably low and only slowly rises with increasing K. Hence, the retrieved neighboring consultations are very selective regarding treatment options and introduce only little noise into the recommendation which results in high MAP@3 values already for small K. However, as stated in 1, primary evaluation criterion is high prediction accuracy as foundation for recommending the potentially best treatment option which is not identical to high agreement with the physician's recommendation.

The identified best K listed in table 6.2 minimizes RMSE with respect to the $overlap \geq 0.95$ criterion. However, as can be seen in figure 6.3 (b), the ratio of the top-3 recommendations which agree with the attending physician's successful recommendations is larger for smaller K than the point where the *overlap* criterion is met. Consequently, mean MAP@3 of the selected K does not exceed 0.54, respectively. Even though the difference is rather small, *Gower similarity* performs slightly better than *Euclidean distance* in terms of outcome prediction and is preferred in the following.

Table 6.2: DR: Inner cross-validation results (5-fold cross-validation). Best K for which RMSE is minimal and the $overlap \geq 0.95$ criterion is met. Additionally, average and standard deviation of the evaluation metrics RMSE, MAP@3, *coverage* and *overlap* are shown.

Metric	**K**	**RMSE**	**MAP@3**	**Coverage**	**Overlap**
DR (Gower)	29.98 (9.51)	0.19 (0.00)	0.54 (0.03)	0.69 (0.04)	0.96 (0.01)
DR (Euclidean)	38.54 (15.63)	0.20 (0.00)	0.54 (0.05)	0.66 (0.07)	0.97 (0.02)
Average efficiency	- (-)	0.29 (0.00)	0.30 (0.02)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	- (-)	0.20 (0.00)	1.00 (0.00)	1.00 (0.00)

6.1.1.3 Missing Value Imputation (DR-Impute)

As stated in chapter 4, the available data is characterized by numerous missing values. The above described patient-data CF utilizes completely imputed versions of treatment history and patient data (*impute 2*). As imputing missing values underlies the risk of introducing noise and corrupted data, the impact of sparsity on the outcome prediction and therapy recommendation

is studied. Due to the inherent capability of handling missing values and the slightly superior performance compared with the *Euclidean* approach, the *Gower similarity* approach, described in section 6.1.1.2, is compared for all three imputation stages summarized in table 4.4.

Figure 6.5 (a) demonstrates the *affinity* prediction superiority of less incomplete consultation representations over sparse representations which becomes particularly apparent with rising neighborhood sizes. The dataset with reduced number of missing values (*impute 1*) in turn shows similar results as the complete consultation representations (*impute 2*). At the selected K, shown in table 6.3, however, no RMSE differences are evident between the three data sets.
Also concerning the capability to rank the successful physician's choice among the top-3 recommendations, utilizing the consultation representations with reduced number of missing values (*impute 1*) shows almost equal results as the complete version (*impute 2*). Interestingly, the ranked therapy recommendation list shows largest overlap with the physician's choice for the entire studied interval of K when using consultation representations without any data imputation (*impute 0*). This is also true for the selected neighborhood size K shown in table 6.3.
As it is the case for MAP@3, also *coverage* and *overlap* do not differ greatly from the above described patient-data CF when comparing the use of the data versions *impute 1* and *impute 2*. Again, only the raw consultation representations (*impute 0*) yields deviating results. *Coverage*, shown in figure 6.6 (b), is overall slightly larger when using the raw data and hence more treatment options are included into the recommendation list. Also *overlap* is already larger for smaller K with raw consultation representations with missing values, which reflects the better MAP@3 results.

Table 6.3: DR-Impute: Inner cross-validation results (5-fold cross-validation). Best K for which RMSE is minimal and the *overlap* ≥ 0.95 criterion is met. Additionally, average and standard deviation of the evaluation metrics RMSE, MAP@3, *coverage* and *overlap* are shown.

Metric	**K**	**RMSE**	**MAP@3**	**Coverage**	**Overlap**
DR-Impute 0 (Gower)	29.43 (7.97)	0.19 (0.00)	0.61 (0.05)	0.70 (0.06)	0.98 (0.01)
DR-Impute 1 (Gower)	33.18 (14.95)	0.19 (0.00)	0.55 (0.05)	0.70 (0.06)	0.97 (0.01)
DR-Impute 2 (Gower)	29.98 (9.51)	0.19 (0.00)	0.54 (0.03)	0.69 (0.04)	0.96 (0.01)
Average efficiency	- (-)	0.29 (0.00)	0.30 (0.02)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	- (-)	0.20 (0.00)	1.00 (0.00)	1.00 (0.00)

Besides the *curse of dimensionality* and the problems associated with attribute relevancy and redundancy referred to above, missing values arise additional issues concerning the similarity measure. At the presence of missing values, vectors with differing dimensions are compared. In case of the conventional CF described in section 6.1.1.1, missing treatments in the consultation representation vectors can be assumed to be NMAR and hence carry information which is reflected in the computed similarities. In contrast, most missing values in the patient data are considered to be MCAR or MAR. This requires partly uncertain assumptions regarding the imputation strategy and renders comparability of the resulting coefficients difficult and unreliable.

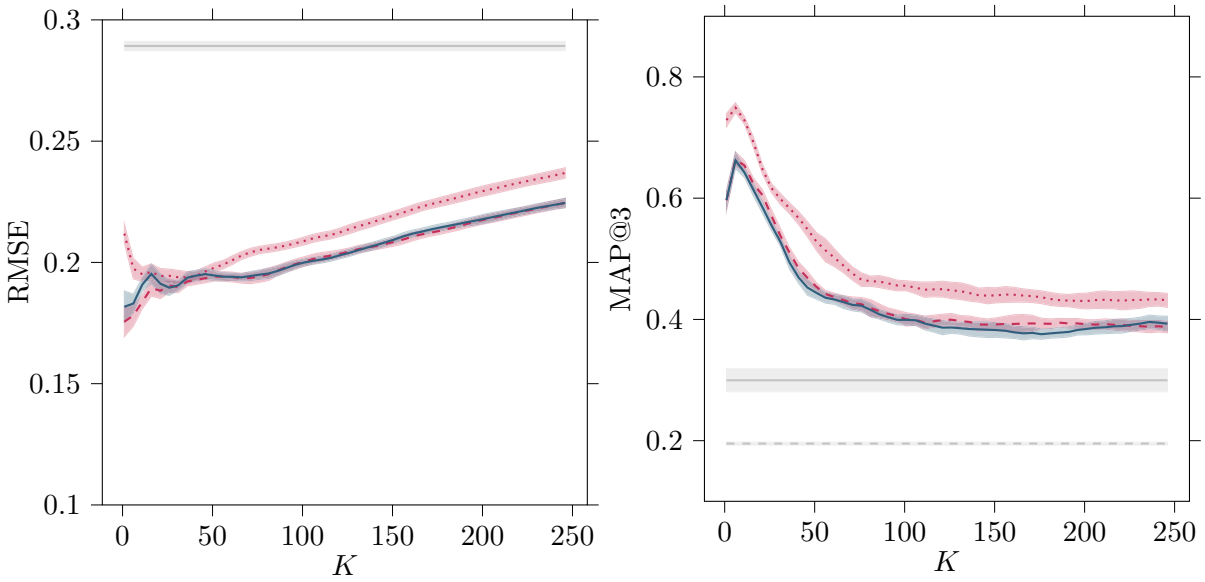


Figure 6.5: DR-Impute: (a) RMSE between estimated and observed outcome and (b) MAP@3 evaluating the ranked list of recommended therapies. *Gower similarity* is compared for the three imputation stages *impute 0* (·····), *impute 1* (- - -), and *impute 2* (——). Additionally, outcome prediction and treatment recommendation based on *average efficiency*, i.e. *affinity* score of each treatment option averaged over all training consultations (——), and *overall popularity*, i.e. frequency of application in the training consultations (- - -), are shown. RMSE and MAP@3 are computed for a neighborhood size range $K \in [1, 250]$.

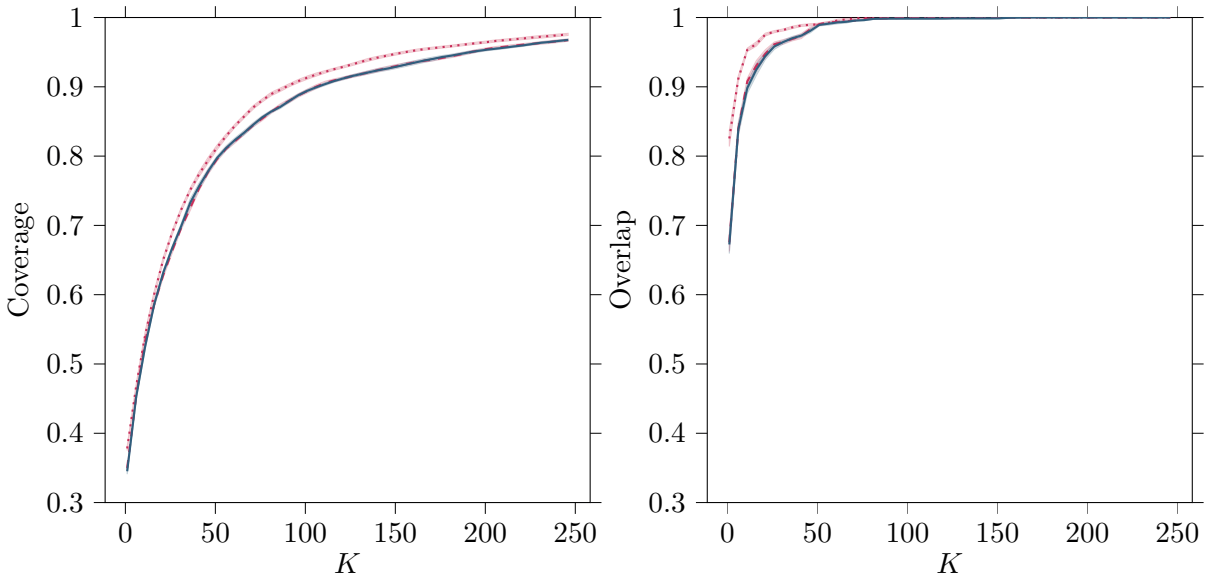


Figure 6.6: DR-Impute: (a) Coverage of available treatment options and (b) ratio of neighbors overlapping the actually applied therapy. *Gower similarity* is compared for the three imputation states *impute 0* (·····), *impute 1* (- - -), and *impute 2* (——). *Coverage* and *overlap* are computed for a neighborhood size range $K \in [1, 250]$.

To conclude, no general superiority of data imputation over dealing with incomplete vectors could be shown. According to the RMSE curve, the imputation strategies are particularly valid with regard to the task of predicting outcomes.

6.1.1.4 Exclusion Rules (DR-Rules)

As introduced in section 5.6, rule sets are implemented in order to integrate literature-based evidence (S3-Guideline [240]) and further local expert recommendations concerning the treatment of *Psoriasis vulgaris*. A post-filtering layer truncates the ranked lists of treatment options according to those rules by excluding presumably inappropriate treatments. In the following, the impact of exclusion rules on the patient-data CF recommendation list is studied. Three combinations of rule sets are compared which differ in the extend to which treatment options are excluded as specified in section 5.6.

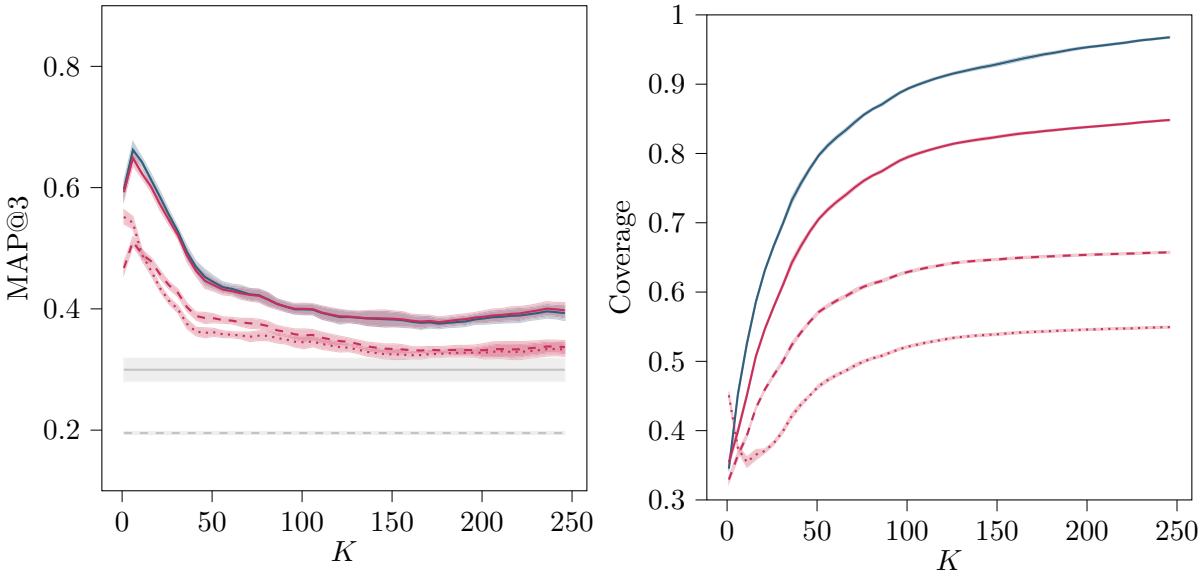


Figure 6.7: DR-Rules: (a) MAP@3 evaluating the ranked list of recommended therapies and (b) *coverage* of available treatment options. Application of no exclusion rules (——) is compared with exclusion rule sets *rules a* (——), *rules b* (- - -), and *rules c* (·····) as described in section 5.6. MAP@3 and *coverage* is computed for a neighborhood size range $K \in [1, 250]$.

As can be seen in figure 6.7 (a) and table 6.4, only excluding treatments which are contraindicated and due to the diagnosed Psoriasis type (*rules a*) has only minor impact on the algorithm's ability to rank the actually applied therapy among the top-3 options. *Coverage*, however, is clearly decreased as shown in figure 6.7 (b). This observation indicates that the patient-data CF algorithm is apparently capable of including those aspects when ranking treatments and the underlying data represents these rules.
Applying rule set *rules b*, which additionally excludes therapies due to the sequence of applied therapies, distinctly impacts the MAP@3 value negatively. Also *coverage* for the selected neighborhood size K is considerably reduced by additional 11% compared with *rules a*. This ob-

servation can be explained with a ground truth not being compliant with the demanded sequence of treatment application. The underlying data foundation either is incomplete or erroneous, or the attending physician was himself not complying with the required treatment sequence.
Additionally excluding treatment options already applied before but aborted, i.e. applying *rules c*, only deteriorates the overall recommendation quality marginally even though *coverage* is further reduced. This small additional MAP@3 deterioration can also be explained with a rule that is not always represented in the ground truth. Nevertheless, as it is already the case for *rules a*, the substantial decrease in *coverage* compared to small MAP@3 deterioration can be linked to exclusion of treatment options which are only rarely recommended by the attending physician. Hence, it can be assumed that the underlying data, which represents the attending physician's recommendations, adhere to those rules.

To summarize, the results show that exclusion of therapies which are contraindicated and due

Table 6.4: DR-Rules: Inner cross-validation results (5-fold cross-validation). Best K for which RMSE is minimal and the *overlap* ≥ 0.95 criterion is met. Additionally, average and standard deviation of the evaluation metrics RMSE, MAP@3, *coverage* and *overlap* are shown.

Metric	**K**	**RMSE**	**MAP@3**	**Coverage**	**Overlap**
No exclusion rules					
DR (Gower)	29.98 (9.51)	0.19 (0.00)	0.54 (0.03)	0.69 (0.04)	0.96 (0.01)
With exclusion rules					
DR-Rules a (Gower)	29.98 (9.51)	0.19 (0.00)	0.53 (0.03)	0.60 (0.04)	0.96 (0.01)
DR-Rules b (Gower)	29.98 (9.51)	0.19 (0.00)	0.43 (0.02)	0.49 (0.03)	0.96 (0.01)
DR-Rules c (Gower)	29.98 (9.51)	0.19 (0.00)	0.41 (0.01)	0.39 (0.02)	0.96 (0.01)
Average efficiency	- (-)	0.29 (0.00)	0.30 (0.02)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	- (-)	0.20 (0.00)	1.00 (0.00)	1.00 (0.00)

to Psoriasis type (*rules a*), but also exclusion of therapies already applied in a patients therapy history (*rules c*) obviously comply with the underlying data. Exclusion of treatment options which do not follow the sequence of therapies described in the S3-Guidelines and extended by the advising clinicians, however, is not in accordance with the ground truth represented by the attending physician's choice.

6.1.1.5 Attribute Selection and Weighting (DR-RBA)

The proposed RBA approach, described in section 5.3.2.1, assigns weights w_d to the individual attributes d, which is equivalent to scaling attributes according to their individual importance. In order to reduce the input space but also to address the above described problems with irrelevant attributes distorting the similarity computation, only important attributes are to be used for consultation comparison. For this purpose, only those attributes assigned with positive weights are taken into account. Finally, *Gower similarity* is computed from the consultations with weighted attributes. 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. The parameters are determined suchlike that RMSE is optimized respecting the *overlap* ≥ 0.95 criterion. The initial attribute weight vector $\mathbf{w}^{init}$ and the weight threshold thr_w for attribute selection are set to 0, resulting in important attributes to have positive weights and negative attributes to be neglected. Concerning K_{RBA}, the best RMSE could be constantly found for $K_{RBA} = 15$.

In figure 6.8 (a) and (b), the previously described *Gower similarity* patient-data CF is compared with the weighted version. Scaling attributes is obviously very beneficial regarding outcome prediction accuracy (RMSE) and quality of the list of top-3 ranked recommendations (MAP@3). Even though especially in the surroundings of K_{RBA} a clear minimum is evident, the prediction error is reduced in total and also MAP@3 is overall increased. This overall improvement is also reflected in the finally selected model, summarized in table 6.5. The DR-RBA approach outperforms the unweighted version for an even smaller neighborhood. Nevertheless, whereas *overlap* remains essentially unchanged, *coverage* is generally larger and the recommender hence tends to be less selective. For the chosen neighborhood size K, however, average *coverage* is identical with 69 %.

Table 6.5: DR-RBA: Inner cross-validation results (5-fold cross-validation). Best K for which RMSE is minimal and the *overlap* ≥ 0.95 criterion is met. Additionally, average and standard deviation of the evaluation metrics RMSE, MAP@3, *coverage* and *overlap* are shown.

Metric	**K**	**RMSE**	**MAP@3**	**Coverage**	**Overlap**
DR (Gower)	29.98 (9.51)	0.19 (0.00)	0.54 (0.03)	0.69 (0.04)	0.96 (0.01)
DR-RBA (Gower)	22.91 (5.28)	0.15 (0.00)	0.64 (0.03)	0.69 (0.04)	0.95 (0.00)
Average efficiency	- (-)	0.29 (0.00)	0.30 (0.02)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	- (-)	0.20 (0.00)	1.00 (0.00)	1.00 (0.00)

Firstly, the results clearly indicate that weighting and selecting attributes according to their assumed importance can improve the prediction accuracy of a neighborhood-based CF approach. The similarity measures underlying the algorithm become more meaningful concerning the given task. Secondly, the results imply that the applied supervised method to learn attribute weights is an appropriate choice and the similarity constraints prove to be justifiable. Nevertheless, it must be kept in mind that redundancies between attributes are not addressed by this method. The RBA does not detect correlations and learns similar weights for dependent attributes.

As described above, for each outer cross-validation iteration, i.e. test patient p, an individual attribute weight vector $\mathbf{w}$ is learned using all other patients' consultations as training data. Considering all 181 iterations, on average the weights of 42.56 (26.77 %) attributes drop below the determined threshold $thr_w = 0$ and are neglected. Consequently, the attribute space is on average reduced to 116.44 attributes. Figures 6.11, 6.12, and 6.13 present the attribute weights. It is noticeable that the variance of the number of dropped values but also the weight values themselves are comparably small for the individual attributes. This observation can be interpreted as a meaningful and reliable selection process.
When comparing patient data and previous treatment attributes, overall, the latter gain more

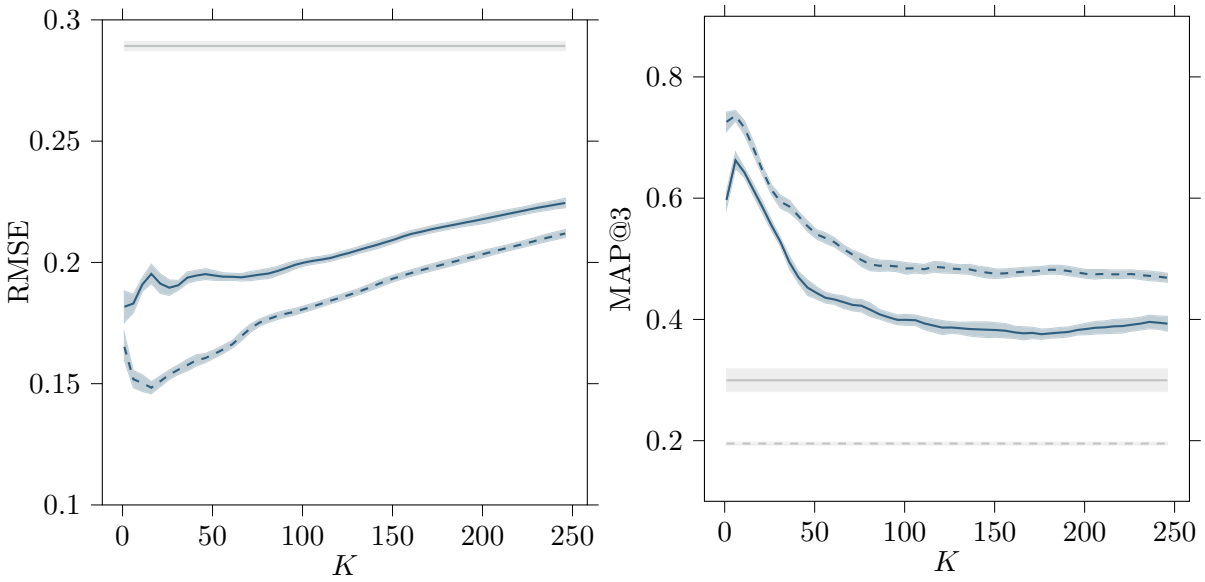


Figure 6.8: DR-RBA: (a) RMSE between estimated and observed outcome and (b) MAP@3 evaluating the ranked list of recommended therapies. *Gower similarity* without (——) and with (- - -) applying attribute selection and weighting are compared. Additionally, outcome prediction and treatment recommendation based on *average efficiency*, i.e. *affinity* score of each treatment option averaged over all training consultations (——), and *overall popularity*, i.e. frequency of application in the training consultations (- - -), are shown. RMSE and MAP@3 are computed for a neighborhood size range $K \in [1, 250]$.

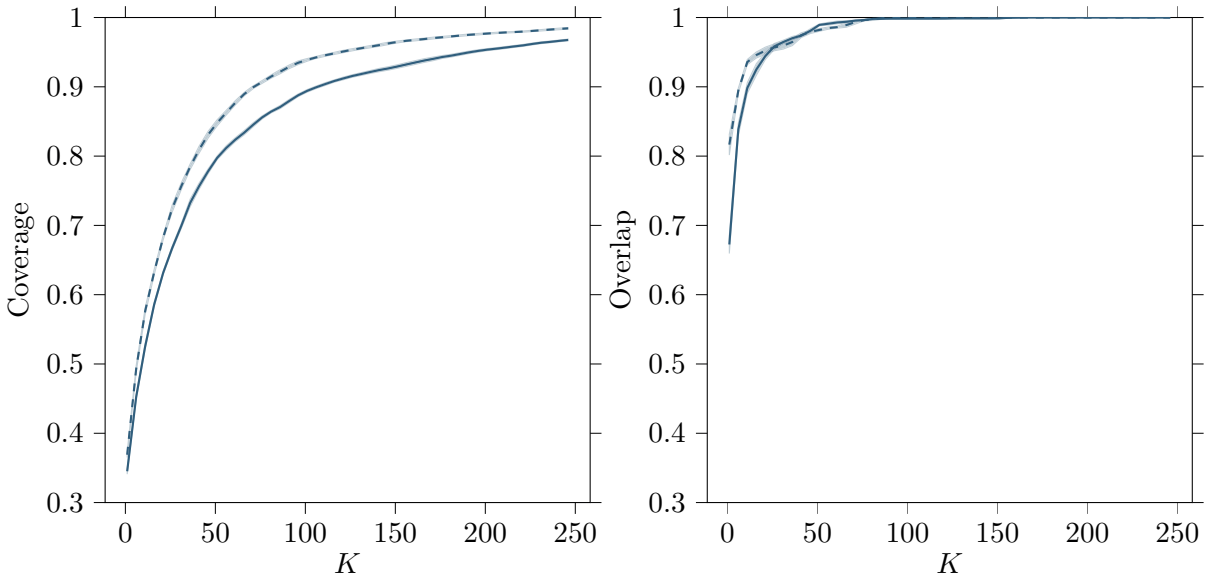


Figure 6.9: DR-RBA: (a) Coverage of available treatment options and (b) ratio of neighbors overlapping the actually applied therapy. *Gower similarity* without (——) and with (- - -) applying attribute selection and weighting are compared. *Coverage* and *overlap* are computed for a neighborhood size range $K \in [1, 250]$.

weight. A straightforward interpretation, which agrees with the conventional CF results shown in section 6.1.1.1, is that therapy history and previous treatment outcome bear crucial information about a patient and potentially effective treatments. Here, especially ADEs, effectiveness and the summarizing *affinity* score of previously applied systemic therapies are assigned large values. $\Delta PASI$ and the actual attending physician's recommendation, on the other hand, are apparently less important. Here, a minor correlation of attribute weights with the frequency of therapy application as shown in figures B.6, B.7 and B.8 is noticeable.
Concerning patient data, especially demographic information, diagnosis information, but also PASI gain comparable large weights. More detailed, especially gender, age, planned child and diagnosed *Psoriasis arthritis* and nail changes are considered to be important. But also family diagnosis and the year of the first diagnosis, which can be seen as an indicator for the length of time the patient is already under treatment, are assigned large weights. Furthermore, especially the weights of some comorbidities such as arterial hypertension, metabolic and mental diseases are particularly striking. Again, the observed weights, however, correlate with the overall occurrence of comorbidities in the data.

In order to evaluate the learned weights, six dermatologists rated the available attributes concerning importance for treatment decisions. The five step ordinal rating scale ranges from *not important at all* to *absolutely essential*. *Spearman's rank correlation* coefficient is computed among all experts and the RBA weights as shown in figure 6.10. In general, correlation between experts and RBA weights are low. As can be expected, the largest correlation is shown for the dermatologist providing the data (*Expert 1*). However, also among the experts correlation is also only moderate in most cases.

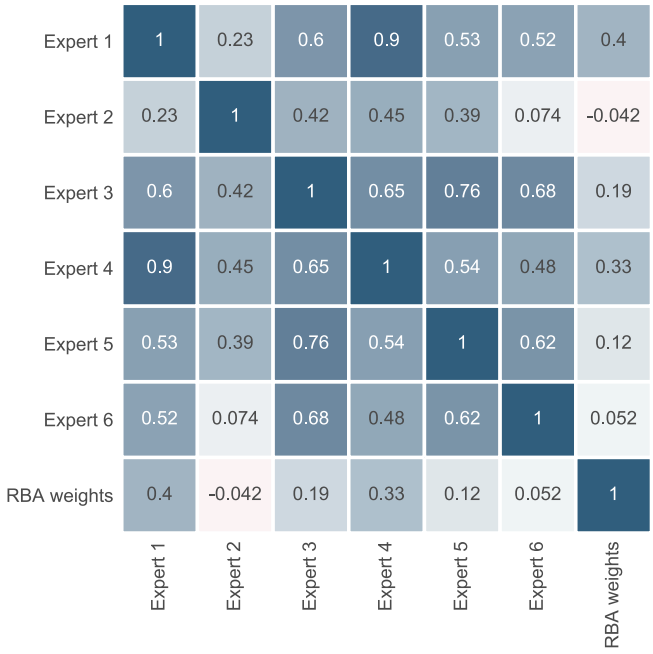


Figure 6.10: Inter-rater agreement concerning importance of attributes for treatment decisions. Estimated importance (RBA weights) and expert ratings are compared.

6.1.1.6 Metric Learning (DR-LMNN)

Instead of only scaling attributes, the LMNN metric learning algorithm described in section 5.4 learns a transformation matrix $\mathbf{L}$ for each outer cross-validation iteration on the basis of the training dataset. *Euclidean distance* is applied to the transformed data in order to compare consultation representations. Two free parameters, additional to the CF neighborhood size K, must be defined: the LMNN neighborhood size K_{LMNN}, which determines the included *target neighbors* and *impostors*, ν, which controls the impact of the competing objectives ϵ_{pull} and ϵ_{push}, and learning rate μ. The best hyperparameter configuration is determined in the inner cross-validation loop (grid search) suchlike that RMSE is optimized respecting the $overlap \geq 0.95$ criterion. Best results could be found for $K_{LMNN} = 10$, $\nu = 0.5$, and $\nu = 0.001$ for the entire range of evaluated K.

When comparing the RMSE curves of the *Euclidean distance* patient-data CF with and without data transformation, the DR-LMNN clearly outperforms the basic approach. The error between estimated and observed outcome quickly approaches a minimum of $RMSE = 0.13$, which is the smallest inner cross-validation error observed for all studied patient-data CF approaches. RMSE only slowly rises with increasing K as can be seen in figure 6.14 (a). *Overlap* is, as the version with consultations represented in the original attribute space, rather large already for small K as shown in figure 6.15 (b). This results in a small neighborhood size K. Furthermore, a large ratio of overlapping treatments with the retrieved neighboring consultations which coincides with small RMSE values is a clear indicator for a meaningful neighborhood. Also the ranked list of treatment options benefits from metric learning. The patient-data CF approach incorporating data transformation is superior to the basic approach not only regarding achievable MAP@3 maximum but maintains a comparable high score for the entire studied range of K. However, what can be observed in figure 6.15 (a), DR-LMNN *coverage* is exceeding the basic *Euclidean distance* patient-data CF especially for rising K. Considering the large MAP@3 score, the patient-data CF algorithm applying data transformation is obviously including more options into the recommendation list, however, is simultaneously capable of ranking those options in accordance with the attending physician's successful choices. Also the scores yielded at the selected neighborhood sizes K, which are summarized in table 6.6, confirm the superiority of the metric learning approach. Particularly the MAP@3 score is the largest compared to all methods according to the inner cross-validation loop results.

Table 6.6: DR-LMNN: Inner cross-validation results (5-fold cross-validation). Best K for which RMSE is minimal and the $overlap \geq 0.95$ criterion is met. Additionally, average and standard deviation of the evaluation metrics RMSE, MAP@3, *coverage* and *overlap* are shown.

Metric	K	RMSE	MAP@3	Coverage	Overlap
DR (Euclidean)	36.00 (7.75)	0.20 (0.00)	0.54 (0.04)	0.69 (0.04)	0.97 (0.01)
DR-LMNN (Euclidean)	25.09 (3.35)	0.14 (0.00)	0.70 (0.02)	0.64 (0.03)	0.96 (0.01)
Average efficiency	- (-)	0.29 (0.00)	0.30 (0.02)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	- (-)	0.20 (0.00)	1.00 (0.00)	1.00 (0.00)

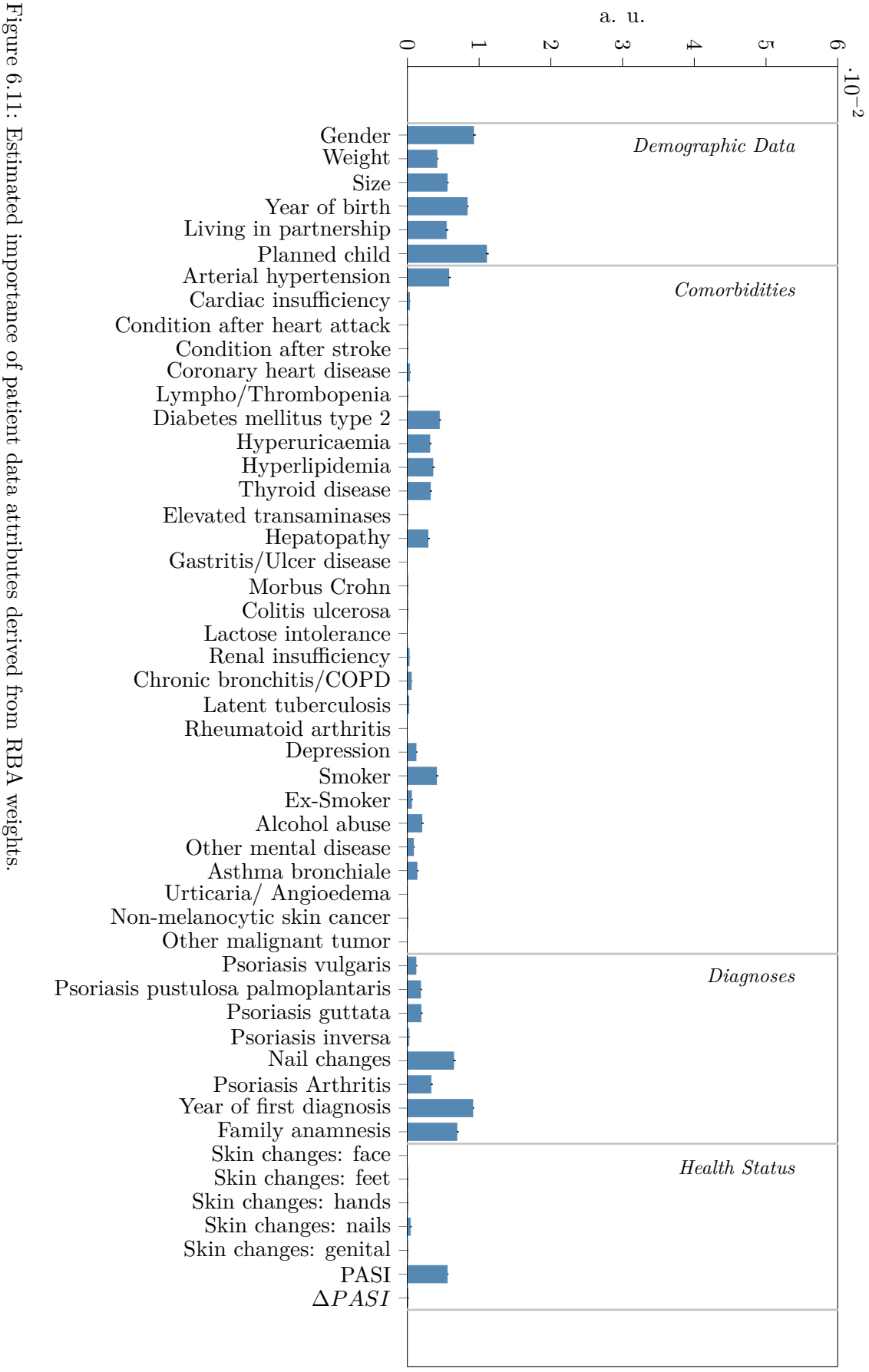


Figure 6.11: Estimated importance of patient data attributes derived from RBA weights.

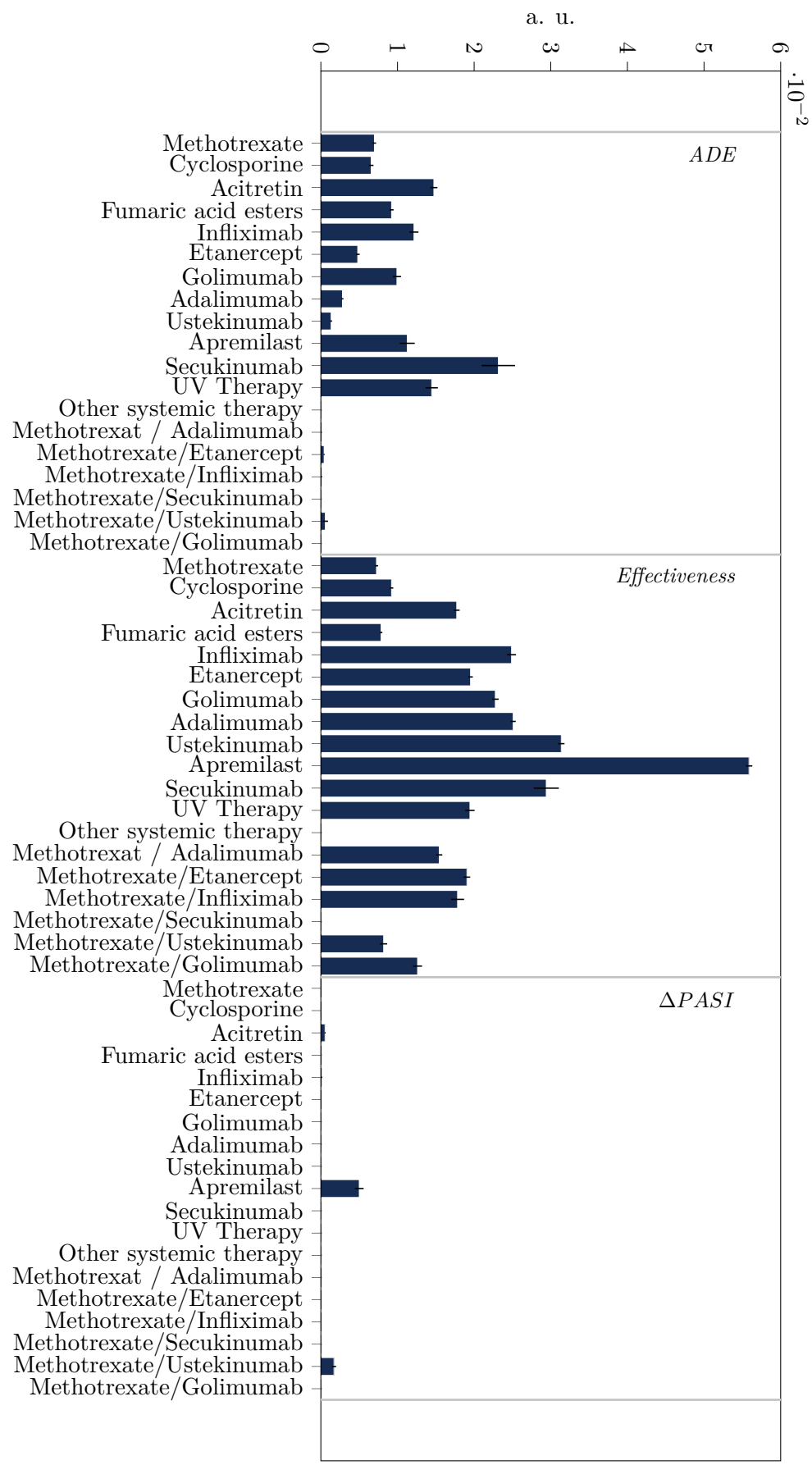


Figure 6.12: Estimated importance of previous treatment outcome attributes (ADEs, effectiveness, $\Delta PASI$) derived from RBA weights.

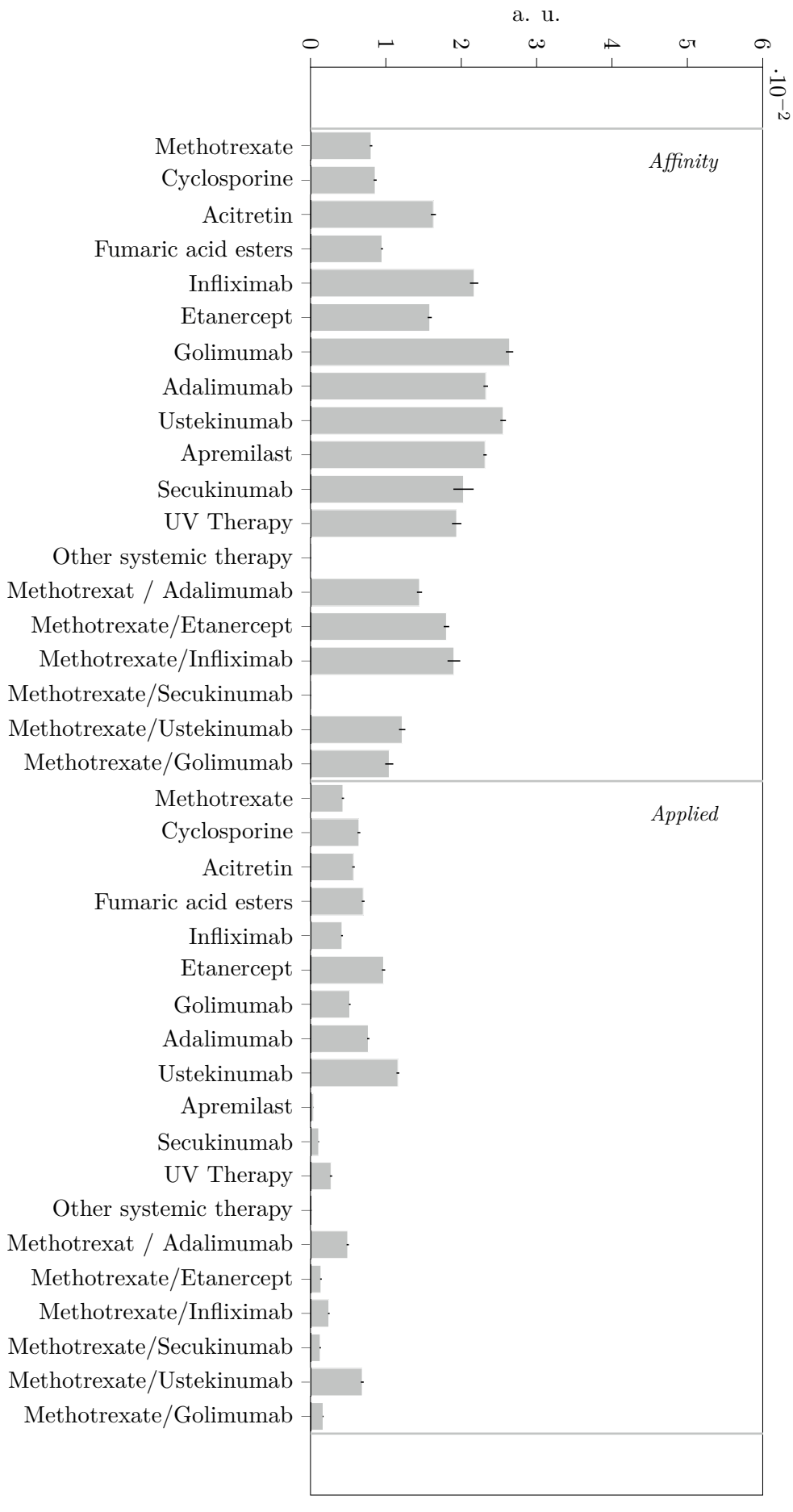


Figure 6.13: Estimated importance of previous treatment outcome attributes (*affinity* score) and previous treatment decision attributes (applied systemic therapy) derived from RBA weights.

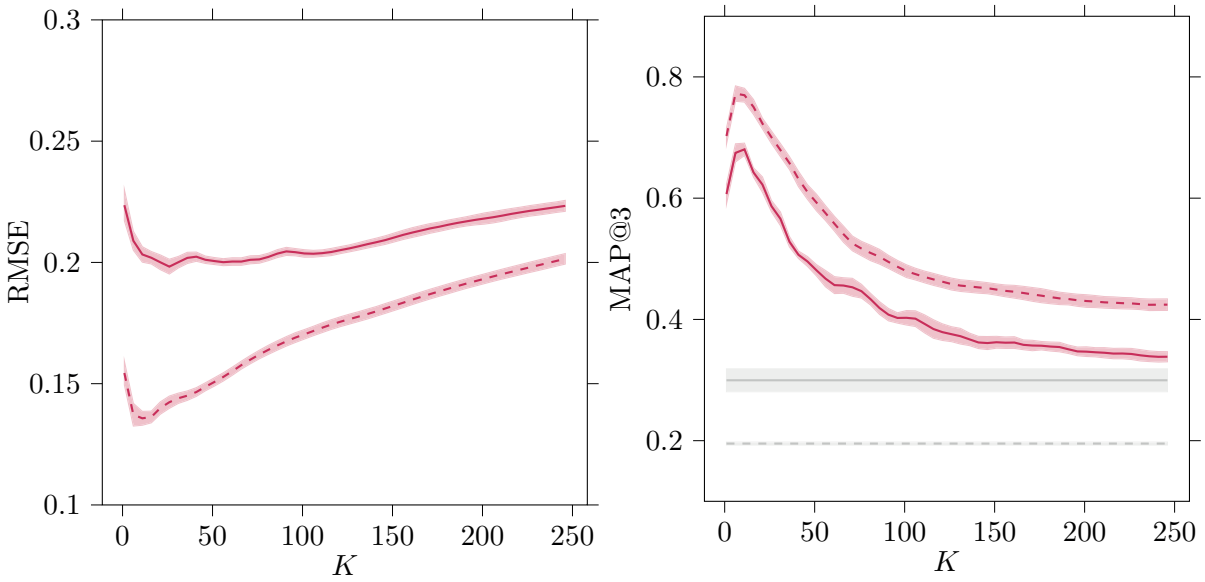


Figure 6.14: DR-LMNN: (a) RMSE between estimated and observed outcome and (b) MAP@3 evaluating the ranked list of recommended therapies. *Euclidean distance* without (——) and with (- - -) applying linear transformation to the data are compared. Additionally, outcome prediction and treatment recommendation based on *average efficiency*, i.e. *affinity* score of each treatment option averaged over all training consultations (——), and *overall popularity*, i.e. frequency of application in the training consultations (- - -), are shown. RMSE and MAP@3 are computed for a neighborhood size range $K \in [1, 250]$.

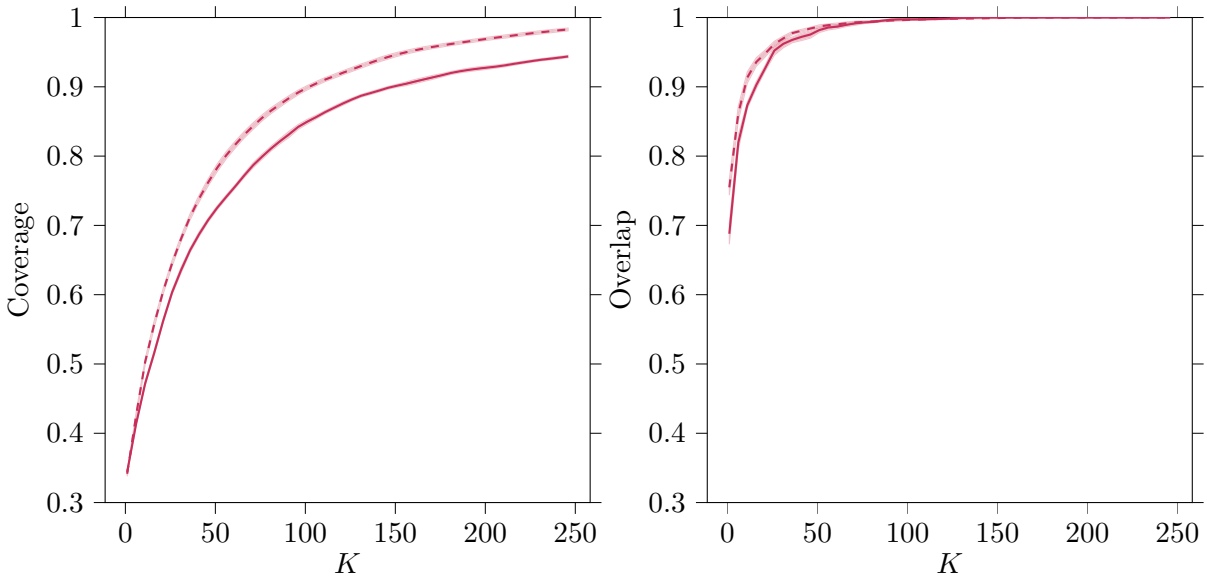


Figure 6.15: DR-LMNN: (a) Coverage of available treatment options and (b) ratio of neighbors overlapping the actually applied therapy. *Euclidean distance* without (- - -) and with (——) applying linear transformation to the data are compared. *Coverage* and *overlap* are computed for a neighborhood size range $K \in [1, 250]$.

Comparable to the RBA approach, metric learning apparently significantly improves both, the prediction accuracy and ranking capability of a neighborhood-based CF. The similarity measures underlying the algorithm become more meaningful and hence powerful selecting and weighting the appropriate consultations. Furthermore, the assumptions regarding similarity, which provide the ground truth for this demonstrated supervised learning methods, prove to be valid.

6.1.2 Sparse Linear Model (SLIM)

In contrast to all previous methods, the linear regression model which is investigated in the following can be considered an item-based approach which learns one distinct model for each treatment option. In order to prevent overfitting by keeping the learned coefficients balanced but also to reduce the overall size of the attribute space, an *elastic net regularization* approach is applied as introduced in 5.4. The impact of the two regularization terms L_1-norm and L_2-norm are controlled by the free parameters β and λ, respectively. As the case with the RBA and LMNN approaches, those parameters are determined within this inner cross-validation loop. The parameters found for each model are summarized in table 6.7.

As can be clearly seen in table 6.7, the performance concerning outcome prediction differs among therapy models. On average, RMSE of the inner cross-validation loops is comparable with the patient-data CF (DR). Nevertheless, some therapy models are capable of achieving clearly superior results, such as the models of *Etanercept*, *Infliximab*, *Adalimumab*, or *Ustekinumab*. Others, such as the *Secukinumab* and *Fumaderm* models, are distinctly inferior to all other methods studied. Also average MAP@3 is in the same range as the patient-data CF even after applying attribute weighting or transformation. A noteworthy feature is the always 100 % *coverage* and *overlap*, as each model produces an outcome estimate for almost every input.

Table 6.7: SLIM: Inner cross-validation results (5-fold cross-validation). Best λ and β for which RMSE on average is minimal for the 5-fold cross-validation, i.e. inner cross-validation loop. Additionally, the average results and standard deviations of RMSE for each therapy model and overall average RMSE and MAP@3 are shown.

Method	λ (10^{-2})	β (10^{-2})	**RMSE**	**MAP@3**
SLIM	- (-)	- (-)	0.18 (0.00)	0.68 (0.02)
Methotrexat	0.75 (0.00)	0.25 (0.00)	0.19 (0.01)	- (-)
Fumaderm	0.47 (0.11)	0.53 (0.11)	0.22 (0.01)	- (-)
Infliximab	0.34 (0.33)	0.15 (0.19)	0.16 (0.02)	- (-)
Etanercept	0.75 (0.02)	0.25 (0.02)	0.15 (0.00)	- (-)
Adalimumab	0.28 (0.28)	0.72 (0.28)	0.16 (0.01)	- (-)
Ustekinumab	0.73 (0.07)	0.27 (0.07)	0.16 (0.00)	- (-)
Secukinumab	0.74 (0.05)	0.26 (0.05)	0.23 (0.01)	- (-)
MTX/Infliximab	0.20 (0.26)	0.54 (0.40)	0.19 (0.04)	- (-)
Average efficiency	- (-)	- (-)	0.29 (0.00)	0.30 (0.02)
Overall popularity	- (-)	- (-)	- (-)	0.20 (0.00)

Considering the inner cross-validation results, learning coefficients directly from the data to model outcome appears to be a promising strategy. Beyond accuracy and quality of the recommendation list, it must be kept in mind that the actual outcome prediction and recommendation process is significantly more effective when utilizing a model-based approach. Nonetheless, these results must be treated with caution, as they are partly based on a very small database. This is especially true e.g. for the *MTX/Infliximab* model. Here, each fold on average only comprises 6 observations, meaning that this model is trained and evaluated on 24 and 6 observations only, respectively.

6.1.3 Gradient-boosted Regression Trees (GBM)

Comparable to the previous methods, the GBM studied in the following can be considered an item-based approach which learns one distinct regression model for each treatment option. As was already detailed in 5.5, many GBM hyperparameters are defined based on preliminary investigations. The remaining free parameters, namely the number of trees to fit n_{trees}, the maximum base learner tree depth d_{max} and the minimum sum of instance weights w_{child}, are selected in the inner cross-validation loop. The parameters found for each model are summarized in table 6.8.

Table 6.8: GBM: Inner cross-validation results (5-fold cross-validation). Best n_{trees}, d_{max} and w_{child} for which RMSE on average is minimal for the 5-fold cross-validation, i.e. inner cross-validation loop. Additionally, the average results and standard deviations of RMSE for each therapy model and overall average RMSE and MAP@3 are shown.

Method	n_{trees}	d_{max}	w_{child}	**RMSE**	**MAP@3**
GBM	- (-)	- (-)	- (-)	0.14 (0.00)	0.40 (0.02)
Methotrexat	59.67 (12.18)	2.69 (1.30)	6.39 (0.94)	0.11 (0.01)	- (-)
Fumaderm	25.00 (0.00)	2.49 (1.29)	5.93 (1.74)	0.19 (0.01)	- (-)
Infliximab	48.48 (7.03)	4.01 (0.26)	3.01 (0.15)	0.14 (0.01)	- (-)
Etanercept	52.90 (19.54)	2.08 (0.53)	3.18 (0.67)	0.14 (0.01)	- (-)
Adalimumab	34.39 (12.39)	2.83 (1.56)	6.93 (0.36)	0.11 (0.01)	- (-)
Ustekinumab	29.70 (9.76)	2.04 (0.42)	6.16 (1.26)	0.14 (0.00)	- (-)
Secukinumab	25.00 (0.00)	2.29 (0.90)	3.45 (1.15)	0.18 (0.01)	- (-)
MTX/Infliximab	64.78 (19.99)	3.10 (0.99)	5.83 (1.01)	0.16 (0.04)	- (-)
Average efficiency	- (-)	- (-)	0.29 (0.00)	0.30 (0.02)	
Overall popularity	- (-)	- (-)	- (-)	0.20 (0.00)	

Equally to the SLIM results, the performance concerning outcome prediction, listed in table 6.8, varies greatly among models. However, compared with the SLIM approach, the individual treatment GBM models provide more accurate outcome predictions. The overall outcome prediction accuracy is, according to the inner cross-validation results, even superior to most of the CF methods. Nevertheless, for treatments for which only inferior SLIM models could be provided (e.g. *Secukinumab* and *Fumaderm*), it is also only possible to model inferior GBM models. With regard to ranking capabilities, however, the GBM shows very weak performance.

To conclude, the GBM's capability to model non-linear relationships within the data benefits RMSE and very accurate outcome predictions can be yielded for some therapy models. This is apparently even true in spite of the limited amount of training data the modeling is based on. Nonetheless, also these results must be considered with caution as partly based on only very little training and validation data. In connection with the GBM approach it becomes particularly clear that a model providing accurate outcome predictions not necessarily provides large MAP@3 scores. Under the assumption of a reliable model, this indicates that in many cases not the most suitable therapy options in terms of *affinity* score optimization was actually applied. Beyond that, as already mentioned above, the model-based approaches in general come along with 100 % *coverage* of therapy options, which are ranked according to predicted outcome, and 100 % *overlap*, which RMSE computation is based on.

6.2 Generalization Performance Evaluation

When considering the outer cross-validation results summarized in table 6.9 and visualized in figures 6.16 and 6.17, especially the large variance of the results becomes apparent. Within each outer cross-validation loop, almost all consultations are available as training data. Only those of test patient p are excluded. Hence, the applied leave-one-patient-out cross-validation approach is, on the one hand, assumed to be almost unbiased. However, the major downside of many small folds is, on the other hand, 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 figures 6.17, is remarkably large and partly spread over the entire value range. However, it must be considered that the large MAP@3 variance is also owned to the characteristic of the evaluation metric itself. As only one actually applied treatment is available for each test consultation, the MAP@3 score for recommendations which meet the ground truth and are ranked among the top-3 can become 1, 0.5, or 0.33, respectively. Actually applied therapies not ranked among the top-3 yields a MAP@3 of 0. Consequently, recommendation lists yielding MAP@3 scores of 0.33 can still be regarded as useful recommendations.

The generalization performance, estimated in the outer leave-one-patient-out cross-validation for all of the proposed algorithms is summarized in table 6.9 and figures 6.16 and 6.16.
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

Table 6.9: Outer cross-validation loop results. Mean and standard deviation of outcome prediction accuracy (RMSE), recommendation list agreement (MAP@3), average *overlap* with applied treatment and *coverage* of treatment options.

Method	**RMSE**	**MAP@3**	**Coverage**	**Overlap**
CF (Cosine)	0.17 (0.13)	0.86 (0.23)	0.51 (0.23)	0.94 (0.17)
CF (Pearson)	0.17 (0.13)	0.85 (0.24)	0.52 (0.23)	0.94 (0.17)
CF (Manhatten)	0.14 (0.09)	0.54 (0.32)	0.89 (0.10)	0.91 (0.21)
CF (Euclidean)	0.14 (0.09)	0.55 (0.32)	0.90 (0.10)	0.92 (0.19)
DR (Gower)	0.18 (0.12)	0.61 (0.33)	0.65 (0.21)	1.00 (0.00)
DR-RBA (Gower)	0.15 (0.11)	0.67 (0.32)	0.66 (0.20)	1.00 (0.00)
DR (Euclidean)	0.20 (0.11)	0.59 (0.36)	0.58 (0.24)	1.00 (0.00)
DR-LMNN (Euclidean)	0.19 (0.12)	0.54 (0.30)	0.70 (0.21)	1.00 (0.00)
DR-Rules a (Gower)	0.18 (0.12)	0.59 (0.34)	0.57 (0.21)	1.00 (0.00)
DR-Rules b (Gower)	0.18 (0.12)	0.48 (0.38)	0.48 (0.22)	1.00 (0.00)
DR-Rules c (Gower)	0.18 (0.12)	0.45 (0.39)	0.40 (0.21)	1.00 (0.00)
DR-Impute 0 (Gower)	0.19 (0.10)	0.65 (0.31)	0.66 (0.21)	1.00 (0.00)
DR-Impute 1 (Gower)	0.18 (0.11)	0.59 (0.32)	0.66 (0.21)	1.00 (0.00)
SLIM	0.18 (0.11)	0.68 (0.33)	1.00 (0.00)	1.00 (0.00)
GBM	0.15 (0.10)	0.31 (0.33)	1.00 (0.00)	1.00 (0.00)
Average efficiency	0.29 (0.13)	0.26 (0.30)	1.00 (0.00)	1.00 (0.00)
Overall popularity	- (-)	0.23 (0.35)	1.00 (0.00)	1.00 (0.00)

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, non-parametric, i.e. distribution free tests are used in both cases although having less statistical power than parametric tests. The omnibus test applied within this work which meets the described data properties is the *Friedman test* [119]. The probability distribution of the Friedman test statistic is approximated by the chi-squared distribution. As both, the number of algorithms to be compared ($k = 17$) and the number of included partitions ($n = 175$ and $n = 154$) are sufficiently large, this distribution assumption can be regarded to be valid and provide reliable p-values. In order to identify which groups are significantly different from each other in case of a rejected null hypothesis, the *Wilcoxon signed-rank test* [382] is used. Moreover, in order to counteract the globally increased likelihood of incorrectly rejected null hypotheses, i.e. an increased Family-wise Error Rate (FWER), which arises with multiple simultaneous tests based on equal samples, the *Bonferoni-Holm*-correction is applied [332]. The individual test samples in each outer cross-validation iteration can be regarded identically distributed, however, cannot be considered independent due to overlapping data. As a consequence, the test results may still be overly optimistic and should be interpreted with caution.

Since the *overlap* of many algorithms is less then 100 %, only those patients for whom all al-

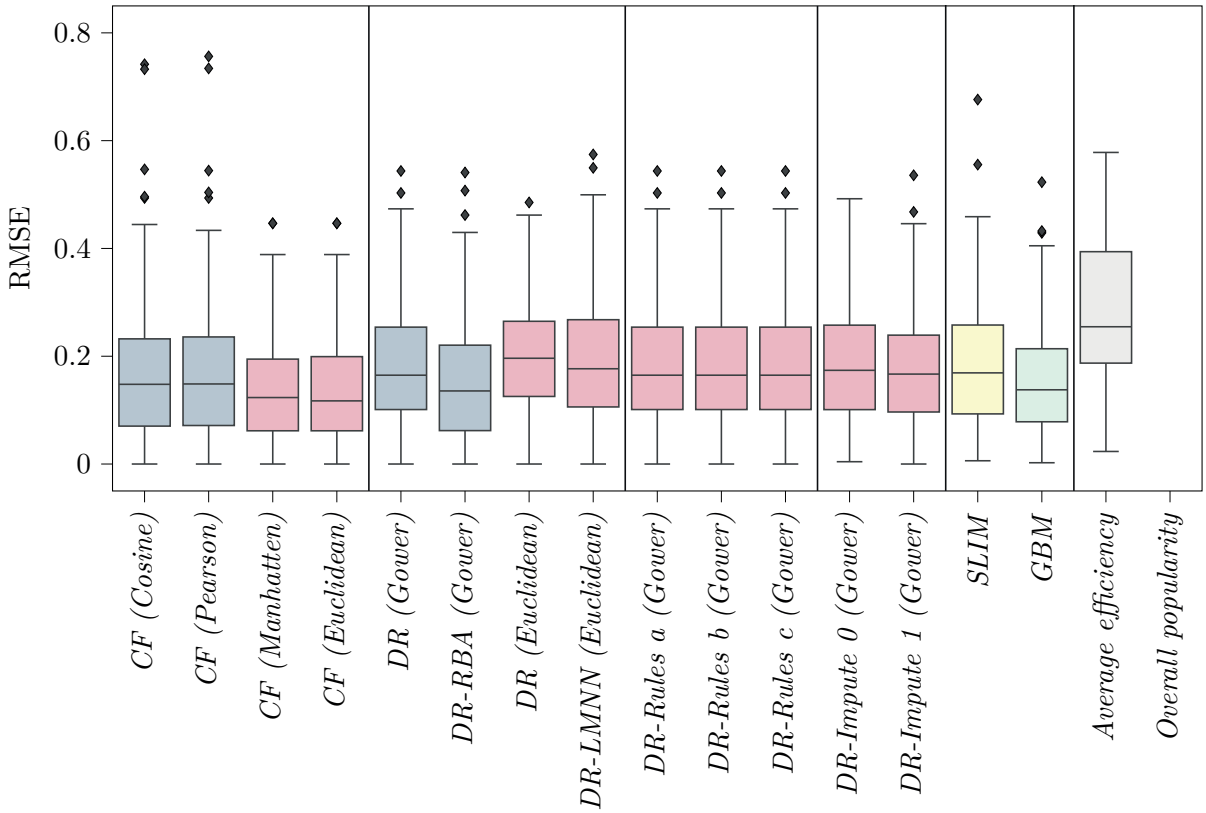


Figure 6.16: Outer cross-validation loop results. Outcome prediction accuracy (RMSE) evaluated for all proposed methods.

gorithms provide RMSE or MAP@3 scores can be considered for statistical testing. As has been verified, imputing the average RMSE or MAP@3 score for each algorithm, respectively, does not change the hypothesis results but only slightly impacts the yielded p-values and renders the test more conservative. Therefore, only the intersection of patients with available RMSE or MAP@3 score are used for the hypothesis testing in the following, encompassing $n = 175$ and $n = 154$ observations, respectively.

Concerning RMSE, the null hypothesis is rejected with test statistic 267.09 and $p = 3.70e-48$ according to the Friedman test. Also regarding the ranking capabilities of the compared algorithms, quantified with the MAP@3 score, significant differences among the evaluated algorithms are evident with test statistic 678.02, $p = 2.03e-133$.

As a conclusion of the findings of the omnibus tests, *Wilcoxon signed-rank tests* are performed in the following on all pairs of algorithms with differing evaluation scores as stated above. RMSE and MAP@3 results are shown in figure 6.18 and 6.19, respectively.

Looking at the results summarized in table 6.9 and p-values in figure 6.18 and 6.19, it becomes obvious that all examined algorithms perform significantly better than the two baseline methods *average affinity* and *overall popularity* in terms of both, outcome prediction and therapy ranking. Hence, it can be concluded that estimating outcome based on local data only is highly beneficial and also the model-based approaches are, in spite of the small training data sizes,

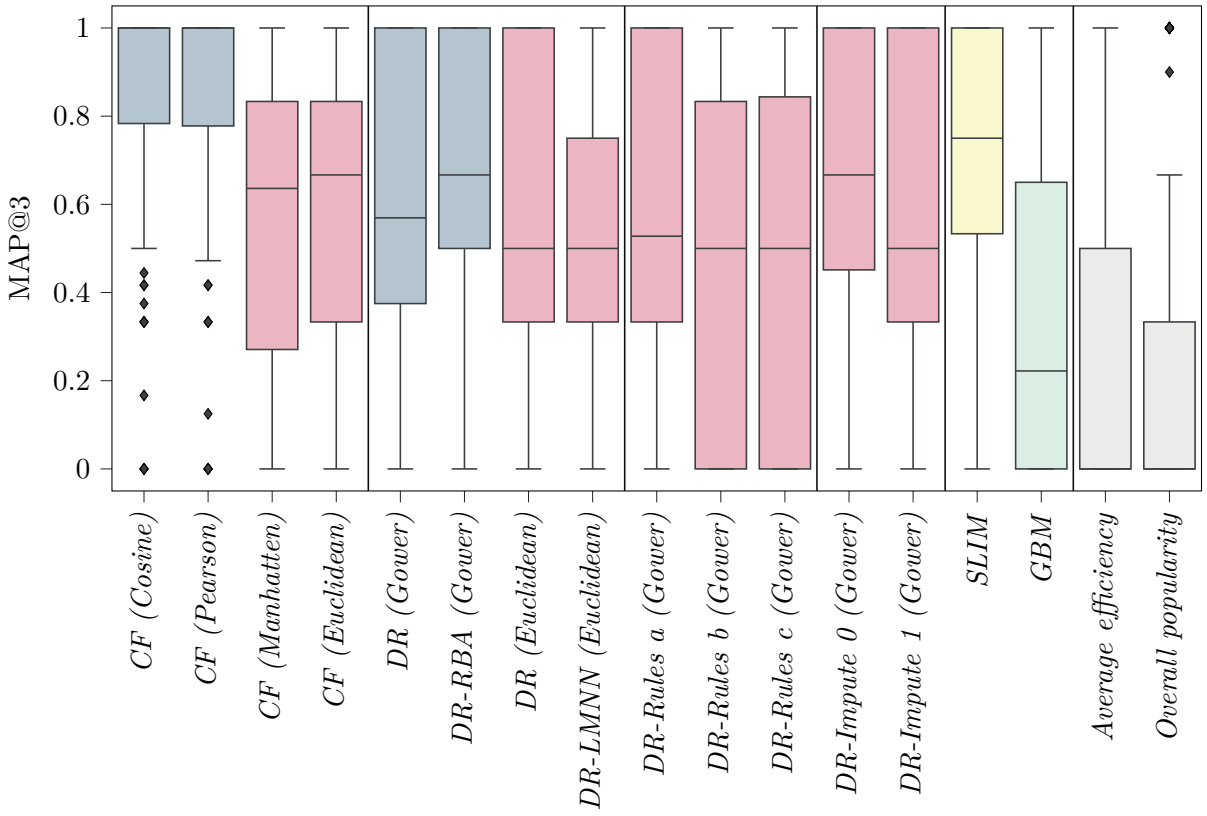


Figure 6.17: Outer cross-validation loop results. Recommendation list agreement (MAP@3) evaluated for all proposed methods.

successful.

In case of the conventional CF, the observed RMSE mean values from the inner cross-validation loop can be reproduced in the outer loop. The generalization performance of the *Minkowski metrics* are superior to the correlation-based similarity measures *Cosine similarity* and *Pearson correlation* coefficient and even outperform all other approaches apart from the *Gower* patient-data CF with attribute weighting and the GBM model. Within the group of *Minkowski metric* approaches, no significant performance difference is evident. Also regarding MAP@3, *coverage* and the *overlap* of prediction and ground truth, inner cross-validation results and estimated generalization performance are comparable concerning the central tendency. Variance of the outer loop results however is, as initially discussed, remarkably large especially for MAP@3. Nevertheless, a statistically significant superiority of the correlation-based conventional CF algorithms over all other evaluated approaches is evident. Within the group of correlation-based methods, no statistically significant difference can be shown. As was already observed for the inner loop, prediction accuracy is improved at the expense of MAP@3 and *vice versa*. One possible explanation is, as already described in section 6.1.1.1, the way how therapy outcomes are treated which have not been applied in common. As large overlap of commonly applied treatments increases similarity in case of *Cosine similarity* and *Pearson correlation*, those approaches are more selective concerning treatments observed in the neighborhood which yields

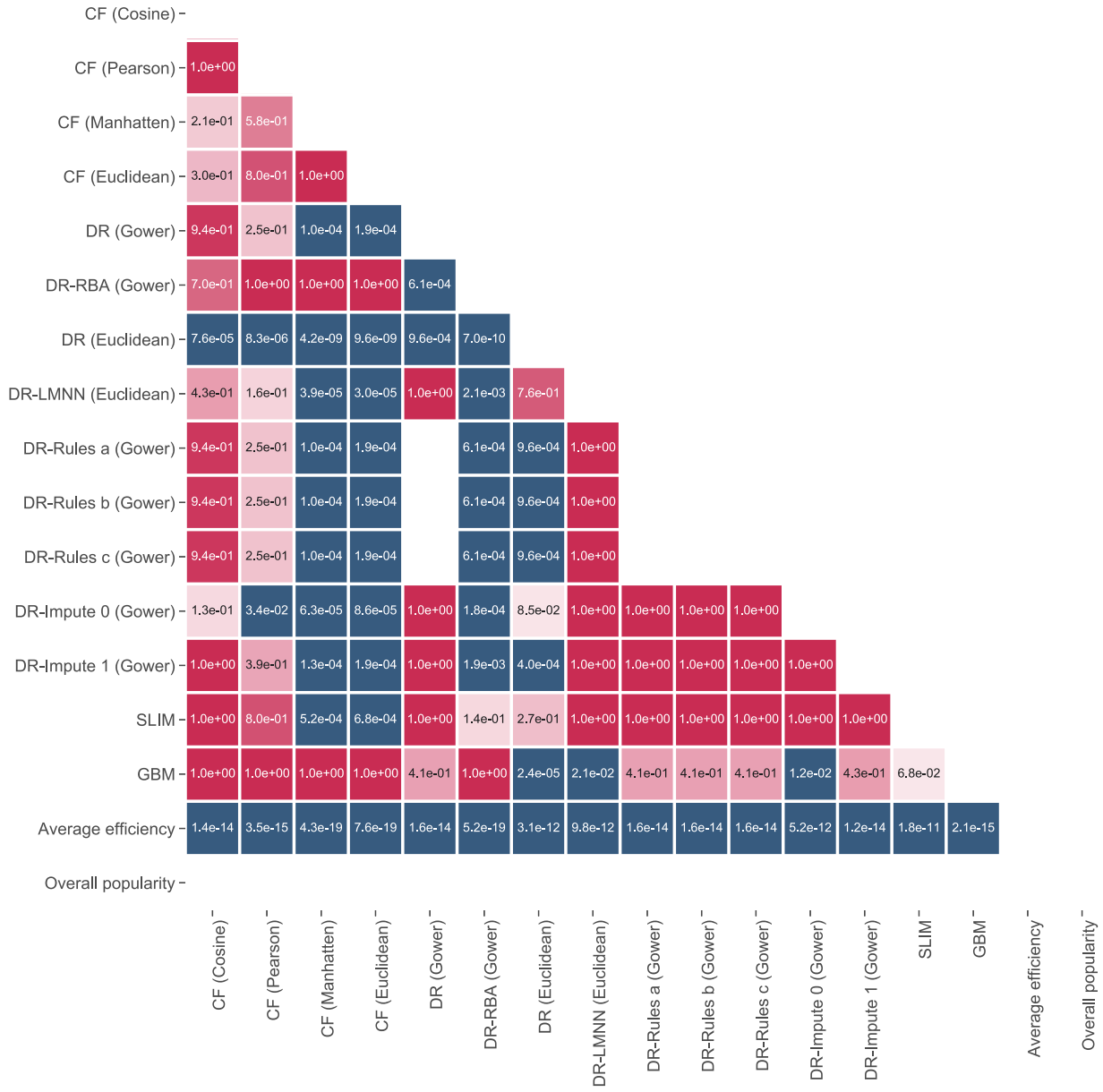


Figure 6.18: p-values of pairwise *post hoc* tests (*Wilcoxon signed-rank tests*), comparing all presented algorithms concerning prediction accuracy (RMSE). Statistical significant performance differences ($p > \alpha$) are colored blue and results from the same distribution are colored red.

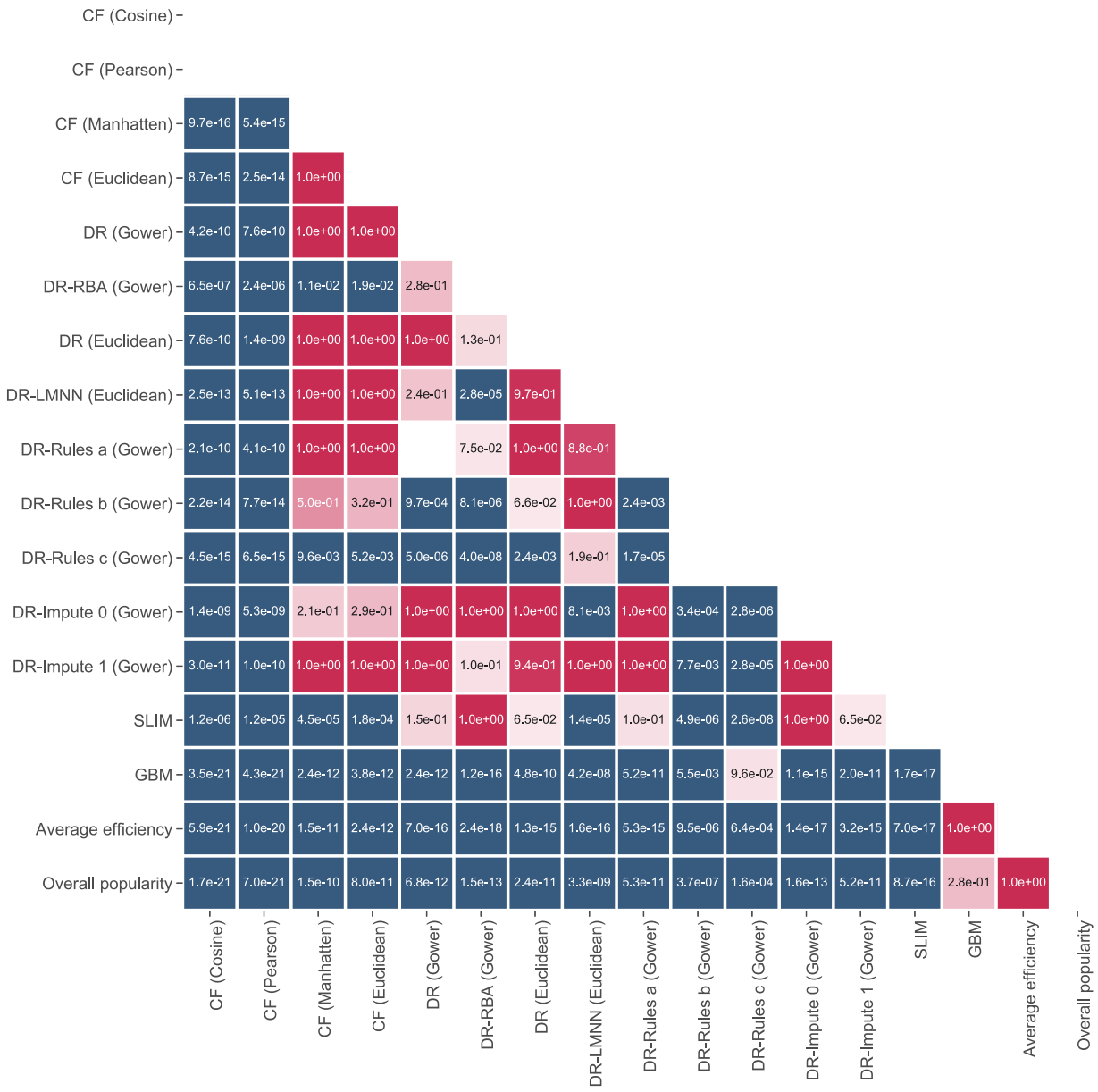


Figure 6.19: p-values of pairwise *post hoc* tests (*Wilcoxon signed-rank tests*), comparing all presented algorithms concerning agreement with the ground truth (MAP@3). Statistical significant performance differences ($p > \alpha$) are colored blue and results from the same distribution are colored red.

larger MAP@3 and lower *coverage*. The *Minkowski metrics*, on the other hand, especially focus on similar outcome when computing similarity which results in small RMSE scores. Those metrics are, however, not sensitive to the number of co-occurring treatments in two vectors to be compared.

According to the patient-data CF generalization performance in table 6.9, *Gower similarity* appears to achieve even better results in the outer than in the inner loop for the selected K in terms of both, outcome prediction and quality of the recommendation list. The patient-data CF approach obviously benefits from the outer loop's larger training data to select the neighborhood from. Similar observations can be also made for the *Euclidean distance* when comparing the outer with the inner cross-validation loop results. Those differences are, however, small and average RMSE and *coverage* are in the same general range as the inner loop results. The *Gower similarity's* superiority over the *Euclidean distance* from the inner cross-validation loop can be confirmed with the generalization performance. Regarding the agreement between ground truth and recommendations, on the other hand, no statistically significant performance difference between *Gower similarity* and the *Euclidean distance* can be shown for the patient-data CF.

Comparing the patient-data CF results from the outer cross-validation loop after applying the proposed RBA algorithm, the benefits of linear attribute scaling shown during model selection can be, to a reduced extent, reproduced in terms of RMSE and MAP@3. However, the improvement of the *Gower similarity* baseline is only statistically significant for MAP@3.

In case of the LMNN approach, the large improvement due to data transformation observed in the inner cross-validation loop cannot be reproduced in the outer loop. No significant performance differences can be shown for RMSE and MAP@3. Both evaluation criteria, even tend to drop in comparison with the *Euclidean distance* baseline. This observation can be explained with a optimistically biased inner cross-validation result as the transformation matrix $\mathbf{L}$ is optimized using the entire training partition in each outer cross-validation iteration p. Thus, also the evaluation partitions of the inner cross-validation loop are contained in this outer loop training partition, which potentially causes overfitting. However, it can be assumed that when using a more comprehensive dataset, the ability to generalize the underlying data patterns increases and also the generalization performance of the LMNN optimization algorithm improves and approaches the results of the inner loop.

The slight benefits of the imputation strategy which is demonstrated in the inner loop cannot be generalized by the leave-one-patient-out cross-validation results. Both, MAP@3 scores and RMSE resulting from applying the raw version (*impute 0*) and any of the imputation strategies *impute 1* and *impute 2*, stem from the equal distributions. Hence, no advantages can be proven.

In contrast, in case of the post-filtering by exclusion rules, the MAP@3 deteriorations of the *Gower similarity* patient-data CF are statistically significant. As already discussed for the inner cross-validation loop, especially *rules a* and *rules c* obviously comply with the underlying ground truth as both only have minor impact on the recommendation list's agreement with the attending physician. In contrast, *rules b* clearly deteriorates the MAP@3 results. Nevertheless, also according to the generalization results, none of the applied exclusion rules is capable of improving the recommendation list as could be expected. Note that the resulting MAP@3

scores of not applying exclusion rules (DR (Gower)) and applying *rules a* as well as the MAP@3 scores of applying *rules b* and *rules c* are partially identical over a wide range of outer loop iterations. Therefore, this test statistic cannot be computed. Moreover, because the application of exclusion rules has no impact on RMSE results, also no comparison among the *Gower similarity* patient-data CF and the versions with exclusion rules are computed.
Finally, as is also shown in table 6.9, also the two applied ML algorithms, SLIM and GBM are comparable to the inner loop results regarding both outcome scores. Here, the trade-off between outcome prediction performance and recommendation list agreement becomes particularly apparent. Whereas SLIM provides outcome predictions comparable to the correlation-based conventional CF and the patient-data CF approaches, this linear model is capable of outperforming the *Minkowski metric* conventional CF in terms of MAP@3. In contrast, the decision tree ensemble achieves comparatively small mean prediction errors which, however, differ only statistically significantly from the *Euclidean distance* patient-data CF. The quality of the ranked therapy list hardly outperforms the *overall popularity* baseline.

6.3 Comparison with Expert Performance

As introduced in section 5.2, the performance of the proposed algorithms and system variants is further compared with the recommendations of human experts. Therefore, the subset of 100 test consultations from different patients and dermatologists' recommendations described in section 4.6 are utilized. As mentioned before, the test dataset comprises 74 consultations in which therapy was actually changed and 26 without change. Overall, therapy recommendations, alternative to the given ground truth, are available from six experts (Dermatologists from different clinics in Germany). In the following, however, only those four experts are included which rated all 100 test consultations. Each expert was asked to prioritize up to three therapy recommendations selected from 20 unique options. These recommendations form top-3 ranked recommendation lists as is output by the recommender system.

In the following, only a selection of the most successful CF approaches are included, namely *Cosine similarity* and *Euclidean distance* conventional CF and the patient-data CFs with and without attribute weighting and attribute space transformation. Training a ML model for each of the 20 unique treatment options is unreliable due to partially only few samples. Comparable to the outer cross-validation loop, an individual CF model is optimized and selected for each of the test consultations. Models are selected on the basis of a 5-fold cross-validation and according to the same criteria as applied in section 6.1. To quantify agreement with the ground truth MAP@3 and *Cohen's Kappa* scores from the highest priority only (κ_1) and either of the recommendations (κ_{all}) are computed. Table 6.10 summarizes the resulting scores from the 100 test consultations in comparison with expert performance and baseline results.

The given results clearly indicate that none of the presented algorithms is capable of providing recommendations with comparable agreement as the human experts. According to this comparison, the *Euclidean distance* conventional CF is the most promising approach which, however, is still only capable of ranking the actually applied treatment on average in 0.54 (0.50) cases among

Table 6.10: Recommendation performance (MAP@3) of the selected algorithms compared to the recommendations of human experts.

Metric	**RMSE**	**MAP@3**	**Coverage**	**Overlap**	κ_1	κ_{all}
CF (Cosine)	0.27 (0.22)	0.37 (0.45)	0.46 (0.21)	0.66 (0.48)	0.12	0.26
CF (Euclidean)	0.17 (0.17)	0.40 (0.43)	0.89 (0.10)	0.90 (0.30)	0.12	0.36
DR (Gower)	0.22 (0.14)	0.28 (0.37)	0.44 (0.15)	0.86 (0.35)	0.04	0.35
DR-RBA (Gower)	0.20 (0.16)	0.34 (0.39)	0.59 (0.17)	0.84 (0.37)	0.04	0.35
DR (Euclidean)	0.22 (0.13)	0.27 (0.36)	0.45 (0.16)	0.82 (0.39)	0.03	0.37
DR-LMNN (Euclidean)	0.22 (0.16)	0.32 (0.38)	0.55 (0.14)	0.83 (0.38)	0.03	0.29
Expert 2	-	0.44	0.14	-	0.30	0.53
Expert 4	-	0.47	0.09	-	0.33	0.49
Expert 5	-	0.43	0.11	-	0.33	0.43
Expert 6	-	0.48	0.10	-	0.35	0.53

the top-3 recommendations (Precision@3). The overall significantly lower MAP@3 scores compared with the estimated generalization performance from section 6.2 has two obvious origins. Firstly, in this experiment 20 instead of 8 therapy options are available for selection. Hence, the overall probability to select the ground truth option is decreased. Secondly, the subset of 100 test consultations can be regarded particularly difficult cases as the majority contain therapy changes compared to the previous consultation. Note that the proportion of consultations in the overall data that contains therapy changes is only 14.24 %. Also noteworthy is the much lower *coverage* with which human experts meet the ground truth. Given that experts give three recommendations per case, *coverage* would amount to 15.00 %. As, however, the number of recommendations decreases with decreasing priority, the true average *coverage* value is even lower with 10.93 %.

7 Further Applications

This chapter summarizes additional own studies addressing further applications and extensions of the proposed therapy recommender system approach. On the one hand, quantification of health status and outcome based on raw vital signs for various conditions is studied in section 7.2 and section 7.3. On the other hand, sentiment analysis methods are applied to patient reviews to extract information on experience with applied treatments in section 7.4. Some of the results described in this chapter are published in [135] and [125].

7.1 Introduction

Successful management and treatment of diseases relies on monitoring outcome such as the effectiveness of therapies. The objective quantification of health status and treatment success by means of clinical scores and parameters is a prerequisite. Such scores and parameters can either be based on questionnaires or derived from vital signs, biosignals, or other markers. Figure 7.1 extents the therapy recommender system inputs by such raw data.

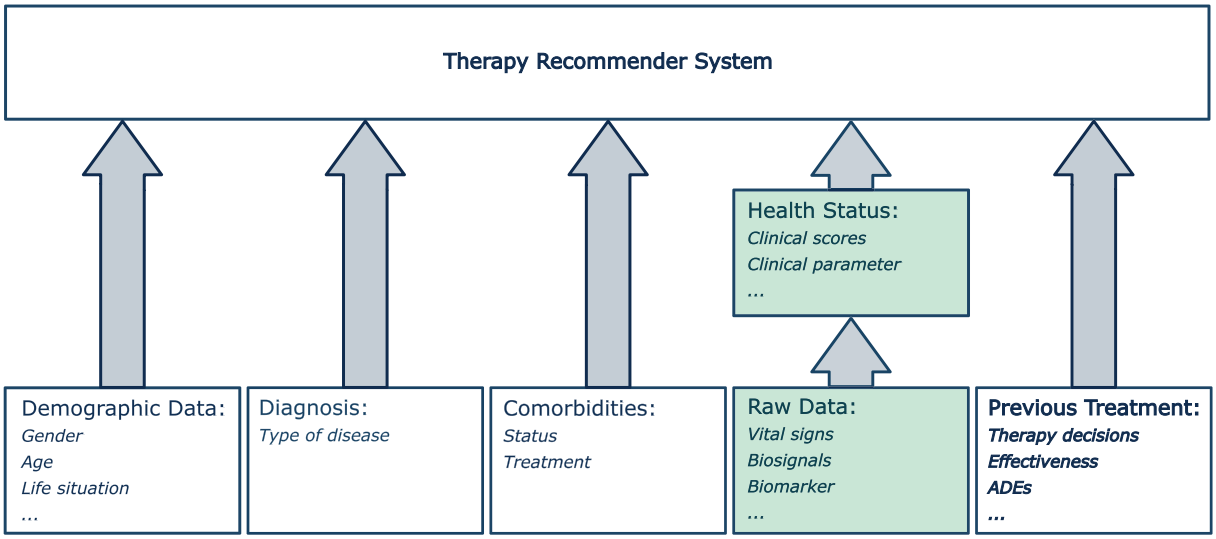


Figure 7.1: Extension and generalization of a therapy recommender system's input.

Moreover, as already mentioned in chapter 1, an additional source of information regarding treatment success and patient satisfaction can be available in the form of written text. To reveal such information, NLP techniques, namely sentiment analysis, can be employed to patient reviews in order to assess experience with applied treatments.

7.2 Sleep Stage Classification

7.2.1 Introduction

Sleep plays a vital role regarding health and wellbeing. In addition to its recovery function, e.g. in the form of an increase in growth hormone levels, human sleep plays a central role in the consolidation of memory and learning as well as in the maintenance of the immune system [331]. However, there is a multitude of sleep disorders, most of which are characterized by daytime sleepiness, difficulties in falling asleep or staying asleep, or the occurrence of abnormalities during sleep. The long-term health consequences of such disorders include an increased risk of high blood pressure, diabetes, obesity and depression as well as an increased risk of heart attack and stroke. Monitoring sleep to prevent and treat serious health problems is therefore crucial. [146]

On average, adult humans spend approximately seven to eight hours a day sleeping. A healthy eight hour sleep contains four or five sleep cycles, each lasting approximately 90 minutes and containing different stages including light sleep (N1 and N2) and deep sleep (N3) NREM sleep stages and rapid eye movement (REM) sleep [331, 187]. The different sleep stages are characterized in table 7.1 along with typical Electroencephalogram (EEG) characteristics. An exemplary *hypnogram*, visualizing the sequence of sleep stages, is shown in figure 7.2 (a). Besides age and individual circadian rhythm, pathological findings and medication intake have significant impact on the sleep architecture [331, 187]. Therefore, in order to assess sleep quality, diagnose sleep disorders and assess treatment outcome, sleep stage classification plays an important role.

Table 7.1: Terminology and characteristics of sleep stages according to the *The AASM Manual for the Scoring of Sleep and Associated Events* [70]

AASM	**Description**	**Characteristics**
W	Wake	Alpha and beta wave activity
N1	Snoozing	Theta wave activity Slow eye movements Declining muscle tone (< W)
N2	Stable sleep	Sleep spindles, K-complexes No eye movements Declining muscle tone (< N1)
N3	Deep sleep	Fraction of delta waves > 20% No eye movements Declining muscle tone (< N2)
REM	Rapid-Eye-Movement	Theta wave activity, Saw-tooth waves Rapid eye movements Lowest average muscle tone

Gold standard in sleep medicine is the Polysomnography (PSG), which includes the recording

of a series of biosignals, such as the EEG, the Electrooculogram (EOG), and the Electromyogram (EMG). Often the Electrocardiogram (ECG), peripheral pulse oximetry as well as respiratory airflow and effort measurements are also recorded. Whole night PSG recordings are segmented into 30 s epochs and manually scored by sleep experts according to standardized guidelines, i.e. the American Association of Sleep Medicine (AASM) Scoring Manual [70]. As the manual PSG scoring is a laborious process and is subject to the experts personal experience and condition (inter-rater agreement $\kappa \approx 0.68$ [77]), automatic sleep stage classification is of great interest and the object of current research [34].

Beyond that, in order to reduce the impact of data acquisition on the patient's sleep, simplify the attachment of sensors and electrodes, and even facilitate home monitoring, reliable sleep stage classification with a reduced number of biosignals is investigated. Besides reducing the number of EEG channels [42], the assessment of sleep quality based on heart rate and respiration only is of special interest as both can be retrieved at various positions and with comparably little impact [166, 278].

7.2.2 Background and Related Work

As fundamental visceral functions, the cardiovascular and respiratory system are regulated by the Autonomous Nervous System (ANS). Due to the mutual activation of the sympathetic and parasympathetic branch, the ANS also plays a central role in the physiology of sleep. The sympathovagal balance of the ANS shows a profound variability related to sleep stages. [267, 353] Consequently, characteristic changes in cardiorespiratory parameters can be a basis to differentiate between sleep stages. Previous studies have focused on the analysis of Heart Rate (HR) and Heart Rate Variability (HRV) as well as on respiratory characteristics and influences. [267, 353, 235, 69]

Especially HRV (*Tachogram*) analysis is a widely used instrument for non-invasive evaluation of the autonomous cardiovascular control [215, 311]. Conventional methods for HRV analysis can be divided into time domain, frequency domain, and non-linear analysis. Whereas the time domain analysis includes statistical methods for the measurement of variability of normal beat-to-beat intervals, the frequency-based methods investigate the distribution of absolute and relative power density within predefined frequency bands [215]. For short time analyses (5 min), a range of three main frequency components are distinguished: Very Low Frequency (VLF) in the range 0 Hz to 0.04 Hz Low-Frequency (LF) in the range 0.04 Hz to 0.15 Hz and High-Frequency (HF) in the range 0.15 Hz to 0.4 Hz [215].

In the following, characteristic observations in cardiorespiratory parameters, depending on sleep stages, are described.
From a cardiorespiratory perspective, NREM sleep can be regarded as period of autonomous stability. The transition from wake via stages N1 and N2 to deep sleep stage N3 is characterized by a progressive increase in parasympathetic regulation and sympathetic inhibition, whereby respective maxima and minima are reached in deep sleep. This shift of the sympathovagal balance can be determined by the progressive decrease of the mean heart rate, as well as the power increase in the HF band and the power decrease in the LF band [267, 353, 235]. The

breathing rate was observed to increase progressively during the transition from wake to deep sleep stage N3 [141, 184].
REM sleep is characterized by instability of the cardiovascular and respiratory system and by immediate outbreaks of varying sympathetic activity. The transition from NREM to REM state is related to significant increase in mean heart rate and the occurrence of irregular breathing patterns. Corresponding to the predominance of the sympathetic nervous system and vagal withdrawal, a power increase in the LF band and a power decrease in the HF band can be observed [267, 353, 69]. Also, an overall increased breathing rate in REM sleep compared to NREM sleep was observed [184].

Table 7.2 lists related works from the scientific literature focusing on sleep stage classification using cardiorespiratory signals. Here, works using HR time series, HRV features as well as respiratory signals from various sources are included. The stated results rather provide a basic performance estimate than facilitate direct comparison as different data sets are used for validation.

Table 7.2: Comparison of related works on sleep stage classification using cardiorespiratory signals.

Ref.	Year	Classifier	Features	Results
Wake, sleep				
[166]	2009	MLP	HRV features, respiration rate features	$Acc. \approx 0.85$
[63]	2015	HMM	HRV features	$Acc. \approx 0.80$
[214]	2018	CNN	HR time series	$\kappa = 0.24 - 0.54$
Wake, N1/N2/N3, REM				
[282]	2007	LDA	HRV features, respiration rate features	$\kappa = 0.45$
[389]	2016	Threshold	Respiration rate features	$\kappa = 0.49$
Wake, N1/N2, N3, REM				
[97]	2013	LDA	HRV features	$Acc. \approx 0.75$
[107]	2015	LDA	HRV features, respiration rate features	$\kappa = 0.49$
[350]	2017	GBM	Respiration rate features	$\kappa = 0.56$
[200]	2018	CNN, SVM	HRV features, respiration rate features	$\kappa = 0.54$
[278]	2019	LSTM	HRV features	$\kappa = 0.61$
[5]	2019	CNN-LSTM-CRF	Respiration rate time series	$\kappa = 0.57$
Wake, N1, N2, N3, REM				
[335]	2020	CNN-LSTM	HR time series, respiration rate time series	$\kappa = 0.59$

7.2.3 Data

All following experiments are conducted on an excerpt of the Sleep Heart Health Study (SHHS)[1], a multi-center cohort study to determine cardiovascular and other consequences of sleep-disordered breathing. The initial examination (SHHS-1) includes the polysomnograms and hypnograms of 6.441 subjects aged $\geq$ 40 years. In this work, only SHHS-1 data from subjects without acute cardiovascular diseases are included (219 subjects) and patients with sleep-disordered breathing (Apnoe-Hypopnoe-Index (AHI) $<$ 5) (182) excluded. Both can be expected to render sleep staging more difficult due to altered sleep architectures. Moreover, 26 subjects are excluded due to bad ECG signal quality, overall resulting in 237 subjects. For each 30 s epoch reference annotations are provided [70], whereas stages N1 and N2 are combined into a single light sleep phase. The resulting class distribution over all subjects and epochs are shown in figure 7.2 (b). Using a subject-wise approach, the data is randomly divided (80 %/20 %) into a train an test set, comprising 190 and 47 subjects, respectively.

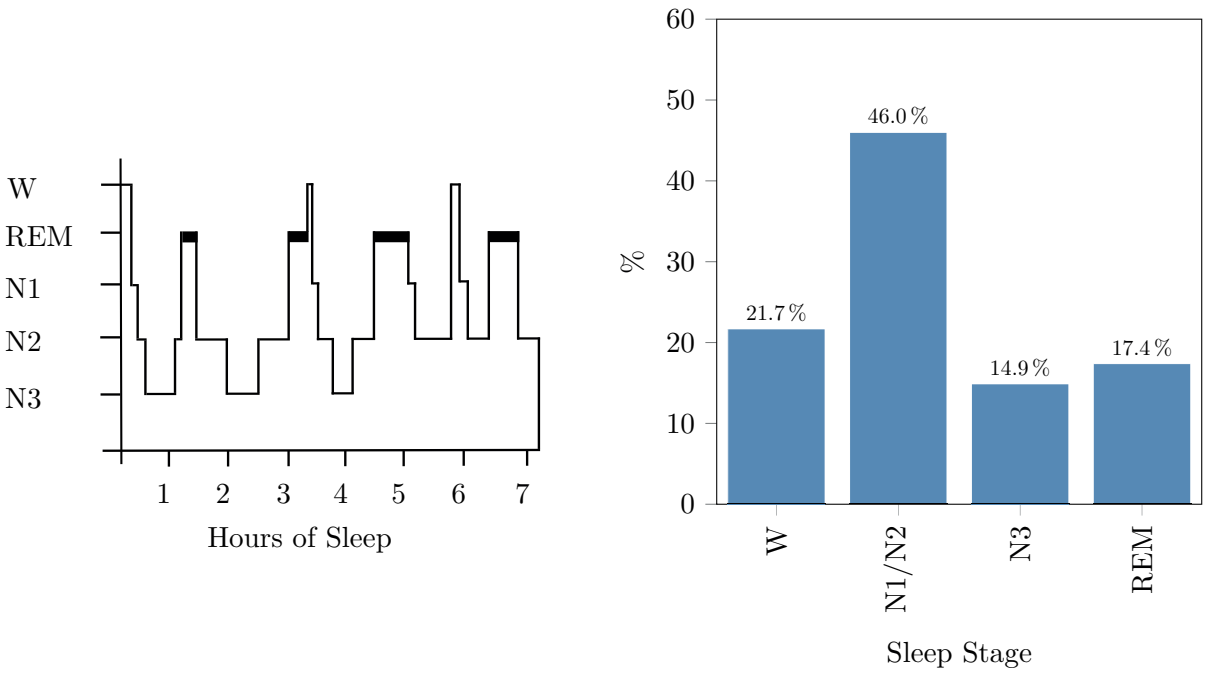


Figure 7.2: Exemplary hypnogram (a) and sleep stage distribution over subjects and epochs (b).

7.2.4 Approaches and Results

Initially, the series of RR intervals, i.e. the *tachogram*, is extracted from the PSG ECG channel and the series of breath-to-breath intervals from the thoracic excursion signal. The approaches for sleep stage classification studied within this work can be divided into methods (i) using features extracted from the RR intervals and breath-to-breath time series and such methods (ii)

[1] `https://sleepdata.org/datasets/shhs`

using the time series directly as input. As sleep cycles feature inherent temporal dependencies, all chosen algorithmic approaches facilitate to model processes or sequences of outputs. All experiments classify input signals into one of the four sleep stages wake, N1/N2, N3, or REM.

7.2.4.1 Feature-based Sleep Stage Classification

For the extraction of HRV and respiratory features, a 180 s sliding window with 30 s step size is applied to the digitized and pre-filtered signals. As listed in table 7.3, a vector of 10 conventional HRV features (1 time domain, 7 frequency domain, 2 non-linear) and two respiratory features are extracted and assigned to each epoch.

Table 7.3: HRV and respiratory features extracted from RR interval and breath-to-breath time series, respectively [215].

	Feature	**Unit**	**Description**
HRV	VLF*	ms^2	Power in the VLF range (<0.04 Hz)
	LF	ms^2	Power in the LF range (0.04 Hz to 0.15 Hz)
	HF	ms^2	Power in the HF range (0.15 Hz to 0.4 Hz)
	Total power	ms^2	Variance of all RR intervals
	LF/HF	–	Ratio LF/HF
	LF norm*	–	Normalized power in the LF frequency range
	HF norm*	–	Normalized power in the HF frequency range
	SD1	ms	Standard deviation 1 derived from Poincaré plot
	SD1	ms	Standard deviation 2 derived from Poincaré plot
	RRI*	ms	Normalized mean RR intervals
Respiration	BBI*	ms	Normalized mean breath-to-breath intervals
	covBBI*	ms	Normalized standard deviation of breath-to-breath intervals

HMMs, as introduced in section 3.3.4, attempt to model a process where a sequence of emitted symbols is observed and an intrinsic underlying pattern of states exists. The aim of classification is to find the most likely hidden state sequence given a HMM and the observed symbol sequence. Transferred to the sleep stage classification problem, each hidden state corresponds to a sleep stage and the observed symbols to the associated feature vector. The HMM is characterized by the compact notation $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$, with the transition probabilities $\mathbf{A}$ between each sleep stages, the emission probabilities $\mathbf{B}$ for each sleep stage, represented by continuous PDF, and the initial state distribution $\mathbf{\Pi}$. The PDFs in $\mathbf{B}$ are characterized by the mean vector μ and covariance matrix $\mathbf{\Sigma}$ of features observed for each sleep stage. Consequently, the HMM λ can be estimated from the training data. Utilizing the observed sequence of symbols, namely the sequence of feature vectors, the classification of sleep stages is implemented as the *decoding* problem described in section 3.3.4 and the *Viterbi* algorithm is used to find the most likely sleep stage sequence. Two implementation are contrasted: using all 12 features listed in table 7.3 and using only a subset of six selected features ($*$) which are determined by preliminary experiments.

As can be seen in table 7.4, the reduced feature set is not inferior to the larger feature space.

Furthermore, the application of a Random Forest (RF) classifier introduced in section F.1.5 and the combination of RF and HMM is studied. Therefore, for an observed feature vector, the classification probability of the DT ensemble is utilized as emission probability. Whereas the RF classifier shows comparable results, the combination of RF and HMM can improve the results.

A further approach to integrate temporal information into the classification is to compute new features from the given series of HRV and respiratory features. To do so, a CNN architecture as described in section 3.3.5.2 is implemented which learns individual filter kernels for each input feature sequence comprising historical observations. In order to maintain interpretability and allow for feature analyses, only a single convolutional layer, no additional pooling layer, and a filter size equal to the considered signal sequence is implemented. As actual classifier, a MLP with two hidden layers is utilized. The network hyperparamters are optimized by means of a 3-fold cross-validation. Best results, which outperform all previous methods, could be achieved with 13 filter kernels for each feature time series and a considered signal length of 13 time steps (6.5 min).

As already stated in section 3.3.5.3, the cyclic structure of RNNs allow to utilize historical information to influence internal states and outputs rather than just taking advantage of information available at a given time step. This makes RNN architectures, namely LSTM networks, particularly suitable for processing sequential data such as sleep stages. Similarly to the CNN approach, for each feature an input sequence of fixed length is fed into the network. As the network can be assumed to learn feature importance, all 12 available features are included. For the sake of simplicity, the applied LSTM network only consists of one LSTM layer and one fully-connected output layer (*Softmax*). For each sequence time step, up to the epoch to be classified, outputs are fed back into the recurrent units and influence the internal states. Besides an uni-directional LSTM network, also a bi-directional LSTM network is implemented and optimized which not only incorporates information from the past but from both past and future. Best results could be yielded with signal length of 120 time steps (60 min) and 15 and 50 LSTM units, respectively. According to the results in table 7.4, the classifiers obviously doesn't benefit from the bi-directional approach.

7.2.4.2 Raw Time Series-based Sleep Stage Classification

The RR interval and breath-to-breath time series are given as sequences with non-equidistant discrete values, since not only the values themselves but also the distances between values depend on the RR and breath-to-breath intervals, respectively. In order to derive continuously sampled sequences as valid classifier input representations, the available time series are linearly interpolated and re-sampled with 4 Hz. According to the reference annotations, the sequences are segmented into 30 s epochs. These 30 s epochs to be classified are further extended by appending the preceding and succeeding four epochs, i.e. 120 s segments, resulting in 270 s sequences as classifier input. Preliminary studies have proven that more input information added to the epoch

[1] Method evaluated with differing test data.

Table 7.4: Feature-based sleep stage classification results.

Method	**Description**	κ
HMM-6	HMM using six dimensional feature vector as emitted symbol.	0.31
HMM-12	HMM using 12 dimensional feature vector as emitted symbol.	0.30
RF-6	RF using six dimensional feature vector as input.	0.29[1]
HMM+RF-6	RF using six dimensional feature vector to predict class probability which is used instead of the probability density functions.	0.34[1]
CNN-6	CNN which learns individual filter kernels for each of the six input feature sequences. The extracted features are classified by a MLP.	0.53
LSTM-12	Uni-directional LSTM using 12 feature sequences as input.	0.49
bi-LSTM-12	Bi-directional LSTM using 12 feature sequences as input.	0.46

Table 7.5: Raw time series-based sleep stage classification results.

Method	**Description**	κ
CNN-RR	CNN architecture using RR interval time series as input.	0.45
CNN-BB	CNN architecture using breath-to-breath interval time series as input.	0.44
MCCNN	Multichannel CNN architecture, which concatenates the CNN-RR and CNN-BB outputs to form MLP classifier input.	0.56
MCCNN-LSTM	Multichannel CNN architecture extended by a bi-directional LSTM layer, using the MLP outputs as input.	0.61

itself is essential for successful classification [335].

Four approaches are investigated. CNN architectures using RR interval time series (1) and breath-to-breath time series (2) individually are implemented and optimized based on [214]. The networks are build of five convolution blocks for hierarchical feature extraction, each consisting of two 1-dimensional convolutional layers, and a MLP with two hidden layers and *softmax* output layer for classification. Furthermore, a multichannel CNN (3) is implemented which concatenates the feature maps generated by the two individual CNNs described above. The concatenated vector is fed into a MLP for classification. Finally, the multichannel CNN is extended by an additional bi-directional LSTM network (4) as proposed in [335]. With the objective to reveal more global patterns, the LSTM layer uses the series of MLP outputs as input. The outputs of the bi-directional LSTM network is finally classified into four sleep stages by a *softmax* layer. The results summarized in table 7.5 indicate that CNNs are powerful tools for classification of sleep stages based on raw RR interval or breath-to-breath interval time series. However, especially in combination with LSTMs to exploit global sequential information, all other examined approaches are outperformed regarding achieved *Cohens's Kappa* scores.

7.2.5 Conclusions

Considering the results from table 7.4, none of the investigated HMMs are capable of achieving classification results comparable with neural network approaches (CNNs, LSTMs). One major

difference is the limitation of first-order HMMs, that the current state depends on the previous state only. Earlier states are not considered. Moreover, the time invariance of HMMs assume constant state, i.e. sleep stage, transition probabilities, which is not reflecting the characteristic of sleep. The proposed neuronal networks, in contrast, exploit much more information from previous or even future time steps and hence can even overcome the time invariance limitation. Especially the CNN approach to learn patterns in sequences of feature time series proves to be superior in this respect.

As already stated in section 7.2.4.2, sleep stage classification using raw RR interval and breath-to-breath interval time series is in no way inferior to the best performing feature-based approach. The CNN is obviously capable of extracting at least as meaningful features from the time series as the conventional HRV and respiratory features. Integrating information about the global temporal sequence of sleep stages by means of a downstream LSTM layers further improves the classification quality significantly and produces state-of-the-art results.

Future works must focus on training and testing classification algorithms using more extensive benchmark data sets including pathological hypnograms and input data. In this context, also the development of classifiers specialized for a subgroup of subjects and patients is a possible approach. Here, future research must deal with the identification of relevant characteristics, such as demographic properties or health conditions, in order to form meaningful groups. Finally, end-to-end classification, utilizing raw ECG and respiration signals as classifier inputs can be focus of further research. According to own preliminary studies using raw ECGs only ($\kappa = 0.16$), there is still potential.

7.3 Parkinson's Disease Patient Gait Assessment

7.3.1 Introduction

The neurodegenerative PD (idiopathic Parkinson syndrome) is the second common neurological disease after Alzheimer's [195]. PD is caused by the progressive necrosis of dopaminergic cells in the brain and is characterized in particular by motor symptoms. The four cardinal symptoms are akinesis or bradykinesis (inhibited or slowed down voluntary motor function), rigor (muscle stiffness), tremor (shaking), and postural instability [122, 89]. However, also non-motoric symptoms, such as cognitive (Dementia) or psychological (Depression) conditions, are associated with PD [122, 89]. Since the cause for the dopaminergic cell necrosis is still unknown, only symptoms are treated. Due to the progressive characteristic of the disease, decrease in effectiveness, and ADEs, typical pharmaceutical treatments must be subject to constant monitoring [111, 283]. Consequently, continuous and objective assessment of severity and treatment response are essential for successful treatment and management. A common tool to assess the severity of PD is the Unified Parkinson's Disease Rating Scale (UPDRS) [362], classifying patients on the scale 0 (no impairment) and 199 (maximal impairment). The UPDRS consists of four parts which are summed up, whereas the part covering motoric examinations (UPDRSM) is of special interest of this research.

The four stated cardinal symptoms affect patient's motion in general and gait in particular. Consequently, gait analysis can be a central component to assess severity and progression of the disease. In the following experiments, sensor recordings of vertical ground reaction forces are utilized to assess the gait of PD patients and healthy control subjects: The objective is to differentiate between patients and healthy control subjects.

7.3.2 Background and Related Work

A literature review revealed five papers which use the same dataset and record selection as this work. From those, three works apply feature extraction based on domain knowledge and two works apply raw data for modeling. Table 7.6 summarizes the identified publications.

Table 7.6: Comparison of related works on Parkinson's gait classification using the same dataset and records as this work.

Ref.	Year	Classifier	Features	Results
[76]	2013	SVM	Frequency domain features	*Acc.* = 91.20 %
[393]	2016	RBF Network	Modeling of gait dynamics	*Acc.* = 96.39 %
[1]	2017	RF	Time domain features and frequency domain features.	*Acc.* = 98.04 %
[156]	2019	CNN	Raw sensor signals.	*Acc.* = 88.70 %
[99]	2020	CNN	Raw sensor signals.	*Acc.* = 98.70 %

7.3.3 Data

In the following studies, the publicly available *Gait in Parkinson's Disease*[1] benchmark dataset is used, which contains gait recordings from 93 PD patients and 73 healthy controls. Underneath each foot eight sensors are placed to measure vertical ground reaction forces, sampled with 100 Hz, as a function of time while walking. Besides each sensor record, the sum of each food is given, resulting in 18 signals for each record. Besides reference walks, were the subjects walked at their usual, self-selected pace for approximately two minutes, the impact of additional cognitive tasks, walking aids, and stimulations was studied in the provided data. In this work, however, only the reference walks are included and combined into an overall dataset. Four patients are excluded as relevant labels or data is not available, resulting in data of 162 patients.

For performance evaluation, subjects are divided into a train (129) and test (33) partition. The distribution of severity of the disease (UPRDRSM) is taken into account in the partitioning.

7.3.4 Approaches and Results

Comparable to the sleep stage classification experiments describe in section 7.2, the studies on gait analysis done in this work can be divided into methods (i) using features extracted from the sensor signals (ii) and such methods using the sensor signals directly as input.

[1] `https://physionet.org/content/gaitpdb/1.0.0/`

Table 7.7: Feature-based gait classification results.

Method	Description	Acc.
RF-673	RF using all available features.	82.1 %
RF-18	RF using best feature subset according to feature selection.	82.1 %

7.3.4.1 Feature-based Classification and Regression

Based on an extensive literature review, gait describing features were identified and extracted from each channel after signal preprocessing (low-pass filtering, denoising, elimination of outliers, normalizing with subjects' weights). The overall 673 extracted features are grouped into the three categories: time domain features (264), which can be further divided into kinetic (220) and spatio-temporal (44) features, frequency domain features (300), and symbolic dynamics (109). For the kinetic features, statistical characteristics are calculated from the preprocessed signals, the center of pressure, and the heel strike and toe off values. Spatio-temporal features are derived from statistics on gait cycle characteristics and asymmetries. Frequency domain features are derived by applying a discrete wavelet transform and computing statistics on the resulting coefficients. Finally, signals are transformed to symbolic sequences, split into word sequences, and statistics on the probability distribution of word types are computed, yielding symbolic dynamics features.

Due to its robustness and state-of-the-art classification results, a RF is employed as classification algorithm. As the importance of the various features assumably varies and features are subject to redundancies, a feature selection algorithm is employed to reduce the feature space dimension.

As benchmark results, all 673 features are applied as classifier input. Furthermore, a wrapper method, namely a Sequential Forward Selection (SFS), is utilized in order to select the most important features. As summarized in table 7.7, the feature selection strategy is not capable of improving classification accuracy compared to using all available features. The RF inherently selects the most important features. However, results prove that same performance could be yielded by providing just a small fraction (2.67 %) of the available features. From 18 selected features, nine stem from time domain, whereas only one is spatio-temporal, five are from time domain, and four are symbolic dynamics features.

7.3.4.2 Raw Time Series-based Classification and Regression

Comparable to section 7.3.4.1, the vertical ground reaction force signals of each channel are preprocessed by normalization with the subjects' weight. Moreover, the signals are divided into segments of defined length with overlap 50 % which are to be classified. Subjects are assign to a class based on majority voting. In this work, segment length of 1 s and 2 s are studied, i.e. 100 and 200 samples. Additionally, each variant is studied with sampling frequency reduced to 50 Hz, i.e. 50 and 100 samples.

A CNN architecture based on [99] is implemented and optimized for the given task. The

Table 7.8: Raw time series-based gait classification results.

Method	Description	Acc.
CNN-1-50	CNN architecture using 1 s segments as input samples with 50 Hz.	97.0 %
CNN-1-100	CNN architecture using 1 s segments as input samples with 100 Hz.	93.9 %
CNN-2-50	CNN architecture using 2 s segments as input samples with 50 Hz.	97.0 %
CNN-2-100	CNN architecture using 2 s segments as input samples with 100 Hz.	93.9 %

network is build of individual 1-dimensional CNNs for each of the 18 channels. Each of those parallel CNNs consist of four convolutional layers and a fully connected layer, whereas two convolutional layers are always followed by a max-pooling layer. The kernel length is always three, whereas the first layer comprises 16 kernels and all the followings 32. The classifier hyperparameters are determined with a grid search by means of a 5-fold cross validation on the training partition. The concatenated CNN outputs are fed into a MLP with two hidden layers for classification. According to the test subject results of the four segmentation variants summarized in table 7.8, the 50 Hz sampling implies to be advantageous. Considering the segment accuracies, however, especially the 2 s variant sampled with 50 Hz is superior to all other approaches.

7.3.5 Conclusions

A large number of gait describing features are described in the literature from which a vast majority are implemented. According to the results from table 7.7, good classification results could be yielded. The applied feature selection is capable of reducing the feature space dimensionality to a large extend while maintaining classification accuracy. Nevertheless, the CNN, which learns features from each of the input signals, is clearly superior independent of he applied segmentation strategy. Longer time periods of 2 s, which contain an entire gait cycle, with simultaneously reduced sampling rate, are advantageous. Suchlike, state-of-the-art classification results can be yielded. According to the results from the literature listed in table 7.6, the feature-based approach is not generally inferior to CNNs. Hence, it can be assumed that emphasis on features and feature selection strategies can further improve the feature-based results.

The demonstrated results distinguish the subjects into PD patients and healthy control subjects. In order to provide a more precise evaluation of treatment effectiveness, future works will focus on reliably assessing the patients' severity of the disease in terms of a clinical score such as the UPDRSM. Based on the described sensor records, regression techniques can be utilized to predict such scores. Own preliminary experiments, which use the features described in section 7.3.4.1 as independent variables of a RF regression algorithm, yields a RMSE of 6.75. The variance in the UPRSM that is declared by the independent variables (coefficient of determination $R^2 = 0.17$) indicates that those predictions require further investigations.

7.4 Drug Review Sentiment Analysis

7.4.1 Introduction

As already introduced in chapter 1, discrepancies in patient cohorts and treatment conditions can have significant impact on the effectiveness and potential risks of ADEs such as side effects. Therefore, post-marketing drug surveillance, i.e. pharmacovigilance, plays a major role concerning drug safety once a drug has been released. Online platforms containing patient experience with pharmaceutical drugs can be regarded a valuable source of information for pharmacovigilance. Additionally, patient-initiated observational studies based on such platforms can facilitate a novel mean to assess the effectiveness of treatment options [380].

However, requirement for automatic processing and analysis of the information contained in large amounts of unstructured information is the transformation of inherent aspects into numerical ratings. One typical way of doing so, in the context of product ratings, is sentiment analysis, which is an extensively studied domain in processing free-text in web media analyses [204]. Sentiment analysis of patient data in general and on drug experience in particular is a challenging research problem that is currently receiving considerable attention. One of the main issues is the lack of annotated data, which is crucial for accurate sentiment classification. Especially, labeled data dealing with distinct aspects, is rare. Moreover, the availability of labeled data is highly domain dependent. Patients suffering from certain conditions are more active in reporting experience on their treatment than others.

In this work (1) the possibility to apply sentiment analysis on drug reviews, and the identification of effectiveness of a drug as well as the severity of side effects caused by a drug using patient reviews is studied. Therefore, classification of side effects and effectiveness is treated as an aspect-based sentiment analysis problem. Furthermore, to address challenges related to the limited data availability, (2) the transferability of the trained models among domains, i.e. diseases, as well as (3) across data sources is studied.

7.4.2 Background and Related Work

Many approaches to sentiment analysis are based on sentiment lexicons. These approaches recognize sentiment terms and patterns of sentiment expressions in natural language texts by matching textual units with opinion words in lexicons annotated for sentiment polarity. However, studies showed that sentiment analysis is often domain-dependent since the polarity of single terms can differ depending on the context they are used in [124, 86]. Furthermore, the language in online forums is highly informal and user-expressed medical concepts are often nontechnical, descriptive, and challenging to extract. Which is why typical lexicons are of limited use for drug review analyses. An alternative approach treats the task as classification problem. Here, machine learning is used to train classifiers on domain-specific data sets to detect the polarity at sentence or document level. Such approaches have the additional advantage to be capable of performing medical sentiment analysis over multiple facets, i.e. sentiments can be learned on specific aspects such as side effects and effectiveness.

Also related works on drug review sentiment analysis can basically be divided into approaches applying lexicons with sentiment scores [237, 225, 297] or such approaches learning sentiments employing supervised classification [128]. Moreover, several studies have attempted to improve domain adaption or cross-domain sentiment classification, although not on drug review aspect-level but among various entities as products, movies or restaurants. In [232] a comprehensive systematic literature review on cross-domain sentiment analysis is presented.

7.4.3 Dataset

Data from two independent webpages for retrieval of user reviews and ratings on drug experience is used. Drugs.com is, according to the provider, the largest and most widely visited pharmaceutical information website providing information for both, consumers and healthcare professionals. It provides user reviews on specific drugs along with related condition and a 10 star user rating reflecting overall user satisfaction. Similarly, Druglib.com is a resource on drug information for both, consumer and healthcare professionals. It comprises considerably fewer reviews but reviews and ratings are provided in a more structured way. Reviews are grouped into reports on the three aspects *benefits*, *side effects* and overall *comment*. Additionally, ratings are available concerning overall satisfaction analogously to drugs.com as well a 5 step *side effects* rating, ranging from *no side effects* to *extremely severe side effects* and a 5 step effectiveness rating ranging from *ineffective* to *very effective*.

User comments and ratings are gathered from both pages using an automatic web crawler, resulting in two data sets comprising 215,063 reviews from Drugs.com and 3,551 reviews from Druglib.com, respectively. Furthermore, three level polarity labels for overall patient satisfaction and three level effectiveness and side effect scores using thresholds as specified in table E.3 are derived. Both data sets are further split into training and test partitions according to a stratified random sampling scheme with the proportion of 75 % and 25 %, respectively. As shown in table E.3, the total number of individual drugs in the Drugs.com data amounts to 6,345 in comparison to the 541 drugs contained in the data derived from Druglib.com. However, the average number of reviews per drug is still considerably higher in the Drugs.com data (58.86) than in the Druglib.com data (7.66). The amount of unique conditions contained in the Druglib.com data, on the other hand, seems to exceed the number of the Drugs.com data. However, it is to be noted that conditions in the latter platform are user inputs in contrast to Drugs.com where conditions are selected from a defined list. Therefore, in this case conditions comprise variations in spelling, synonyms and combination of conditions.

7.4.4 Approaches

The objective of this study is threefold:

1. Prediction of the overall patients' satisfaction with applied medications and sentiments on side effects and effectiveness by employing classification-based sentiment analyses.

2. Evaluating the transferability of models among medical domains, i.e conditions, by learning

a model on data from one condition (source domain) to classify overall patient satisfaction in data from another condition (target domain).

3. Evaluating the transferability of models across data sources, i.e Drugs.com and Druglib.com, by learning a model on reviews from one data source (source data) to classify overall patient satisfaction and sentiments on side effects in data from another source (target data).

Whereas for the first two tasks the ground truth is available for both data sets, distinct reviews covering the aspects side effects and effectiveness along with labels are only available for the Druglib.com data. To evaluate the transferability of side effects prediction models across data sets, 400 randomly picked samples from the Drugs.com data were manually labeled concerning side effects by two independent annotators. The inter-rater agreement measured with the *Cohens's Kappa* statistic is 81.84 %which is considered as very strong agreement. The annotators discussed all mismatching entities and agreed on a consensus.

Both approaches, sentiment analysis regarding overall patients' satisfaction and the aspect-based analysis of patients' sentiments on side effects and medication effectiveness were converted to classification problems. In case of overall patient satisfaction, the user ratings are converted to three disjoint classes representing the polarity of a patient's sentiment regarding the applied medication (negative, neutral, positive). In addition, also the severity of side effects and the level of effectiveness were transferred to three disjoint classes as described in table table E.3.
For all prediction tasks, a n-grams approach is applied in order to represent the user reviews. That means both, single tokens, e.g. words, (unigrams) as well as two or more adjacent tokens (bigrams, trigrams), e.g. 2- or 3-word expressions, are used to derive features for classification. Based on the total collection of occurring n-grams, i.e. the corpus, each review can be represented as a sparse vector of token counts.
Initially, all reviews are preprocessed according to a standard scheme: Alphabetic characters are transferred to lowercase and special characters, punctuation and numbers are removed. Subsequently, the preprocessed documents are tokenized on spaces to obtain the overall vocabulary and a feature space representations of each review. No stop words are removed from the texts. However, to reduce the feature space, terms that have a relative document frequency higher than a given threshold are discarded when building the vocabulary.
Using the extracted feature representations, LogR is employed for building sentiment models for the various prediction tasks. Model hyperparameters are tuned using a 5-fold cross-validation grid search on the respective training data, targeting the best *Cohens's Kappa* score. Optimized hyperparameter include n-gram number of adjacent tokens, token document frequency threshold, and LogR regularization strength. As shown in table E.2, besides the annotated subset from the Drugs.com data, labels are considerably unbalanced. To compensate for this disproportionate distribution, classifications errors are penalized with a weight inversely proportional to its class frequency during training.
All experiments are evaluated by computing confusion matrices and deriving both, accuracy and *Cohens's Kappa* scores.

Table 7.9: In-domain sentiment analysis.

Aspect	**Source**	**Acc.**	κ
Overall rating	Drugs.com	92.24	83.99
Overall rating	Druglib.com	68.73	27.60
Overall rating (all)	Druglib.com	75.39	43.62
Benefits (Effectiveness)	Druglib.com	77.70	44.13
Side effects	Druglib.com	77.12	60.22

7.4.5 Experiments and Results

7.4.5.1 In-domain Sentiment Analysis

In an initial experiment, overall performance when applying sentiment analysis to drug reviews is studied. Therefore, one model for each data set (Drugs.com and Druglib.com), to classify overall patient satisfaction reviews, is trained and evaluated utilizing the corresponding training and test data. Additionally, as in case of the Druglib.com data the *comments* section might only contain supplementary remarks, a combination of all three reports (*benefits*, *side effects* and *comments*) of a patient on a respective drug were concatenated to represent the overall patient satisfaction review.

Furthermore, the expression of sentiments on the two aspects *side effects* and *effectiveness* within patient generated texts are studied. Therefore, two LogR models are optimized and trained on the *benefits* and *side effects* training data derived from Druglib.com, respectively. Both, predicted *effectiveness* and *side effect* labels are compared against the actual labels obtained from the user ratings.

As detailed in table 7.9, overall patient satisfaction can be mined from patient texts with very high accuracy and *Cohens's Kappa* score in case of the Drugs.com data. The significantly worse performance reported for the Druglib.com data is assumed to have two main reasons. First, the data set is considerably smaller, which hampers the modelling. Moreover, the *comments* section is mainly used for supplementary information on personal experience and drug application and not explicitly for comments on satisfaction. When combining all three aspects, i.e. patient reports, classification performance can be improved over the previous result. In both approaches concerning the Druglib.com data the largest error contribution results from neutral ratings classified as positive which cannot be improved by data combination. The performance improvement, however, results from the reduction of misclassified negative ratings.

Sentiment analysis related to the specific aspects *effectiveness* and *side effects* shows promising results. Especially the *side effects* comments seem to provide valuable features that facilitate mining sentiments on side effects. Here, errors are mainly due to misclassification of neighbouring classes, namely excessive missclassification as *mild / moderate side effects*. In case of *effectiveness* classification the larges error contribution stems from *marginally / moderately effective* reviews classified as *considerably / highly effective*, whereas *considerably / highly effective* labeled reviews can be classified correctly with 95 % accuracy. However, it must be kept in mind that also

comments on *benefits* not necessarily relate to effectiveness only but may also encompass other aspects.

7.4.5.2 Cross-domain Sentiment Analysis

In this experiment the performance of models built on data from one condition, i.e. the source domain, and evaluated on data related to other conditions, i.e. the target domain is studied. To do so, overall patient satisfaction models are trained on drug review subsets related to one selected condition only. These domain models are then evaluated on other condition related subsets. Domains, i.e. subsets of particular conditions, are selected by extracting five of the most frequent disorders present in the Druglib.com data set from diverse medical fields. These are Contraception (38,436), Depression (12,164), Pain (8,245), Anxiety (7,812) and Diabetes, Type 2 (3,362), with frequency in descending order. In-domain performances, i.e. training and testing of data from the same condition, are reported as averaged 5-fold cross-validation results.

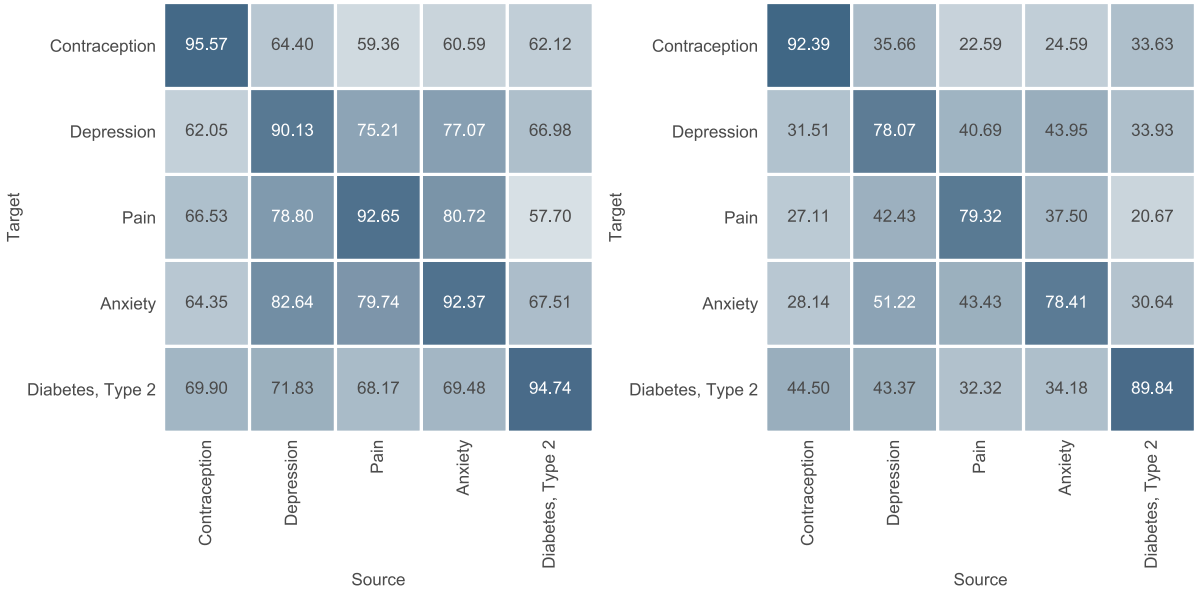


Figure 7.3: Cross-domain sentiment analysis results: (a) Accuracy and (b) *Cohen's Kappa* of sentiment predictions.

The results summarized in figures 7.3 (a) and (b) demonstrate that the selected training domain has considerable impact on the classifier performance when applied to data from other domains. Especially, in-domain training and testing clearly outperforms all cross-domain setups. This finding clearly emphasizes the hypothesis of domain-specific vocabulary. For Contraception and Diabetes, even the overall rating classification using the entire data could be outperformed. However, the model trained on Depression data only seems to generalize better on the other domain data than e.g. a model trained on Diabetes data only. Furthermore, there are combinations showing better performances than others, e.g. Depression and Anxiety compared to Contraception and Anxiety, which is assumed to be due to underlying coherences of side effects or expressions and domain specific vocabulary used by patients. Moreover, the medical field

dealing with Depression and Anxiety is closely related. From drugs concerning Depression (115) and Anxiety (81), 33 drugs are applied in both conditions whereas for Contraception (181) and Anxiety there is no overlap. Furthermore, the confusion matrices showed that main classification errors occurred on neutrally labeled reviews for all domain combinations. Transferring the task to a binary classification problem without classification of neutral entities would result in substantially higher accuracy and *Cohens's Kappa* values.

7.4.6 Cross-data Sentiment Analysis

Finally, the transferability of the trained models among data sources is studied. Overall patient satisfaction models are trained on both associated training data sets and evaluated on drug reviews from the other, independent data source test set. As discussed in section 7.4.5.1, in case of the Druglib.com data a combination of all three reports (*benefits*, *side effects* and *comments*) were concatenated to represent the overall patient satisfaction review. Additionally, the performance of a classifier trained on *side effect* comments from the Druglib.com data is evaluated on the manually annotated data from Drugs.com.

Table 7.10: Cross-data sentiment analysis.

Aspect	**Train Source**	**Test Source**	**Acc.**	κ
Overall Rating	Drugs.com	Druglib.com	75.29 %	0.48
Overall Rating	Druglib.com	Drugs.com	70.06 %	0.27
Side Effects	Druglib.com	Drugs.com	49.75 %	0.26

Transferring a sentiment model trained on the significantly larger Drugs.com data to the Druglib.com data shows promising classification capabilities. Evaluating the model trained on the much smaller Druglib.com data with the Drugs.com data, however, doesn't perform satisfactorily. We assume such findings, on the one hand, to result from the limited training data size. On the other hand, differing data properties are likely to restrict the transferability. As stated previously, in contrast to the Druglib.com data Drugs.com reviews are highly unstructured covering multiple aspects in an entire review.

As summarized in table 7.10, applying the model trained on the *side effect* aspect to the Drugs.com reviews also performs poorly. The largest fraction of the classification error stems from reviews labeled as reporting *No* or *severe / extremely severe side effects* as *mild / moderate*. The features extracted from the Druglib.com data obviously don't contain sufficient discriminating power to classify the unstructured Drugs.com review which are not dealing with a single aspects only. Utilizing a larger training data set, leading to less ambiguous features, might improve the results.

7.4.7 Conclusions

Within this preliminary work, the application of machine learning based sentiment analysis of patient generated drug reviews was studied. Depending on aspect and data source promising

classification results could be obtained. Concerning model portability, in-domain (i.e. condition) training and evaluation shows very good classification results, the performance of models trained on one specific condition and tested on another condition, varies among domains. However, conditions which belong to similar medical fields and are partly treated with equal medications also show higher potentials for model transferability. Cross-data evaluation, i.e. training and testing classifiers on data from different sources, is only unsatisfactorily possible with the applied classifier and features. The application of a more sophisticated classifier, namely a LSTMs network, was additionally investigated. Instead of representing each review in a single vector, the sequence of words (*one-hot-encoding*) is exploited. Moreover, word embeddings are applied to yield denser word representations and a reduced feature space dimension: (1) A readily trained *GloVe* embedding [265], (2) a *GloVe* embedding adapted to the given corpus, and (3) a *Word2Vec* embedding [223] trained on the given data. However, neither the LSTMs network nor the word embeddings were capable of outperforming the basic and n-gramm an LogR approach. The results clearly indicate that especially aspect-based sentiment analysis requires more extensive data sets to extract features with sufficient generalization capabilities.

8 Conclusion

In section 8.1, the results presented in chapter 6 are discussed, related to the hypotheses formulated in chapter 1, and general findings are named. The following section 8.2 highlights problems and challenges and gives some suggestions and ideas for future research directions. Section 8.3, finally, summarize this thesis.

8.1 Discussion and Generalization

Considering the generalization performance demonstrated in chapter 6, it can be concluded that the first two research questions formulated in section 1.2 can be answered positively for all proposed algorithms and hence the two associated hypotheses are not rejected. According to the demonstrated outcome prediction performance, it is possible to predict outcome of therapy options more accurately than average outcome, which is represented by the *average affinity* baseline. The therapy recommendations derived from these predictions clearly outperform the *overall popularity* baseline which answers the second research question. The third hypothesis, however, is rejected. MAP@3 scores and *Cohen's Kappa* results achieved by all of the proposed algorithms are clearly below the ones resulting from the expert recommendations.

Nevertheless, it can be concluded that the proposed approaches are capable of supplementing a physician's experience and external evidence with practice-based evidence from local cohorts, as proposed in [209], and can provide evidence where it is missing otherwise. The proposed therapy recommendation approach is capable of automatically providing actionable recommendations at the time and location of decision-making, as stated as essential features of CDSSs according to [168].

A general advantage of deriving recommendations from outcome predictions is independence from the popularity of a drug. The treatment options that are potentially most successful with respect to an addressed outcome objective are recommended. This allows the selected algorithm to be optimized with regard to the respective treatment outcome. As was shown in section 2.2, the majority of works in the literature optimize and evaluate treatment decision support regarding agreement with expert recommendations or guidelines instead of outcome.

With this purpose in mind, both, the neighborhood-based CF methods, which estimate outcome and rank treatment options based on local data only, but also the model-based approaches show, in spite of the small training data sizes, great potential. However, in terms of the endpoints formulated in section 1.2, outcome prediction accuracy (RMSE) and agreement between ground truth and top-3 recommendation list (MAP@3), the much simpler CF algorithms are in no way inferior to the more sophisticated model-based approaches given the available data. The essential

strength of the CF approaches is twofold. On the one hand, the modeling based on local data clearly benefits accuracy when predicting outcome of the actually applied therapy. On the other hand, only treatment options are included into the recommendation list which are observed in that neighborhood of the target patient. Hence, CFs additionally feature the selection of a subset of therapy options which improves the recommendation quality, i.e. MAP@3. To summarize, outcome prediction accuracy and ranking capability benefit, a feature which is not given by the ML approaches.

Beyond this advantage, the CF methods bring the additional value of being very intuitive. Predictions and recommendations are, beyond attribute importance estimates, additionally 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. An exemplary Graphical User Interface (GUI) (dashboard), which is developed within the context of this work, is demonstrated in appendix B.5. Whereas figure C.1 shows the data input and presentation forms, figure C.2 shows the output of an exemplary therapy recommendation list in the form of bar charts with *affinity* score predictions. For a selected therapy option, summary statistics from the local neighborhood on which the recommendation is based are visualized in pie charts. This gives insight into decision-making and can additionally serve as a basis for integration of patient values and preferences into treatment decisions. Both are important features to push acceptance of such CDSSs as reported in the literature about CDSSs [29, 321]. Nonetheless, such interpretability issues are hardly addressed in the related works which were identified in sections 2.2 and 2.3. The demonstrated algorithms optimize recommendations in terms of the *affinity* score. Nevertheless, also other outcome measures, e.g. each of the indicators described in chapter 4, can be applied individually. Providing treatment recommendations based on a selected outcome aspect can facilitate to chose a treatment which meets a distinct patient preference such a low risk of ADEs.

One particular strength of the ML approaches is their superiority in terms of scalability as soon as larger amounts of data are to be processed. Computation complexity for prediction and ranking at run-time is negligible for trained models in comparison with the CF approaches. The very powerful GBM regression model provides accurate outcome predictions, however, at the expense of low MAP@3 scores. The SLIM recommender, on the other hand, provides better ranking capabilities. Yet, outcome prediction is inferior to many of the other proposed algorithms.

When comparing conventional and patient-data CF, it is shown that the therapy history seems to be a particularly important attribute which actually justifies the conventional CF approach. Even though the *cold start* issue is a limiting difficulty. For the practical application this means that this treatment history along with associated outcomes must be thoroughly documented. Considering either of the evaluation criteria, the patient-data CF approaches are clearly inferior to the conventional approaches. Extending the attribute space by additional patient characteristics is obviously not beneficial. There are two data properties that basically contribute to the

Table 8.1: Qualitative comparison of the proposed algorithms regarding the aspects scalability, interpretability, and the two evaluation criteria outcome prediction accuracy and recommendation quality.

Method	**Prediction**	**Ranking**	**Interpretability**	**Scalability**
CF (Cosine)	+	++	++	−
CF (Pearson)	+	++	++	−
CF (Manhatten)	++	−	++	−
CF (Euclidean)	++	−	++	−
DR (Gower)	−	+	++	−
DR-RBA (Gower)	+	++	++	−
DR (Euclidean)	−−	−	++	−
DR-LMNN (Euclidean)	−−	−	++	−
DR-Rules a (Gower)	−	−	++	−
DR-Rules b (Gower)	−	−−	++	−
DR-Rules c (Gower)	−	−−	++	−
DR-Impute 0 (Gower)	−	+	++	−
DR-Impute 1 (Gower)	−	−−	++	−
SLIM	−	+	+	+
GBM	++	−−	+	+
Average efficiency	−−	−−	−−	+
Overall popularity	−−	−−	−−	+

observed performance difference. Firstly, the significantly larger attribute space (25 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. 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. Based on the given data, however, results cannot be improved by the proposed supervised attribute scaling or attribute space transformation methods. It must be noted that the performance and reliability of attribute weights depend not only on the *a priori* assumptions of similarity and dissimilarity, but in particular on sufficient and meaningful training data. Also the implemented imputation strategies apparently do not favor the performance of the patient-data CF but the *Gower similarity* is obviously sufficiently capable of coping with the missing values. However, because of the prerequisite to use complete datasets in order to apply SLIM and LMNN, the proposed imputation strategies can be recommended for such patient data methods.

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 between recommendations and actually and successfully applied treatments. As was shown, the inter-rater agreement renders the ground truth of applied treatments rather unreliable regarding the MAP@3 results.

The uncertainty concerning the validity of the ground truth is also shown in the context of the post-filtering. The exclusion rules are hardly represented by the data. Due to the limited reliability of the recommendation ground truth, algorithm selection should be based on outcome prediction accuracy rather than the ranking of treatment options. Though, it can be expected that larger data volumes, especially if embedding experience from different physicians and facilities, will also render the MAP@3 score a more trustworthy evaluation criterion. With respect to the stated focus but also considering the interpretability issue, the conventional CF using *Minkowski metric* can be considered as the overall preferable algorithm.

A qualitative comparison of all presented algorithms regarding the discussed aspects scalability, interpretability, and the two evaluation criteria outcome prediction accuracy and recommendation quality, is provided in table 8.1. Beyond that, the advantages and disadvantages of the proposed approaches are summarized and compared in table D.1.

8.2 Future Perspectives

The major challenge and limitation of this work is the small data foundation on which it is based but also the low quality inherent from manual data extraction and structuring. This problem is reinforced by the comparatively large number of therapy options and the unbalanced distribution of applied treatments, which makes reliable modeling even more difficult. Two factors determine the demand for a large data foundation. 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. Benchmark datasets with suitable longitudinal data are unfortunately not available, which in turn emphasizes the uniqueness of this work. Furthermore, due to the poor relationship between data quantity and attribute space size, the *curse of dimensionality* is an omnipresent problem of this work. This becomes particularly evident as the additional comprehensive patient data offers hardly any advantages over the approaches using treatment history only. But also the potential of attribute weighting (RBA), attribute transformation (LMNN), and embedded attribute selection (SLIM, GBM) can hardly be exploited. It may be expected, though, that the model-based approaches, as well as RBA and LMNN, will develop their capabilities with larger and more representative data volumes.

Another critical issue is the aspect of only partially observed (*hidden*) ground truth [221], meaning that only outcome for one recommended and applied treatment option per consultation is available. On the background of the 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. Whereas both limitations can be countered by a large dataset that covers a wide variety of patients and treatment options, the hidden ground truth can also be tackled by samples rated by multiple experts.

As the essential prerequisite for successful modeling and consultation comparisons is to identify

the most important attributes, a more detailed analysis using attribute selection methods on a more comprehensive dataset is highly recommended. In this context, future work could also focus on exploring determining factors and attribute weights by including expert knowledge. It must be kept in mind, thought, that attribute importance as well as similarity of clinical cases is often subjective. To avoid reliability problems as mentioned above, parameters should hence be based on majority votes or consensus from large-scale surveys.

Essential for the recommended neighborhood-based CF methods is the identification of a neighborhood which is characterized by a high degree of homogeneity. Within this work, the demonstrated CF algorithms determine homogeneity by similarity measures. However, also indicators measuring purity in DTs such as *Shannon entropy* or *Gini index* (3.3.2) can be investigated. Finally, instead of defining a fixed neighborhood size K, utilizing a similarity or homogeneity threshold to determine the neighborhood size could be studied.

A noteworthy challenge in connection with the practical applicability of the neighborhood-based CF methods is the calculation effort at runtime. The user-based CF requires searching the entire consultation database which becomes increasingly impracticable as the amount of data increases. Conceivable approaches to tackle this challenge is the application of *k-dimensional trees* or predefined clusters of patients or consultations which are searched. *k-means* clustering, for example, involves the additional attraction of deriving prototypes or templates of patient representations that can be used for interpretation purposes. Nevertheless, there always is a trade-off between prediction accuracy and scalability. Cluster-based methods may exhibit better scalability than bare neighborhood-based modeling but may not satisfactorily address the prediction for patients with rare characteristics.

Model-based approaches typically are not subject to efficiency issues during the prediction phase. As a consequence, larger datasets also bear the potential to make more complex ML models applicable and attractive. Especially deep neural networks ("Deep Learning") are currently experiencing great popularity because of their good performance and capability to learn difficult patterns. They are, however, dependent on large training data sets and interpretability or providing explanation of the model's reasoning becomes difficult. A key drawback of all proposed algorithms is there reduced capability to consider the temporal dependencies of consultations. The sequence of consultations can be considered as observations over a defined period of time resulting in temporal sequences of varying length. According to the taxonomy of tasks which make use of the ordering property of sequential data formulated in [90], a reasonable choice is to consider treatment recommendation as *sequence classification*. Here, a single class label is predicted which characterizes the entire (multivariate) input sequence. Exemplary algorithms capable of performing such tasks while considering time dependencies are e.g. HMMs [277] but also RNNs such as LSTMs [157] or Gated Recurrent Units (GRUs) [66], which are successfully applied in other domains such as time-series classification. Besides interpretability issues, again the required data volume is the exclusion criterion of such methods within the scope of this thesis.

It is important to bear in mind that clinical data today is an expensive asset. In particular, feedback on interventions from longitudinal observations is difficult to obtain and often associated with long time constants. This shortage also concerns development and evaluation of CDSSs, which can be considered as the 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.
One alternative to larger datasets could be, at least for research and development purposes, to further simplify the task, e.g. by recommending drug groups. In case of the present example, conventional, biopharmaceutical and other medications could be recommended instead of individual pharmaceuticals. Biopharmaceutical drugs could be further grouped according to their mechanism, namely into TNF-α antagonists, IL-12/13 antibody, and IL-17 antibody drugs.
However, it can be assumed that the digitization of the health care system will continue to advance in the coming years. As a consequence, also EHRs will be increasingly captured and interchanged in standardized formats (HL7[1], FHIR[2]). But also government initiatives such as the *Telematikinfrastruktur*[3], which aims to digitally network all actors in the healthcare system, and the *Digitale-Versorgung-Gesetz* will drive the development of health apps (*Digitale Gesundheitsanwendungen*[4]) and digitalized health data. Nevertheless, the question of whether more data alone is beneficial depends heavily on the underlying structure of the data and cannot be generalized.

The envision of a recommender system, as introduced and visualized in section 1.2, implements a closed feedback loop which ensures that the proposed system learns from every patient at every consultation. To do so, application specific data, such as the condition related attributes, recommended and applied treatments, and the associated outcomes are recorded in a structured and standardized fashion. The goal is to gain experience for the full range of relevant treatment options from many similar patients and a large variety of patients. Additionally, in order to avoid bias, the system collects data at multiple facilities (multi-center) with different geographic locations. Assuming that data is collected from a system in operation, a strategically expedient action could be to provide recommendations or other decision support based on heuristics and rules until a data-driven approach provides added value. Such static systems, however, require careful maintenance in order to keep the knowledge source updated and to preserve the benefit for the user. Real benefit of such a CDSS can only be delivered if is is, as already discussed in chapter 2, seamlessly integration into the clinical workflow and into the healthcare ecosystem. This benefit must be recognizable and measurable such as in terms of saved time or money or improved outcome.

With this work, the author hopes to contribute ideas and basics to the state of the art, development and application of therapy recommendation systems. Overall, it is hoped that such systems, in a reliable and clinically evaluated form, will find their way into practical medicine

[1] `https://www.hl7.org/`
[2] `https://www.hl7.org/fhir/`
[3] `https://www.gematik.de/telematikinfrastruktur/`
[4] `https://www.bfarm.de/DE/Medizinprodukte/DVG/_node.html`

in order to improve health care in terms of objectivity, safety and patient satisfaction.

8.3 Summary

Within this thesis, an exemplary CDSS is developed which provides individualized pharmaceutical drug and drug combination recommendations for patients suffering from *Psoriasis*. Therefore, data representing patients and consultations were extracted from health records and transformed into a structured format. These representations allow for descriptive statistical analyses and development and evaluation of data-driven prediction algorithms. The intention is to predict patient-specific treatment outcomes in order to derive recommendations. Suchlike, treatments are supposed to be recommended independent from overall popularity or average efficiency but personalized to a patient and consultation. Moreover, the data-driven approach does without domain knowledge, adapts to the underlying data and is capable of improving with a growing database.

CF algorithms, derived from the RS domain, but also state of the art ML algorithms are adapted to the problem at hand and are evaluated regarding prediction accuracy (RMSE). In order to measure outcome, a summarizing score, denoted as *affinity* score, is developed which combines multiple outcome aspects as *efficiency*, relative change of the PASI and ADEs. Additionally, the proposed algorithms' capability to rank treatment options suchlike that the potentially most optimal treatments are preferred is studied. The ranked therapy lists are rated by MAP@3, a score derived from IR. Two different versions of input information are contrasted: approaches solely relying on treatment history and approaches which incorporate a wide range of patient describing attributes to represent a patient and consultation. As especially in the latter representations missing values are pervasive, different imputation strategies are investigated, depending on the mechanisms underlying the missing values. Finally, various combinations of evidence-based and expert-based exclusion rules are implemented in order to filter potentially inadequate treatment options from the recommendation lists. All algorithms and variations are optimized and evaluated in a nested cross-validation loop in order to use the limited amount of available data efficiently.

The estimated generalization performance shows that all proposed algorithms are capable of outperforming the baseline predictions and recommendations. The results also show that the much simpler CF algorithms are in no way inferior to the more sophisticated model-based ML approaches given the available data. This can be attributed to their ability to simultaneously predict outcome and select a subset of treatment options based on similar cases.
Concerning outcome prediction and considering interpretabililty of recommendations it was shown that the conventional CF approaches utilizing information on treatment history only and using *Minkowski* metrics are the preferred algorithms. The conventional CF utilizing correlation-based similarity measures, on the other hand, outperforms all other studied approaches regarding ranking quality of recommendations. As is further shown, neither in terms of the estimated outcome prediction, nor in terms of recommendation agreement with the ground truth, generalization performance of the CF algorithms benefits from the incorporation of ad-

ditional patient data. Given the available data, this limitation could not be eliminated with the investigated attribute weighting and data transformation approaches. Also the model-based approaches, which also depend on sufficient and informative data, are in total inferior regarding the addressed major objectives: good prediction accuracy and interpretable recommendation. Nevertheless, when comparing recommendations generated by different versions of the proposed therapy recommender system with those provided by human experts, the recommender system is inferior. Inter-rater agreement (*Cohen's Kappa*) between automatically generated recommendations and given ground truth is even worse than the already small agreement between experts and ground truth which was revealed in an own preliminary study.

Beyond the recommender system algorithms, further own studies, which address the quantification of health status and outcome based on raw vital signs, are demonstrated in this thesis. Those include sleep stage classification based on cardiorespiratory signals and PD gait assessment. Finally, also the application of sentiment analysis methods to patient reviews, targeting the extraction of information on experience with applied treatments, was studied and is described.

References

[1] Açıcı, K. et al. 'A random forest method to detect parkinson's disease via gait analysis'. In: *EANN 2017: Engineering Applications of Neural Networks*. Vol. 744. Cham: Springer, 2017, pp. 609–619.

[2] Adomavicius, G. and Tuzhilin, A. 'Context-aware recommender systems'. In: *Recommender Systems Handbook*. 2nd ed. New York: Springer, 2015, pp. 191–226.

[3] Agarwal, D. and Chen, B. C. 'Regression-based latent factor models'. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*. Paris, France, 2009, pp. 19–27.

[4] Aggarwal, C. C. *Recommender Systems*. 1st ed. Basel, Switzerland: Springer, 2012.

[5] Aggarwal, K. et al. 'A Structured Learning Approach with Neural Conditional Random Fields for Sleep Staging'. In: *Proceedings of the 2018 IEEE International Conference on Big Data, Big Data 2018*. Seattle, Washington, USA, 2018, pp. 1318–1327.

[6] Akl, E. A. et al. 'Living systematic reviews: 4. Living guideline recommendations'. In: *Journal of Clinical Epidemiology* 91 (2017), pp. 47–53.

[7] Alder, H. et al. 'Computer-Based Diagnostic Expert Systems in Rheumatology: Where Do We Stand in 2014?' In: *International Journal of Rheumatology* 2014 (2014), pp. 1–10.

[8] Alfonso, L. J., Herrero, M. A. and Núñez, L. 'A dose-volume histogram based decision-support system for dosimetric comparison of radiotherapy treatment plans'. In: *Radiation Oncology* 10 (2015), pp. 1–9.

[9] Ambady, N. and Rosenthal, R. 'Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis'. In: *Psychological Bulletin* 111.2 (1992), pp. 256–274.

[10] Amin Morid, M., Liu Sheng, O. R. and Abdelrahman, S. 'Leveraging Time Series Data in Similarity Based Healthcare Predictive Models: The Case of Early ICU Mortality Prediction'. In: *Proceedings of the 23rd Americas Conference on Information Systems*. Boston, Massachusetts, USA, 2017.

[11] Ampudia-Blasco, F. J. et al. 'A decision support tool for appropriate glucose-lowering therapy in patients with type 2 diabetes'. In: *Diabetes Technology and Therapeutics* 17.3 (2015), pp. 194–202.

[12] Augustin, M. et al. 'Disease severity, quality of life and health care in plaque-type psoriasis: A multicenter cross-sectional study in Germany'. In: *Dermatology* 216.4 (2008), pp. 366–372.

[13] Augustin, M. et al. 'Co-morbidity and age-related prevalence of psoriasis: Analysis of health insurance data in Germany'. In: *Acta Dermato-Venereologica* 90.2 (2010), pp. 147–151.

[14] Avorn, J. 'The psychology of clinical decision making - Implications for medication use'. In: *New England Journal of Medicine* 378 (2018), pp. 689–691.

[15] AWMF - Association of the Scientific Medical Societies in Germany. *Regelwerk Leitlinien: Stufenklassifikation*. URL: `https://www.awmf.org/en/clinical-practice-guidelines/awmf-guidance/cpg-development/awmf-regelwerk-01-planung-und-organisation/po-stufenklassifikation.html` (visited on 27/09/2020).

[16] Backhaus, K. et al. *Multivariate Analysemethoden - Eine anwendungsorientierte Einführung*. 14th ed. Berlin Heidelberg: Springer, 2016.

[17] Bal, M. et al. 'Performance Evaluation of the Machine Learning Algorithms Used in Inference Mechanism of a Medical Decision Support System'. In: *ScientificWorldJournal* 2014 (2014), pp. 1–15.

[18] Banning, M. 'A review of clinical decision making: Models and current research'. In: *Journal of Clinical Nursing* 17.2 (2008), pp. 187–195.

[19] Barnett, G. O. et al. 'DXplain: An Evolving Diagnostic Decision-Support System'. In: *JAMA: The Journal of the American Medical Association* 258.1 (1987), pp. 67–74.

[20] Barratt, A. 'Evidence Based Medicine and Shared Decision Making: The challenge of getting both evidence and preferences into health care'. In: *Patient Education and Counseling* 73.3 (2008), pp. 407–412.

[21] Bates, D. W. et al. 'Ten Commandments for Effective Clinical Decision Support : Making the Practice of Evidence-based Medicine a Reality'. In: *Journal of American Medical Informatics Association* 10.6 (2003), pp. 523–530.

[22] Batista, G. E. and Monard, M. C. 'An analysis of four missing data treatment methods for supervised learning'. In: *Applied Artificial Intelligence* 17.5-6 (2003), pp. 519–533.

[23] Bauer-Mehren, A. et al. 'Network analysis of unstructured EHR data for clinical research'. In: *Proceedings of the AMIA Joint Summits on Translational Science*. San Francisco, California, USA, 2013, pp. 14–18.

[24] Beaulieu-Jones, B. K. and Moore, J. H. 'MISSING DATA IMPUTATION IN THE ELECTRONIC HEALTH RECORD USING DEEPLY LEARNED AUTOENCODERS.' In: *Proceedings of the Pacific Symposium on Biocomputing*. Fairmont Orchid, Hawaii, 2017, pp. 207–218.

[25] Beeler, P. E., Bates, D. W. and Hug, B. L. 'Clinical decision support systems'. In: *Swiss medical weekly* 144 (2014), pp. 1–7.

[26] Begum, S. et al. 'Case-based reasoning systems in the health sciences: A survey of recent trends and developments'. In: *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 41.4 (2011), pp. 421–434.

[27] Bell, R. M. and Koren, Y. 'Scalable collaborative filtering with jointly derived neighborhood interpolation weights'. In: *Proceedings of the IEEE International Conference on Data Mining, ICDM*. New Orleans, Louisiana, USA, 2007, pp. 43–52.

[28] Bellet, A., Habrard, A. and Sebban, M. 'A Survey on Metric Learning for Feature Vectors and Structured Data'. In: *arXiv.org* (2013). URL: `http://arxiv.org/abs/1306.6709`.

[29] Berner, E. S. *Clinical Decision Support Systems*. 3rd ed. Basel, Switzerland: Springer International Publishing, 2016.

[30] Bero, L. et al. 'Factors associated with findings of published trials of drug-drug comparisons: Why some statins appear more efficacious than others'. In: *PLoS Medicine* 4.6 (2007), pp. 1001–1010.

[31] Bindoff, I. et al. 'The potential for intelligent decision support systems to improve the quality and consistency of medication reviews'. In: *Journal of Clinical Pharmacy and Therapeutics* 37.4 (2012), pp. 452–458.

[32] Bindoff, I. K. et al. 'Development of an intelligent decision support system for medication review'. In: *Journal of Clinical Pharmacy and Therapeutics* 32.1 (2007), pp. 81–88.

[33] Bishop, C. M. *Pattern Recognition and Machine Learning*. 1st ed. New York: Springer, 2006.

[34] Boostani, R., Karimzadeh, F. and Nami, M. 'A comparative review on sleep stage classification methods in patients and healthy individuals'. In: *Computer Methods and Programs in Biomedicine* 140 (2017), pp. 77–91.

[35] Boriah, S., Chandola, V. and Kumar, V. 'Similarity Measures for Categorical Data: A Comparative Evaluation'. In: *Proceedings of the 2008 SIAM International Conference on Data Mining, SDM'08*. Atlanta, Georgia, USA, 2008, pp. 243–254.

[36] Breese, J., Heckerman, D. and Kadie, C. 'Empirical Analysis of Predicive Algorithms for Collaborative Filtering'. In: *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, UAI'98*. Madison, Wisconsin, USA, 1998, pp. 43–52.

[37] Breiman, L. et al. *Classification and Regression Trees*. Monterey, California, USA: Wadsworth and Brooks, 1984.

[38] Breiman, L. 'Arcing classifiers'. In: *Annals of Statistics* 26.3 (1998), pp. 801–849.

[39] Breiman, L. 'Bagging predictors'. In: *Machine Learning* 24.2 (1996), pp. 123–140.

[40] Breiman, L. 'Pasting Small Votes for Classification in Large Databases and On-Line'. In: *Machine Learning* 103 (1999), pp. 85–103.

[41] Breiman, L. 'Random forests'. In: *Machine learning* (2001), pp. 5–32.

[42] Bresch, E., Großekathöfer, U. and Garcia-Molina, G. 'Recurrent deep neural networks for real-time sleep stage classification from single channel EEG'. In: *Frontiers in Computational Neuroscience* 12 (2018), pp. 1–12.

[43] Bright, T. J. et al. 'Effect of clinical decision-support systems: A systematic review'. In: *Annals of Internal Medicine* 157.1 (2012), pp. 29–43.

[44] Brown, S.-A. A. 'Patient Similarity: Emerging Concepts in Systems and Precision Medicine'. In: *Frontiers in Physiology* 7 (2016), pp. 1–6.

[45] Buchkowsky, S. S. and Jewesson, P. J. 'Industry Sponsorship and Authorship of Clinical Trials over 20 Years'. In: *Annals of Pharmacotherapy* 38.4 (2004), pp. 579–585.

[46] Buciluǎ, C., Caruana, R. and Niculescu-Mizil, A. 'Model compression'. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06*. Philadelphia, Pennsylvania, USA, 2006, pp. 535–541.

[47] Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM). *Pharmakovigilanz*. URL: `https://www.bfarm.de/DE/Arzneimittel/Pharmakovigilanz/%7B%5C_%7Dnode.html` (visited on 09/01/2020).

[48] Burke, R. 'Hybrid Recommender Systems: Survey and Experiments'. In: *User Modeling and User-Adapted Interaction* 12.4 (2002), pp. 331–370.

[49] Calero Valdez, A. et al. 'Recommender Systems for Health Informatics: State-of-the-Art and Future Perspectives'. In: *Machine Learning for Health Informatics. Lecture Notes in Computer Science*. Cham: Springer, 2016, pp. 391–414.

[50] Campbell-Scherer, D. 'Multimorbidity: A challenge for evidence-based medicine'. In: *Evidence-Based Medicine* 15.6 (2010), pp. 165–166.

[51] Campillo-Gimenez, B. et al. 'Improving Case-Based Reasoning Systems by Combining K-Nearest Neighbour Algorithm with Logistic Regression in the Prediction of Patients' Registration on the Renal Transplant Waiting List'. In: *PLoS ONE* 8.9 (2013), pp. 1–10.

[52] Canny, J. 'Collaborative Filtering with Privacy via Factor Analysis'. In: *Proceeding for the 25th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '02*. Tampere, Finland, 2002, pp. 238–245.

[53] Celi, L. A., Zimolzak, A. J. and Stone, D. J. 'Dynamic clinical data mining: Search engine-based decision'. In: *Journal of Medical Internet Research* 16.6 (2014), pp. 1–7.

[54] Chan, A. W. et al. 'Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles'. In: *Journal of the American Medical Association* 291.20 (2004), pp. 2457–2465.

[55] Chan, L. W. et al. 'Machine learning of patient similarity: A case study on predicting survival in cancer patient after locoregional chemotherapy'. In: *Proceedings of the 2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2010*. Hong Kong, 2010, pp. 467–470.

[56] Chan, L. W. et al. 'PubMed-supported clinical term weighting approach for improving inter-patient similarity measure in diagnosis prediction'. In: *BMC Medical Informatics and Decision Making* 15.1 (2015), pp. 1–8.

[57] Chawla, N. V. and Davis, D. A. 'Bringing big data to personalized healthcare: A patient-centered framework'. In: *Journal of General Internal Medicine* 28 (2013), pp. 660–665.

[58] Chen, C. et al. 'A guideline-based decision support for pharmacological treatment can improve the quality of hyperlipidemia management'. In: *Computer Methods and Programs in Biomedicine* 97.3 (2010), pp. 280–285.

[59] Chen, J. H. and Altman, R. B. 'Automated physician order recommendations and outcome predictions by data-mining electronic medical records'. In: *Proceedings of the AMIA Joint Summits on Translational Science proceedings*. San Francisco, California, USA, pp. 206–210.

[60] Chen, J. H. and Altman, R. B. 'Mining for clinical expertise in (undocumented) order sets to power an order suggestion system'. In: *Proceedings of the AMIA Joint Summits on Translational Science proceedings*. San Francisco, California, USA, 2013, pp. 34–38.

[61] Chen, S., Ma, B. and Zhang, K. 'On the similarity metric and the distance metric'. In: *Theoretical Computer Science* 410.24-25 (2009), pp. 2365–2376.

[62] Chen, Y., Elenee Argentinis, J. D. and Weber, G. 'IBM Watson: How Cognitive Computing Can Be Applied to Big Data Challenges in Life Sciences Research.' In: *Clinical therapeutics* 38.4 (2016), pp. 688–701. URL: `http://www.sciencedirect.com/science/article/pii/S0149291815013168`.

[63] Chen, Y., Zhu, X. and Chen, W. 'Automatic sleep staging based on ECG signals using hidden Markov models'. In: *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015*. Milan, Italy, 2015, pp. 530–533.

[64] Chi, C. L. et al. 'Optimal decision support rules improve personalize warfarin treatment outcomes'. In: *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2016*. Orlando, Florida, USA, 2016, pp. 2594–2597.

[65] Chiang, W.-H. et al. 'Drug Recommendation toward Safe Polypharmacy'. In: *arXiv.org* (2018). URL: `https://arxiv.org/abs/1803.03185`.

[66] Cho, K. et al. 'Learning phrase representations using RNN encoder-decoder for statistical machine translation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*. Doha, Qatar, 2014, pp. 1724–1734.

[67] Choi, I. et al. 'Incidence and treatment costs attributable to medication errors in hospitalized patients'. In: *Research in Social and Administrative Pharmacy* 12.3 (2016), pp. 428–437.

[68] Choi, S. S., Cha, S. H. and Tappert, C. C. *A survey of binary similarity and distance measures*. Tech. rep. 2009, pp. 80–85.

[69] Chouchou, F. and Desseilles, M. 'Heart rate variability: A tool to explore the sleeping brain?' In: *Frontiers in Neuroscience* 8 (2014), pp. 1–9.

[70] Civan, M. M. et al. 'Regulatory volume decrease by cultured non-pigmented ciliary epithelial cells'. In: *Experimental Eye Research* 54.2 (1992), pp. 181–191.

[71] Cost, S. and Salzberg, S. 'A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features'. In: *Machine Learning* 10.1 (1993), pp. 57–78.

[72] Cottrell, S. S. 'A simple method for finding the scattering coefficients of quantum graphs'. In: *Journal of Mathematical Physics* 56.9 (2015), pp. 1–34.

[73] Cover, T. M. and Hart, P. E. 'Nearest Neighbor Pattern Classification'. In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[74] Croskerry, P. 'Clinical cognition and diagnostic error: Applications of a dual process model of reasoning'. In: *Advances in Health Sciences Education* 14 (2009), pp. 27–35.

[75] Daemen, A. and De Moor, B. 'Development of a kernel function for clinical data'. In: *Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society: Engineering the Future of Biomedicine, EMBC 2009*. Minneapolis, Minnesota, USA, 2009, pp. 5913–5917.

[76] Daliri, M. R. 'Chi-square distance kernel of the gaits for the diagnosis of Parkinson's disease'. In: *Biomedical Signal Processing and Control* 8.1 (2013), pp. 66–70.

[77] Danker-Hopfe, H. et al. 'Interrater reliability between scorers from eight European sleep laboratories in subjects with different sleep disorders'. In: *Journal of Sleep Research* 13.1 (2004), pp. 63–69.

[78] David, G., Bernstein, L. and Coifman, R. R. 'Generating evidence based interpretation of hematology screens via anomaly characterization'. In: *Open Clinical Chemistry Journal* 4.1 (2011), pp. 10–16.

[79] Davis, D. A. et al. 'Time to CARE: A collaborative engine for practical disease prediction'. In: *Data Mining and Knowledge Discovery* 20.3 (2010), pp. 388–415.

[80] Dayton, C. S. et al. 'Evaluation of an Internet-based decision-support system for applying the ATS/CDC guidelines for tuberculosis preventive therapy'. In: *Medical Decision Making* 20.1 (2000), pp. 1–6.

[81] De Clercq, P. A. et al. 'Approaches for creating computer-interpretable guidelines that facilitate decision support'. In: *Artificial Intelligence in Medicine* 31.1 (2004), pp. 1–27.

[82] De Croon, R. et al. 'Health Recommender Systems: Systematic Review'. In: *Journal of Medical Internet Research* 23.6 (June 2021), e18035.

[83] Deerwester, S. et al. 'Indexing by latent semantic analysis'. In: *Journal of the American Society for Information Science* 41.6 (1990), pp. 391–407.

[84] Deitelzweig, S. B. et al. 'Reviewing a clinical decision aid for the selection of anticoagulation treatment in patients with nonvalvular atrial fibrillation: Applications in a US managed care health plan database'. In: *Clinical Therapeutics* 36.11 (2014), pp. 1566–1573.

[85] Del Mar, C., Doust, J. and Glasziou, P. *Clinical Thinking: Evidence, Communication and Decision-Making*. 1st ed. Oxford, England: BMJ Books, 2007.

[86] Denecke, K. and Denecke, K. 'Sentiment Analysis from Medical Texts'. In: *Health Web Science*. Cham: Springer, 2015, pp. 83–98.

[87] Deshpande, M. and Karypis, G. 'Item-based top-N recommendation algorithms'. In: *ACM Transactions on Information Systems* 22.1 (2004), pp. 143–177.

[88] Deshpande, R. R. et al. 'Knowledge-driven decision support for assessing dose distributions in radiation therapy of head and neck cancer'. In: *International Journal of Computer Assisted Radiology and Surgery* 11.11 (2016), pp. 2071–2083.

[89] Deuschl, G., Oertel, W. and Reichmann, H. *S3-Leitlinie Idiopathisches Parkinson-Syndrom*. 2016. URL: `www.dgn.org` (visited on 25/09/2020).

[90] Dietterich, T. G. 'Machine learning for sequential data: A review'. In: *Lecture Notes in Computer Science*. Berlin Heidelberg: Springer, 2002, pp. 15–30.

[91] Domingos, P. and Pazzani, M. 'On the Optimality of the Simple Bayesian Classifier under Zero-One Loss'. In: *Machine Learning* 29.2-3 (1997), pp. 103–130.

[92] Dong, Y. and Peng, C. Y. J. 'Principled missing data methods for researchers'. In: *SpringerPlus* 2.1 (2013), pp. 1–17.

[93] Draper, B., Kaito, C. and Bins, J. 'Iterative Relief'. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. Vol. 6. Madison, 2003, pp. 62–62.

[94] Duan, L., Street, W. N. and Xu, E. 'Healthcare information systems: Data mining methods in the creation of a clinical recommender system'. In: *Enterprise Information Systems* 5.2 (2011), pp. 169–181.

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

[96] Ebadollahi, S. et al. 'Predicting Patient's Trajectory of Physiological Data using Temporal Trends in Similar Patients: A System for Near-Term Prognostics'. In: *Proceedings of the AMIA Symposium*. Washington, DC, USA, 2010, pp. 192–196.

[97] Ebrahimi, F. et al. 'Automatic sleep staging using empirical mode decomposition, discrete wavelet transform, time-domain, and nonlinear dynamics features of heart rate variability signals'. In: *Computer Methods and Programs in Biomedicine* 112.1 (2013), pp. 47–57.

[98] Eccher, C., Seyfang, A. and Ferro, A. 'Implementation and evaluation of an Asbru-based decision support system for adjuvant treatment in breast cancer'. In: *Computer Methods and Programs in Biomedicine* 117.2 (2014), pp. 308–321.

[99] El Maachi, I., Bilodeau, G. A. and Bouachir, W. 'Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait'. In: *Expert Systems with Applications* 143 (2020).

[100] Elliott, J. H. et al. 'Living systematic review: 1. Introduction—the why, what, when, and how'. In: *Journal of Clinical Epidemiology* 91 (2017), pp. 23–30.

[101] Exarchos, T. P. et al. 'Patient specific cardiovascular risk assessment and treatment decision support based on multiscale modelling and medical guidelines'. In: *Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2011*. Boston, Massachusetts, USA, 2011, pp. 838–841.

[102] Faries, D. E. et al. 'Local control for identifying subgroups of interest in observational research: Persistence of treatment for major depressive disorder'. In: *International Journal of Methods in Psychiatric Research* 22.3 (2013), pp. 185–194.

[103] Finlay, A. Y. and Khan, G. K. 'Dermatology Life Quality Index (DLQI)—a simple practical measure for routine clinical use'. In: *Clinical and Experimental Dermatology* 19.3 (1994), pp. 210–216.

[104] Fleiss, J. L., Levin, B. and Paik, M. C. 'The Measurement of Interrater Agreement'. In: *Statistical Methods for Rates and Proportions* (2004), pp. 598–626.

[105] Folino, F. and Pizzuti, C. 'A comorbidity-based recommendation engine for disease prediction'. In: *Proceedings of the IEEE Symposium on Computer-Based Medical Systems*. Perth, 2010, pp. 6–12.

[106] Folino, F. and Pizzuti, C. 'A recommendation engine for disease prediction'. In: *Information Systems and e-Business Management* 13.4 (2015), pp. 609–628.

[107] Fonseca, P. et al. 'Sleep stage classification with ECG and respiratory effort'. In: *Physiological Measurement* 36.10 (2015), pp. 2027–2040.

[108] Fortin, M. et al. 'Randomized controlled trials: Do they have external validity for patients with multiple comorbidities?' In: *Annals of Family Medicine* 4.2 (2006), pp. 104–108.

[109] Fox, G. H. *File: Psoriasis guttata.jpg*. 2010. URL: `https://commons.wikimedia.org/wiki/File:Psoriasis%7B%5C_%7Dguttata.jpg` (visited on 19/06/2020).

[110] Fox, G. H. *File:Psoriasis manum.jpg*. 2010. URL: `https://commons.wikimedia.org/wiki/File:Psoriasis%7B%5C_%7Dmanum.jpg` (visited on 19/06/2020).

[111] Fox, S. H. et al. 'International Parkinson and movement disorder society evidence-based medicine review: Update on treatments for the motor symptoms of Parkinson's disease'. In: *Movement Disorders* 33.8 (2018), pp. 1248–1266.

[112] Frankovich, J., Longhurst, C. A. and Sutherland, S. M. 'Evidence-based medicine in the EMR era'. In: *New England Journal of Medicine* 365.19 (2011), pp. 1758–1759.

[113] Fredriksson, T. and Pettersson, U. 'Severe psoriasis — Oral therapy with a new retinoid'. In: *Dermatology* 157.4 (1978), pp. 238–244.

[114] Freund, Y. and Schapire, R. E. 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting'. In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139.

[115] Freund, Y. and Schapire, R. E. 'Experiments with a New Boosting Algorithm'. In: *Proceedings of the 13th International Conference on Machine Learning*. 1996, pp. 148–156.

[116] Friedman, J., Hastie, T. and Tibshirani, R. 'Additive logistic regression: A statistical view of boosting'. In: *Annals of Statistics* 28.2 (2000), pp. 337–407.

[117] Friedman, J. H. 'Greedy function approximation: A gradient boosting machine'. In: *Annals of Statistics* 29.5 (2001), pp. 1189–1232.

[118] Friedman, J. H. 'Stochastic gradient boosting'. In: *Computational Statistics and Data Analysis* 38.4 (2002), pp. 367–378.

[119] Friedman, M. 'The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance'. In: *Journal of the American Statistical Association* 32.200 (1937), pp. 675–701.

[120] Gallego, B. et al. 'Bringing cohort studies to the bedside: Framework for a 'green button' to support clinical decision-making'. In: *Journal of Comparative Effectiveness Research* 4.3 (2015), pp. 191–197.

[121] Garg, A. X. et al. 'Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review'. In: *Journal of the American Medical Association* 293.10 (2005), pp. 1223–1238.

[122] Gerlach, M., Reichmann, H. and Riederer, P. *Die Parkinson-Krankheit: Grundlagen, Klinik, Therapie*. 4th ed. Wien: Springer, 2007.

[123] Gers, F. A., Schmidhuber, J. and Cummins, F. 'Learning to forget: Continual prediction with LSTM'. In: *Neural Computation* 12.10 (2000), pp. 2451–2471.

[124] Goeuriot, L. et al. 'Sentiment lexicons for health-related opinion mining'. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, IHI'12*. Miami, Florida, USA, 2012, pp. 219–225.

[125] Goldammer, M. et al. 'Specializing CNN Models for Sleep Staging based on Heart Rate'. In: *Proceedings of the 47th Computing in Cardiology Conference, CinC '20*. Rimini, Italy, 2020.

[126] Goldberg, D. et al. 'Using collaborative filtering to Weave an Information tapestry'. In: *Communications of the ACM* 35.12 (1992), pp. 61–70.

[127] Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning*. Cambridge, Massachusetts, USA: MIT Press, 2016. URL: `http://www.deeplearningbook.org`.

[128] Gopalakrishnan, V. and Ramaswamy, C. 'Patient opinion mining to analyze drugs satisfaction using supervised learning'. In: *Journal of Applied Research and Technology* 15.4 (2017), pp. 311–319.

[129] Gottlieb, A. et al. 'A method for inferring medical diagnoses from patient similarities'. In: *BMC Medicine* 11.194 (2013), pp. 1–9.

[130] Goud, R., Hasman, A. and Peek, N. 'Development of a guideline-based decision support system with explanation facilities for outpatient therapy'. In: *Computer Methods and Programs in Biomedicine* 91.2 (Aug. 2008), pp. 145–153.

[131] Gower, J. C. 'A General Coefficient of Similarity and Some of Its Properties'. In: *Biometrics* 27.4 (1971), pp. 857–871.

[132] Graham, J. W. 'Missing data analysis: Making it work in the real world'. In: *Annual Review of Psychology* 60 (2009), pp. 549–576.

[133] Gräßer, F., Malberg, H. and Zaunseder, S. 'Neighborhood Optimization for Therapy Decision Support'. In: *Current Directions in Biomedical Engineering* 5.1 (2019), pp. 1–4.

[134] Gräßer, F. et al. 'Application of recommender system methods for therapy decision support'. In: *Proceedings of the 18th IEEE International Conference on e-Health Networking, Applications and Services, Healthcom 2016*. Munich, Germany, 2016, pp. 1–6.

[135] Gräßer, F. et al. 'Aspect-Based sentiment analysis of drug reviews applying cross-Domain and cross-Data learning'. In: *Proceedings of the 2018 International Conference on Digital Health, DH '18*. Lyon, France, 2018, pp. 121–125.

[136] Gräßer, F. et al. 'Therapy Decision Support Based on Recommender System Methods'. In: *Journal of Healthcare Engineering* 2017 (2017), pp. 1–12.

[137] Grasso, D. J., Ford, J. D. and Lindhiem, O. 'A Patient-Centered Decision-Support Tool Informed by History of Interpersonal Violence: "Will This Treatment Work for Me?"' In: *Journal of Interpersonal Violence* 31.3 (2016), pp. 465–480.

[138] Griffiths, C. E. and Barker, J. N. 'Pathogenesis and clinical features of psoriasis'. In: *Lancet* 370.9583 (2007), pp. 263–271.

[139] Groves, M. 'Understanding clinical reasoning: The next step in working out how it really works'. In: *Medical Education* 46.5 (2012), pp. 444–446.

[140] Gunawardana, A. and Shani, G. 'A survey of accuracy evaluation metrics of recommendation tasks'. In: *Journal of Machine Learning Research* 10 (2009), pp. 2935–2962.

[141] Gutierrez, G. et al. 'Respiratory rate variability in sleeping adults without obstructive sleep apnea'. In: *Physiological Reports* 4.17 (2016), pp. 1–9.

[142] Haas, P. J. *Medizinische Informationssysteme und Elektronische Krankenakten*. 1st ed. Berlin Heidelberg: Springer, 2005.

[143] Hall, M. A. and Smith, L. A. 'Feature subset selection: a correlation based filter approach. Progress in Connectionist-based Information Systems'. In: *Proceedings of the International Conference on Neural Information Processing and Intelligent Information Systems, ICONIP 1997*. Dunedin, New Zealand, 1997, pp. 855–858.

[144] Halldorsson, B. V. et al. 'A Clinical Decision Support System for the Diagnosis, Fracture Risks and Treatment of Osteoporosis'. In: *Computational Intelligence Techniques in Medicine* 2015.189769 (2015), pp. 1–7.

[145] Hao, F. and Blair, R. H. 'A comparative study: Classification vs. user-based collaborative filtering for clinical prediction'. In: *BMC Medical Research Methodology* 16.172 (2016), pp. 1–14.

[146] Harding, K. and Feldman, M. 'Sleep Disorders and Sleep Deprivation: An Unmet Public Health Problem'. In: *Journal of the American Academy of Child & Adolescent Psychiatry* 47.4 (2008), pp. 473–474.

[147] Hartge, F., Wetter, T. and Haefeli, W. E. 'A similarity measure for case based reasoning modeling with temporal abstraction based on cross-correlation'. In: *Computer Methods and Programs in Biomedicine* 81.1 (2006), pp. 41–48.

[148] Hassan, S. and Syed, Z. 'From netflix to heart attacks: Collaborative filtering in medical datasets'. In: *Proceedings of the 1st ACM International Health Informatics Symposium, IHI'10*. Arlington, Virginia, USA, 2010, pp. 128–134.

[149] Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning*. 2nd ed. New York: Springer, 2009.

[150] Hemens, B. J. et al. 'Computerized clinical decision support systems for drug prescribing and management: A decision-maker-researcher partnership systematic review'. In: *Implementation Science* 6.89 (2011), pp. 1–17.

[151] Henriques, J. et al. 'Prediction of Heart Failure Decompensation Events by Trend Analysis of Telemonitoring Data'. In: *IEEE Journal of Biomedical and Health Informatics* 19.5 (2015), pp. 1757–1769.

[152] Herlocker, J. L. et al. 'An Algorithmic Framework for Performing Collaborative Filtering'. In: *SIGIR Forum* 51.2 (2017), pp. 227–234.

[153] Herlocker, J. L. et al. 'Evaluating collaborative filtering recommender systems'. In: *ACM Transactions on Information Systems* 22.1 (2004), pp. 5–53.

[154] Hernando, M. E. et al. 'Evaluation of DIABNET, a decision support system for therapy planning in gestational diabetes'. In: *Computer Methods and Programs in Biomedicine* 62.3 (2000), pp. 235–248.

[155] Hielscher, T. et al. 'Using participant similarity for the classification of epidemiological data on hepatic steatosis'. In: *Proceedings of the IEEE Symposium on Computer-Based Medical Systems*. New York, New York, USA, 2014, pp. 1–7.

[156] Hoang, N. S. et al. 'Gait classification for Parkinson's Disease using Stacked 2D and 1D Convolutional Neural Network'. In: *Proceedings of the International Conference on Advanced Technologies for Communications*. Hanoi, Vietnam: IEEE Computer Society, Oct. 2019, pp. 44–49.

[157] Hochreiter, S. and Schmidhuber, J. 'Long Short-Term Memory'. In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[158] Holzinger, A. 'Interactive machine learning for health informatics: when do we need the human-in-the-loop?' In: *Brain Informatics* 3.2 (2016), pp. 119–131.

[159] Hübner, U. et al. *IT-Report Gesundheitswesen, Schwerpunkt – Wie reif ist die IT in deutschen Krankenhäusern?* Tech. rep. Osnabrück: Hochschule Osnabrück, 2018, pp. 1–98.

[160] Iguyon, I. and Elisseeff, A. 'An introduction to variable and feature selection'. In: *Journal of Machine Learning Research* 3 (2003), pp. 1157–1182.

[161] John, G. H., Kohavi, R. and Pfleger, K. 'Irrelevant Features and the Subset Selection Problem'. In: *Proceedings of the Eleventh International Machine Learning Conference*. New Brunswick, New Jersey, 1994, pp. 121–129.

[162] Jungen, D. et al. 'Cost-of-illness of psoriasis – results of a German cross-sectional study'. In: *Journal of the European Academy of Dermatology and Venereology* 32.1 (2018), pp. 174–180.

[163] Kahneman, D. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

[164] Kang, H. 'The prevention and handling of the missing data'. In: *Korean Journal of Anesthesiology* 64.5 (2013), pp. 402–406.

[165] Kaplan, R. M. and Frosch, D. L. 'Decision making in medicine and health care'. In: *Annual Review of Clinical Psychology* 1.1 (2005), pp. 525–556.

[166] Karlen, W., Mattiussi, C. and Floreano, D. 'Sleep and wake classification with ECG and respiratory effort signals'. In: *IEEE Transactions on Biomedical Circuits and Systems* 3.2 (2009), pp. 71–78.

[167] Karvounis, E. C. et al. 'A decision support system for the treatment of patients with ventricular assist device support'. In: *Methods of Information in Medicine* 53.2 (2014), pp. 121–136.

[168] Kawamoto, K. et al. 'Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success'. In: *BMJ* 330.765 (2005), pp. 1–8.

[169] Kim, D. and Yum, B. J. 'Collaborative filtering based on iterative principal component analysis'. In: *Expert Systems with Applications* 28.4 (2005), pp. 823–830.

[170] Kim, W. B., Jerome, D. and Yeung, J. 'Diagnosis and Management of Psoriasis'. In: *Canadian Family Physician* 63.4 (2017), pp. 278–285.

[171] Kira, K. and Rendell, L. A. 'A Practical Approach to Feature Selection'. In: *Proceedings of the International Conference on Machine Learning*. Aberdeen, Scottland, 1992, pp. 249–256.

[172] Klenk, S. et al. 'Determining patient similarity in medical social networks'. In: *Proceedings of the First International Workshop on Web Science and Information Exchange in the Medical Web, MedEx 2010*. Raleigh, North Carolina, USA, 2010, pp. 6–13.

[173] Knorr-Held, L. *Analysis of Incomplete Multivariate Data*. Boca Raton: CRC Press, 1997.

[174] Kohavi, R., Langley, P. and Yun, Y. 'The utility of feature weighting in nearest-neighbor algorithms'. In: *Proceedings of the Ninth European Conference on Machine Learning*. Prague, Czech Republic, 1997, pp. 85–92.

[175] Kohn, L. T., Corrigan, J. M. and Donaldson, M. S. *To Err is Human: Building a Safer Health System*. 1st ed. Washington, DC, USA: The National Academies Press, 2000.

[176] Komkhao, M., Lu, J. and Zhang, L. 'Determining Pattern Similarity in a Medical Recommender System'. In: *Proceedings of the International Conference on Data and Knowledge Engineering, ICDKE 2012*. Wuyishan, Fujian, Chinal, 2012, pp. 103–114.

[177] Kononenko, I., Šimec, E. and Robnik-Šikonja, M. 'Overcoming the myopia of inductive learning algorithms with RELIEFF'. In: *Applied Intelligence* 7.1 (1997), pp. 39–55.

[178] Konstan, J. A. et al. 'Applying Collaborative Filtering to Usenet News'. In: *Communications of the ACM* 40.3 (1997), pp. 77–87.

[179] Koren, Y. 'Collaborative filtering with temporal dynamics'. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '09*. Paris, France, 2009, pp. 447–456.

[180] Koren, Y. 'Factor in the neighbors: Scalable and accurate collaborative filtering'. In: *ACM Transactions on Knowledge Discovery from Data* 4.1 (2010), pp. 1–24.

[181] Koren, Y. 'Factorization meets the neighborhood: A multifaceted collaborative filtering model'. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'08*. Las Vegas, Nevada, USA, 2008, pp. 426–434.

[182] Koren, Y., Bell, R. and Volinsky, C. 'Matrix factorization techniques for recommender systems'. In: *Computer* 42.8 (2009), pp. 30–37.

[183] Koutkias, V. et al. 'Constructing Clinical Decision Support Systems for Adverse Drug Event Prevention: A Knowledge-based Approach'. In: *Proceedings of the AMIA Annual Symposium*. Washington, DC, USA, 2010, pp. 402–406.

[184] Krieger, J. et al. 'Breathing during sleep in normal middle-aged subjects'. In: *Sleep* 13.2 (1990), pp. 143–154.

[185] Kropf, M. et al. 'Evaluation of a Clinical Decision Support Rule-set for Medication Adjustments in mHealth-based Heart Failure Management'. In: *Studies in Health Technology and Informatics* 212 (2015), pp. 81–87.

[186] Krulwich, B. 'LIFESTYLE FINDER: Intelligent User Profiling Using Large-Scale Demographic Data'. In: *AI Magazine* 18.2 (1997), pp. 37–45.

[187] Kryger, M. H., Dement, W. C. and Roth, T. *Principles and practice of sleep medicine*. 6th ed. Elsevier, 2017.

[188] Kubben, P. et al. 'An evidence-based mobile decision support system for subaxial cervical spine injury treatment'. In: *Surgical Neurology International* 2.32 (2011), pp. 1–4.

[189] Kulis, B. 'Metric Learning: A Survey'. In: *Foundations and Trends® in Machine Learning* 5.4 (2013), pp. 287–364.

[190] Kuncheva, L. I. and Alpaydin, E. *Combining Pattern Classifiers: Methods and Algorithms*. 2nd ed. New York: John Wiley & Sons, 2014, p. 384.

[191] Landis, J. R. and Koch, G. G. 'The Measurement of Observer Agreement for Categorical Data'. In: *Biometrics* 33.1 (1977), pp. 159–174.

[192] Langley, R. G., Krueger, G. G. and Griffiths, C. E. 'Psoriasis: Epidemiology, clinical features, and quality of life'. In: *Annals of the Rheumatic Diseases* 64.2 (2005), pp. 18–23.

[193] Larkin, I. et al. 'Association between academic medical center pharmaceutical detailing policies and physician prescribing'. In: *JAMA - Journal of the American Medical Association* 317.17 (2017), pp. 1785–1795.

[194] Lattar, H. et al. 'Health Recommender Systems: A Survey'. In: *Smart Innovation, Systems and Technologies* 146 (2020), pp. 182–191.

[195] Lee, A. and Gilbert, R. M. 'Epidemiology of Parkinson Disease'. In: *Neurologic Clinics* 34.4 (2016), pp. 955–965.

[196] Lee, J., Maslove, D. M. and Dubin, J. A. 'Personalized mortality prediction driven by electronic medical data and a patient similarity metric'. In: *PLoS ONE* 10.5 (2015), pp. 1–13.

[197] Leeper, N. J. et al. 'Practice-Based Evidence: Profiling the Safety of Cilostazol by Text-Mining of Clinical Notes'. In: *PLoS ONE* 8.5 (2013), pp. 1–8.

[198] Levenson, R. M. et al. 'Pigeons (Columba livia) as trainable observers of pathology and radiology breast cancer images'. In: *PLoS ONE* 10.11 (2015), pp. 1–21.

[199] Li, L. et al. 'Identification of type 2 diabetes subgroups through topological analysis of patient similarity'. In: *Science Translational Medicine* 7.311 (2015), pp. 1–16.

[200] Li, Q. et al. 'Deep learning in the cross-time frequency domain for sleep staging from a single-lead electrocardiogram'. In: *Physiological Measurement* 39 (2018), pp. 1–12.

[201] Lin, H. C. et al. 'Development of a real-time clinical decision support system upon the web mvc-based architecture for prostate cancer treatment'. In: *BMC Medical Informatics and Decision Making* 11 (2011), pp. 1–11.

[202] Lin, J. H. and Haug, P. J. 'Exploiting missing clinical data in Bayesian network modeling for predicting medical problems'. In: *Journal of Biomedical Informatics* 41.1 (2008), pp. 1–14.

[203] Linden, G., Smith, B. and York, J. 'Amazon.com Recommendations: Item-to-Item Collaborative Filtering'. In: *IEEE Internet Computing* 7.1 (2003), pp. 76–80.

[204] Liu, H., Mei, J. and Xie, G. 'Towards collaborative chronic care using a clinical guideline-based decision support system'. In: *Studies in Health Technology and Informatics* 180.1 (2012), pp. 492–496.

[205] Liu, H. and Yu, L. 'Toward integrating feature selection algorithms for classification and clustering'. In: *IEEE Transactions on Knowledge and Data Engineering* 17.4 (2005), pp. 491–502.

[206] Liu, J. and Finkelstein, J. 'Introducing pharmacogenomic decision support for medication risk assessment in people with polypharmacy'. In: *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*. Kansas City, Missouri, USA, 2017, pp. 1803–1808.

[207] Loghmanpour, N. A., Druzdzel, M. J. and Antaki, J. F. 'Cardiac health risk stratification system (CHRiSS): A Bayesian-based decision support system for left ventricular assist device (LVAD) therapy'. In: *PLoS ONE* 9.11 (2014), pp. 1–10.

[208] Loke, P. Y., Chew, L. and Yap, C. W. 'Pilot study on developing a decision support tool for guiding re-administration of chemotherapeutic agent after a serious adverse drug reaction'. In: *BMC Cancer* 11 (2011), pp. 1–6.

[209] Longhurst, C. A., Harrington, R. A. and Shah, N. H. 'A 'green button' for using aggregate patient data at the point of care'. In: *Health Affairs* 33.7 (2014), pp. 1229–1235.

[210] Loughrey, J. and Cunningham, P. 'Overfitting in Wrapper-Based Feature Subset Selection: The Harder You Try the Worse it Gets'. In: *Research and Development in Intelligent Systems XXI* (2007), pp. 33–43.

[211] Lowsky, D. J. et al. 'A K-nearest neighbors survival probability prediction method'. In: *Statistics in Medicine* 32.12 (2013), pp. 2062–2069.

[212] Lu, X., Huang, Z. and Duan, H. 'Supporting adaptive clinical treatment processes through recommendations'. In: *Computer Methods and Programs in Biomedicine* 107.3 (2012), pp. 413–424.

[213] Makary, M. A. and Daniel, M. 'Medical error-the third leading cause of death in the US'. In: *BMJ (Online)* 353 (2016), pp. 1–5.

[214] Malik, J., Lo, Y. L. and Wu, H. T. 'Sleep-wake classification via quantifying heart rate variability by convolutional neural network'. In: *arXiv.org* (2018). URL: `https://arxiv.org/abs/1808.00142`.

[215] Malik, M. et al. 'Heart rate variability. Standards of measurement, physiological interpretation, and clinical use'. In: *European Heart Journal* 17.3 (1996), pp. 354–381.

[216] Manson, J. A. E. et al. 'Algorithm and mobile app for menopausal symptom management and hormonal/non-hormonal therapy decision making: A clinical decision-support tool from The North American Menopause Society'. In: *Menopause* 22.3 (2015), pp. 247–253.

[217] McEvoy, M. D. et al. 'A Smartphone-based decision support tool improves test performance concerning application of the guidelines for managing regional anesthesia in the patient receiving antithrombotic or thrombolytic therapy'. In: *Anesthesiology* 124.1 (2016), pp. 186–198.

[218] McIsaac, W. J., Moineddin, R. and Ross, S. 'Validation of a decision aid to assist physicians in reducing unnecessary antibiotic drug use for acute cystitis'. In: *Archives of Internal Medicine* 167.20 (2007), pp. 2201–2206.

[219] MediaJet. *File:An Arm Covered With Plaque Type Psoriasis.jpg*. URL: `https://commons.wikimedia.org/wiki/File:An%7B%5C_%7DArm%7B%5C_%7DCovered%7B%5C_%7DWith%7B%5C_%7DPlaque%7B%5C_%7DType%7B%5C_%7DPsoriasis.jpg` (visited on 23/03/2020).

[220] Mehrholz, J. 'Wissenschaft erklärt: Evidenzstufen – Studien nach ihrer Qualität einordnen'. In: *Ergopraxis* 3.6 (2010), p. 14.

[221] Mei, J. et al. 'Outcome-driven Evaluation Metrics for Treatment Recommendation Systems'. In: *Studies in Health Technology and Informatics* 210 (2015), pp. 190–194.

[222] Michel, M. et al. *Improving Patient Safety Using ATHENA-Decision Support System Technology: The Opioid Therapy for Chronic Pain Experience*. Rockville: Agency for Healthcare Research and Quality (US), 2008.

[223] Mikolov, T. et al. 'Efficient estimation of word representations in vector space'. In: *arXiv.org* (2013). URL: `http://arxiv.org/abs/1301.3781`.

[224] Miller, R. A., Pople, H. E. and Myers, J. D. 'Internist-I, an Experimental Computer-Based Diagnostic Consultant for General Internal Medicine'. In: *New England Journal of Medicine* 307.8 (1982), pp. 468–476.

[225] Mishra, A., Malviya, A. and Aggarwal, S. 'Towards Automatic Pharmacovigilance: Analysing Patient Reviews and Sentiment on Oncological Drugs'. In: *Proceedings of the 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015*. Atlantic City, New Jersey, USA, 2015, pp. 1402–1409.

[226] Mitchell, T. M. *Machine Learning*. 1st ed. New York: McGraw-Hill, 1997.

[227] Miyo, K. et al. 'Development of case-based medication alerting and recommender system: A new approach to prevention for medication error'. In: *Studies in Health Technology and Informatics* 129 (2007), pp. 871–874.

[228] Mohri, T. and Tanaka, H. *An optimal weighting criterion of case indexing for both numeric and symbolic attributes*. Tech. rep. 1994, pp. 123–127.

[229] Moja, L. et al. 'Effectiveness of computerized decision support systems linked to electronic health records: A systematic review and meta-analysis'. In: *American Journal of Public Health* 104.12 (2014), pp. 1–11.

[230] Molnar, C. *Interpretable Machine Learning*. 2020. URL: `https://christophm.github.io/interpretable-ml-book/` (visited on 27/09/2020).

[231] Morimoto, T. et al. 'Adverse durg events and medication errors: Detection and classification methods'. In: *Quality and Safety in Health Care* 13.4 (2004), pp. 306–314.

[232] Al-Moslmi, T. et al. 'Approaches to Cross-Domain Sentiment Analysis: A Systematic Literature Review'. In: *IEEE Access* 5 (2017), pp. 16173–16192.

[233] Mrowietz, U. et al. 'Definition of treatment goals for moderate to severe psoriasis: A European consensus'. In: *Archives of Dermatological Research* 303.1 (2011), pp. 1–10.

[234] Mulder, M. J. et al. 'Towards personalised intra-arterial treatment of patients with acute ischaemic stroke: A study protocol for development and validation of a clinical decision aid'. In: *BMJ Open* 7.3 (2017), pp. 1–5.

[235] Murali, N. S., Svatikova, A. and Somers, V. K. 'Cardiovascular physiology and sleep'. In: *Frontiers in Bioscience* 8 (2003), pp. 636–652.

[236] Mustaqeem, A., Anwar, S. M. and Majid, M. 'A modular cluster based collaborative recommender system for cardiac patients'. In: *Artificial Intelligence in Medicine* 102 (2020), pp. 1–12.

[237] Na, J.-C. and Kyaing, W. Y. M. 'Sentiment Analysis of User-Generated Content on Drug Review Websites'. In: *Journal of Information Science Theory and Practice* 3.1 (2015), pp. 6–23.

[238] Nachtigall, I. et al. 'Long-term effect of computer-assisted decision support for antibiotic treatment in critically ill patients: A prospective 'before/after' cohort study'. In: *BMJ Open* 4.12 (2014), pp. 1–10.

[239] Nast, A. et al. 'Low prescription rate for systemic treatments in the management of severe psoriasis vulgaris and psoriatic arthritis in dermatological practices in Berlin and Brandenburg, Germany: Results from a patient registry'. In: *Journal of the European Academy of Dermatology and Venereology* 22.11 (2008), pp. 1337–1342.

[240] Nast, A. et al. *S3 - Leitlinie zur Therapie der Psoriasis vulgaris Update 2017*. 2017. URL: `https://www.awmf.org/leitlinien/detail/ll/013-001.html` (visited on 27/09/2020).

[241] Nathan, A. J. and Scobell, A. 'Pattern classification with missing data: a review'. In: *Neural Computing and Applications* 19 (2009), pp. 263–282.

[242] Newgard, C. D. and Lewis, R. J. 'Missing data: How to best account for what is not known'. In: *JAMA - Journal of the American Medical Association* 314.9 (2015), pp. 940–941.

[243] Ng, K. et al. 'Personalized Predictive Modeling and Risk Factor Identification using Patient Similarity.' In: *Proceedings of the AMIA Joint Summits on Translational Science*. San Francisco, California, USA, 2015, pp. 132–136.

[244] Niehoff, K. M. et al. 'Development of the Tool to Reduce Inappropriate Medications (TRIM): A Clinical Decision Support System to Improve Medication Prescribing for Older Adults'. In: *Pharmacotherapy* 36.6 (2016), pp. 694–701.

[245] Nielsen, P. B. et al. 'Improvement of anticoagulant treatment using a dynamic decision support algorithm: A Danish Cohort study'. In: *Thrombosis Research* 133.3 (2014), pp. 375–379.

[246] Ning, X. and Karypis, G. 'SLIM: Sparse LInear Methods for top-N recommender systems'. In: *Proceedings of the IEEE International Conference on Data Mining, ICDM 2011*. Vancouver, Canada, 2011, pp. 497–506.

[247] Ning, X. and Karypis, G. 'Sparse Linear Methods with Side Information for top-N recommendations'. In: *Proceedings of the 21st Annual Conference on World Wide Web Companion, WWW '12*. Lyon, France, 2012, pp. 581–582.

[248] Nuckols, T. K. et al. 'The effectiveness of computerized order entry at reducing preventable adverse drug events and medication errors in hospital settings: A systematic review and meta-analysis'. In: *Systematic Reviews* 3 (2014), pp. 1–12.

[249] O'Mara-Eves, A. et al. 'Using text mining for study identification in systematic reviews: A systematic review of current approaches'. In: *Systematic Reviews* 4.5 (2015), pp. 1–22.

[250] Oluoch, T. et al. 'Effect of a clinical decision support system on early action on immunological treatment failure in patients with HIV in Kenya: A cluster randomised controlled trial'. In: *The Lancet HIV* 3.2 (2016), pp. 76–84.

[251] Osterloh, F. 'Ruf nach mehr Unabhängigkeit'. In: *Deutsches Ärzteblatt* 115.10 (2018), pp. 424–426.

[252] Owasirikul, W. et al. 'Prediction of shape diameter undergoing coil embolization of saccular intracranial aneurysm treatment using a hybrid decision support system'. In: *Australasian Physical and Engineering Sciences in Medicine* 36.2 (2013), pp. 177–191.

[253] Pai, S. and Bader, G. D. 'Patient Similarity Networks for Precision Medicine'. In: *Journal of Molecular Biology* 430.18 (2018), pp. 2924–2938.

[254] Panahiazar, M. et al. 'Using EHRs for Heart Failure Therapy Recommendation Using Multidimensional Patient Similarity Analytics'. In: *Studies in Health Technology and Informatics* 210 (2015), pp. 369–373.

[255] Pandey, B. and Mishra, R. B. 'Knowledge and intelligent computing system in medicine'. In: *Computers in Biology and Medicine* 39.3 (2009), pp. 215–230.

[256] Parimbelli, E. et al. 'Patient similarity for precision medicine: A systematic review'. In: *Journal of Biomedical Informatics* 83 (2018), pp. 87–96.

[257] Park, Y. J., Kim, B. C. and Chun, S. H. 'New knowledge extraction technique using probability for case-based reasoning: Application to medical diagnosis'. In: *Expert Systems* 23.1 (2006), pp. 2–20.

[258] Park, Y. J. and Tuzhilin, A. 'The long tail of recommender systems and how to leverage it'. In: *Proceedings of the 2008 ACM Conference on Recommender Systems, RecSys' 08*. Lausanne, Switzerland, 2008, pp. 11–18.

[259] Paterek, A. 'Improving regularized singular value decomposition for collaborative filtering'. In: *Proceedings of KDD Cup and Workshop*. San Jose, California, USA, 2007, pp. 39–42.

[260] Patzer, R. E. et al. 'IChoose kidney: A clinical decision aid for kidney transplantation versus dialysis treatment'. In: *Transplantation* 100.3 (2016), pp. 630–639.

[261] Pazzani, M. J. 'Framework for collaborative, content-based and demographic filtering'. In: *Artificial Intelligence Review* 13.5 (1999), pp. 393–408.

[262] Pedersen, A. B. et al. 'Missing data and multiple imputation in clinical epidemiological research'. In: *Clinical Epidemiology* 9 (2017), pp. 157–166.

[263] Pedreira, C. E. et al. 'New decision support tool for treatment intensity choice in childhood acute lymphoblastic leukemia'. In: *IEEE Transactions on Information Technology in Biomedicine* 13.3 (2009), pp. 284–290.

[264] Peng, H., Long, F. and Ding, C. 'Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.8 (2005), pp. 1226–1238.

[265] Pennington, J., Socher, R. and Manning, C. D. 'GloVe: Global vectors for word representation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*. Doha, Qatar, 2014, pp. 1532–1543.

[266] Penny, K. I. and Chesney, T. 'Imputation methods to deal with missing values when data mining trauma injury data'. In: *Proceedings of the International Conference on Information Technology Interfaces, ITI 2006*. Cavtat/Dubrovnik, Croatia, 2006, pp. 213–218.

[267] Penzel, T. et al. 'Modulations of heart rate, ECG, and cardio-respiratory coupling observed in polysomnography'. In: *Frontiers in Physiology* 7 (2016), pp. 1–15.

[268] Persson, M. et al. 'Evaluation of a computer-based decision support system for treatment of hypertension with drugs: Retrospective, nonintervention testing of cost and guideline adherence'. In: *Journal of Internal Medicine* 247.1 (2000), pp. 87–93.

[269] Pevnick, J. M. et al. 'Effect of electronic prescribing with formulary decision support on medication tier, copayments, and adherence'. In: *BMC Medical Informatics and Decision Making* 14 (2014), pp. 1–12.

[270] Plumb, G., Molitor, D. and Talwalkar, A. 'Model agnostic supervised local explanations'. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*. Montreal, Canada, 2018, pp. 2520–2529.

[271] Polikar, R. 'Ensemble based systems in decision making'. In: *IEEE Circuits and Systems Magazine* 6.3 (2006), pp. 21–44.

[272] Pruszydlo, M. G. et al. 'Development and evaluation of a computerised clinical decision support system for switching drugs at the interface between primary and tertiary care'. In: *BMC Medical Informatics and Decision Making* 12 (2012), pp. 1–8.

[273] Pudil, P. et al. 'Floating search methods for feature selection with nonmonotonic criterion functions'. In: *Pattern Recognition* 2 (2002), pp. 279–283.

[274] Qian, B. et al. ‘A relative similarity based method for interactive patient risk prediction’. In: *Data Mining and Knowledge Discovery* 29.4 (2015), pp. 1070–1093.

[275] Quinlan, J. R. ‘Induction of decision trees’. In: *Machine Learning* 1.1 (1986), pp. 81–106.

[276] Quinlan, J. R. *C4.5: Programs for Machine Learning*. 1st ed. San Francisco: Morgan Kaufmann Publishers, 1993, p. 302.

[277] Rabiner, L. R. ‘A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition’. In: *Proceedings of the IEEE* 77.2 (1989), pp. 257–286.

[278] Radha, M. et al. ‘Sleep stage classification from heart-rate variability using long short-term memory neural networks’. In: *Scientific Reports* 9 (2019), pp. 1–11.

[279] Raghavan, S. R., Ladik, V. and Meyer, K. B. ‘Developing decision support for dialysis treatment of chronic kidney failure’. In: *IEEE Transactions on Information Technology in Biomedicine* 9.2 (2005), pp. 229–238.

[280] Raschka, S. ‘Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning’. In: *arXiv.org* (2018). URL: `https://arxiv.org/abs/1811.12808`.

[281] Ratzel, R. et al. ‘(Muster-)Berufsordnung für die in Deutschland tätigen Ärztinnen und Ärzte’. In: *Deutsches Ärzteblatt* (2019), pp. 1–9.

[282] Redmond, S. J. et al. ‘Sleep staging using cardiorespiratory signals’. In: *Somnologie* 11.4 (2007), pp. 245–256.

[283] Reichmann, H. ‘Modern treatment in Parkinson’s disease, a personal approach’. In: *Journal of Neural Transmission* 123.1 (2016), pp. 73–80.

[284] Resnick, P. et al. ‘GroupLens: An open architecture for collaborative filtering of netnews’. In: *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, CSCW 1994*. Chapel Hill, North Carolina, USA, 1994, pp. 175–186.

[285] Reunanen, J. ‘Overfitting in feature selection: Pitfalls and solutions’. PhD thesis. Aalto University, Helsinki, Finland, 2012.

[286] Reunanen, J. ‘Overfitting in Making ComparisonsBetween Variable Selection Method’. In: *Journal of Machine Learning Research* 3 (2003), pp. 1371–1382.

[287] Ribeiro, M. T., Singh, S. and Guestrin, C. ‘"Why should i trust you?" Explaining the predictions of any classifier’. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA, 2016, pp. 1135–1144.

[288] Ricci, F. et al. *Recommender Systems Handbook*. 2nd ed. New York: Springer, 2015.

[289] Richards, H. L. et al. ‘Patients with psoriasis and their compliance with medication’. In: *Journal of the American Academy of Dermatology* 41.4 (1999), pp. 581–583.

[290] Rodriguez-Maresca, M. et al. ‘Implementation of a computerized decision support system to improve the appropriateness of antibiotic therapy using local microbiologic data’. In: *BioMed Research International* 2014 (2014), pp. 1–9.

[291] Rokach, L. and Maimom, O. *Data Mining and Knowledge Discovery Handbook*. 2nd ed. New York: Springer, 2010.

[292] Rollman, B. L. et al. 'A randomized trial using computerized decision support to improve treatment of major depression in primary care'. In: *Journal of General Internal Medicine* 17.7 (2002), pp. 493–503.

[293] Rubin, D. B. 'Inference and missing data'. In: *Biometrika* 63.3 (1976), pp. 581–592.

[294] Rubin, D. B. *Multiple Imputation for Nonresponse in Surveys*. New Jersey: John Wiley & Sons, 1986.

[295] Russell, S. J. and Norvig, P. *Künstliche Intelligenz: ein moderner Ansatz*. 3rd ed. München: Pearson, Higher Education, 2012.

[296] Sackett, D. L. et al. 'Evidence based medicine: what it is and what it isn't'. In: *BMJ* 312 (1996), pp. 71–72.

[297] Salas-Zárate, M. D. P. et al. 'Sentiment Analysis on Tweets about Diabetes: An Aspect-Level Approach'. In: *Computational and Mathematical Methods in Medicine* 2017 (2017), pp. 1–10.

[298] Sarwar, B., Karypis, G. and Konstan, J. 'Item-Based Collaborative Filtering Recommendation'. In: *Proceedings of the 10th international conference on World Wide Web, WWW '01*. Hong Kong, 2001, pp. 285–295.

[299] Sarwar, B. et al. 'Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems'. In: *Proceedings of the Fifth International Conference on Computer and Information Science, ICIS 2002*. Seoul, Korea, 2002, pp. 27–28.

[300] Sarwar, B. et al. 'Application of Dimensionality Reduction in Recommender System'. In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '00, WebKDD-2000 Workshop*. Boston, Massachusetts, USA, 2000, pp. 1–12.

[301] Schäfer, H. et al. 'Towards health (Aware) recommender systems'. In: *Proceedings of the 2017 International Conference on Digital Health*. London, England, 2017, pp. 157–161.

[302] Schäfer, I. et al. 'Epidemiologie der Psoriasis in Deutschland - Auswertung von Sekundärdaten einer gesetzlichen Krankenversicherung'. In: *Gesundheitswesen* 73.5 (2011), pp. 308–313.

[303] Schapire, R. E. 'The Strength of Weak Learnability'. In: *Machine Learning* 5.2 (1990), pp. 197–227.

[304] Scheitel, M. R. et al. 'Effect of a novel clinical decision support tool on the efficiency and accuracy of treatment recommendations for cholesterol management'. In: *Applied Clinical Informatics* 8.1 (2017), pp. 124–136.

[305] Scherf, M. and Brauer, W. *Feature Selection by Means of a Feature Weighting Approach (Technical Report No. FKI22197)*. Tech. rep. 1997, pp. 1–22.

[306] Schmitt, J. et al. 'Efficacy and safety of systemic treatments for moderate-to-severe psoriasis: Meta-analysis of randomized controlled trials'. In: *British Journal of Dermatology* 170.2 (2014), pp. 274–303.

[307] Schneeweiss, S. 'Learning from big health care data'. In: *New England Journal of Medicine* 370.23 (2014), pp. 2161–2163.

[308] Schnurrer, J. U. and Frölich, J. C. 'Zur Häufigkeit und Vermeidbarkeit von tödlichen unerwünschten Arzneimittelwirkungen'. In: *Internist* 44.7 (2003), pp. 889–895.

[309] Schurink, C. A. et al. 'Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units'. In: *Lancet Infectious Diseases* 5.5 (2005), pp. 305–312.

[310] Sezgin, E. and Özkan, S. 'A systematic literature review on Health Recommender Systems'. In: *Proceedings of the 2013 E-Health and Bioengineering Conference, EHB 2013*. Iasi, Romania, 2013, pp. 1–4.

[311] Shaffer, F. and Ginsberg, J. P. 'An Overview of Heart Rate Variability Metrics and Norms'. In: *Frontiers in Public Health* 5 (2017), pp. 1–17.

[312] Sharafoddini, A., Dubin, J. A. and Lee, J. 'Patient Similarity in Prediction Models Based on Health Data: A Scoping Review'. In: *JMIR Medical Informatics* 5.1 (2017), pp. 1–17.

[313] Sharafoddini, A. et al. 'A new insight into missing data in intensive care unit patient profiles: Observational study'. In: *Journal of Medical Internet Research* 7.1 (2019), pp. 1–19.

[314] Shardanand, U. and Maes, P. 'Social information filtering'. In: *Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '95*. Denver, Colorado, USA, 1995, pp. 210–217.

[315] Shekelle, P. G. et al. 'Validity of the agency for healthcare research and quality clinical practice guidelines: How quickly do guidelines become outdated?' In: *Journal of the American Medical Association* 286.12 (2001), pp. 1461–1467.

[316] Shivade, C. et al. 'A review of approaches to identifying patient phenotype cohorts using electronic health records'. In: *Journal of the American Medical Informatics Association* 21.2 (2014), pp. 221–230.

[317] Shortliffe, E. *Computer-based medical consultations: MYCIN*. 1st ed. New York: Elsevier, 1976.

[318] Shortliffe, E. H. 'Computer Programs to Support Clinical Decision Making'. In: *JAMA: The Journal of the American Medical Association* 258.1 (1987), pp. 61–66.

[319] Shrivastava, V. K. et al. 'First review on psoriasis severity risk stratification: An engineering perspective'. In: *Computers in Biology and Medicine* 63 (2015), pp. 52–63.

[320] Sill, J. et al. 'Feature-Weighted Linear Stacking'. In: *arXiv.org* (2009). URL: `http://arxiv.org/abs/0911.0460`.

[321] Sim, I. E. A. 'Clinical Decision Support Systems for the Practice of Evidence-Based Medicine'. In: *JAMIA: Journal of the American Medical Informatics Association* 8.6.6 (2001), pp. 527–534.

[322] Singhi, S. K. and Liu, H. 'Feature subset selection bias for classification learning'. In: *Proceedings of the 23rd international conference on Machine learning, ICML '06*. Pittsburgh, Pennsylvania, USA, 2006, pp. 849–856.

[323] Skevofilakas, M. T. et al. 'A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines'. In: *Proceedings of the 27th Annual International Conference of the IEEE Engineering in Medicine and Biology, EMBC 2005*. Shanghai, China, 2005, pp. 2429–2432.

[324] Smith, W. P. et al. 'A decision aid for intensity-modulated radiation-therapy plan selection in prostate cancer based on a prognostic Bayesian network and a Markov model'. In: *Artificial Intelligence in Medicine* 46.2 (2009), pp. 119–130.

[325] Sodsee, S. and Komkhao, M. 'Evidence-based Medical Recommender Systems : A Review'. In: 4.September (2013), pp. 114–120.

[326] Somol, P. et al. 'Floating search methods in feature selection'. In: *Pattern Recognition Letters* 15.11 (1994), pp. 1119–1125.

[327] Somol, P., Novovicova, J. and Pudil, P. 'Efficient Feature Subset Selection and Subset Size Optimization'. In: *Pattern Recognition Recent Advances*. 1st ed. IntechOpen, 2010. Chap. 4, pp. 75–98.

[328] Sönnichsen, A. et al. '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'. In: *Trials* 17 (2016), pp. 1–9.

[329] Stern, R. S. et al. 'Psoriasis is common, carries a substantial burden even when not extensive, and is associated with widespread treatment dissatisfaction'. In: *Journal of Investigative Dermatology Symposium Proceedings* 9.2 (2004), pp. 136–139.

[330] Strub, F., Mary, J. and Gaudel, R. 'Hybrid Collaborative Filtering with Autoencoders'. In: *arXiv.org* (2016). URL: `https://arxiv.org/abs/1603.00806`.

[331] Stuck, B. A. et al. *Praxis der Schlafmedizin*. 3rd ed. Berlin Heidelberg: Springer, 2018.

[332] Sture Holm. 'A Simple Sequentially Rejective Multiple Test Procedure'. In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70.

[333] Su, X. and Khoshgoftaar, T. M. 'A Survey of Collaborative Filtering Techniques'. In: *Advances in Artificial Intelligence* 2009 (2009), pp. 1–19.

[334] Suebnukarn, S., Rungcharoenporn, N. and Sangsuratham, S. 'A Bayesian decision support model for assessment of endodontic treatment outcome'. In: *Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology and Endodontology* 106.3 (2008), e48–e58.

[335] Sun, H. et al. 'Sleep staging from electrocardiography and respiration with deep learning'. In: *Sleep* 43.7 (July 2020), pp. 1–26.

[336] Sun, J. et al. 'A system for mining temporal physiological data streams for advanced prognostic decision support'. In: *Proceedings of the IEEE International Conference on Data Mining, ICDM '10*. Sydney, Australia, 2010, pp. 1061–1066.

[337] Sun, J. et al. 'Localized supervised metric learning on temporal physiological data'. In: *Proceedings of the International Conference on Pattern Recognition*. Istanbul, Turkey, 2010, pp. 4149–4152.

[338] Sun, J. et al. 'Supervised patient similarity measure of heterogeneous patient records'. In: *ACM SIGKDD Explorations Newsletter* 14.1 (2012), pp. 16–24.

[339] Sun, Y. 'Iterative RELIEF for feature weighting: Algorithms, theories, and applications'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (2007), pp. 1035–1051.

[340] Suner, A. et al. 'Correctreatment: A web-based decision support tool for rectal cancer treatment that uses the analytic hierarchy process and decision tree'. In: *Applied Clinical Informatics* 6.1 (2015), pp. 56–74.

[341] Suo, Q. et al. 'Deep patient similarity learning for personalized healthcare'. In: *IEEE Transactions on Nanobioscience* 17.3 (2018), pp. 219–227.

[342] Sutton, A. J. 'Publication bias'. In: *The handbook of research synthesis and meta-analysis*. 2nd ed. New York: Russell Sage Foundation, 2009, pp. 435–452.

[343] Sutton, R. T. et al. 'An overview of clinical decision support systems: benefits, risks, and strategies for success'. In: *npj Digital Medicine* 3.17 (2020), pp. 1–10.

[344] Syed, M. E. 'Attribute weighting in K-nearest neighbor classification'. PhD thesis. University of Tampere, 2014, pp. 1–55.

[345] Symeonidis, P. and Zioupos, A. *Matrix and Tensor Factorization Techniques for Recommender Systems*. 1st ed. New York: Springer, 2016.

[346] T.G., H. 'An efficient random decision tree algorithm for case-based reasoning systems'. In: *Proceedings of the 24th International Florida Artificial Intelligence Research Society, FLAIRS - 24*. Florida, California, USA, 2011, pp. 401–406.

[347] Takács, G. et al. 'Scalable Collaborative Filtering Approaches for Large Recommender Systems'. In: *Journal of Machine Learning Research* 10 (2009), pp. 623–656.

[348] Tan, S. et al. 'Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation'. In: *arXiv.org* (2018). URL: `https://arxiv.org/abs/1710.06169`.

[349] Tang, J., Alelyani, S. and Liu, H. 'Feature Selection for Classification: A Review'. In: *Data Classification*. 1st ed. New York: Chapman and Hall/CRC, 2014, pp. 37–64.

[350] Tataraidze, A. et al. 'Estimation of a priori probabilities of sleep stages: A cycle-based approach'. In: *Proceedings of the 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2017*. Jeju Island, Korea, 2017, pp. 3745–3748.

[351] The Lancet Oncology. 'Clinical decision making: more than just an algorithm'. In: *The Lancet Oncology* 18.12 (2017), p. 1.

[352] Thomson, P. et al. 'A computerised guidance tree (decision aid) for hypertension, based on decision analysis: Development and preliminary evaluation'. In: *European Journal of Cardiovascular Nursing* 5.2 (2006), pp. 146–149.

[353] Tobaldini, E. et al. 'Heart rate variability in normal and pathological sleep'. In: *Frontiers in Physiology* 4 (2013), pp. 1–11.

[354] Trafton, J. A. et al. 'Designing an automated clinical decision support system to match clinical practice guidelines for opioid therapy for chronic pain'. In: *Implementation Science* 5.1 (2010), p. 26.

[355] Tran, T. N. T. et al. 'Recommender systems in the healthcare domain: state-of-the-art and research issues'. In: *Journal of Intelligent Information Systems* (2020), pp. 1573–7675.

[356] Tranchevent, L. C. et al. 'Predicting clinical outcome of neuroblastoma patients using an integrative network-based approach'. In: *Biology Direct* 13.1 (2018), pp. 1–13.

[357] Traupe, H. and Robra, B.-P. *Themenheft 11 "Schuppenflechte"*. Tech. rep. Robert Koch-Institut, 2007, pp. 1–10.

[358] Trimble, M. and Hamilton, P. 'The thinking doctor: Clinical Decision making in contemporary medicine'. In: *Clinical Medicine, Journal of the Royal College of Physicians of London* 16.4 (2016), pp. 343–346.

[359] Troyanskaya, O. et al. 'Missing value estimation methods for DNA microarrays'. In: *Bioinformatics* 17.6 (2001), pp. 520–525.

[360] Tsafnat, G. et al. 'The automation of systematic reviews'. In: *BMJ (Online)* 345 (2013), pp. 1–2.

[361] Tversky, A. and Kahneman, D. 'Judgment under uncertainty: Heuristics and biases'. In: *Judgment under Uncertainty* (2013), pp. 3–20.

[362] *UPDRS (Unified Parkinson's Disease Rating Scale)*. URL: `https://neurologienetz.de/fachliches/skalen-scores/updrs-unified-parkinsons-disease-rating-scale` (visited on 12/08/2020).

[363] Urbanowicz, R. J. et al. 'Relief-based feature selection: Introduction and review'. In: *Journal of Biomedical Informatics* 85 (2018), pp. 189–203.

[364] Varma, S. and Simon, R. 'Bias in error estimation when using cross-validation for model selection'. In: *BMC Bioinformatics* 7 (2006), pp. 1–8.

[365] Vert, J. P., Tsuda, K. and Schölkopf, B. 'A Primer on Kernel Methods'. In: *Kernel Methods in Computational Biology*. 1st ed. Cambridge, Massachusetts, USA: MIT Press, 2004. Chap. 1, pp. 35–70.

[366] Very, a. N. J. a. et al. 'Medication-related Clinical Decision Support in Computerized Provider Order Entry Systems : A Review'. In: *Journal of the American Medical Informatics Association* 14.1 (2007), pp. 29–40.

[367] Viterbi, A. J. 'Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm'. In: *IEEE Transactions on Information Theory* 13.2 (1967), pp. 260–269.

[368] Vranas, K. C. et al. 'Identifying distinct subgroups of ICU patients: A machine learning approach'. In: *Critical Care Medicine* 45.10 (2017), pp. 1607–1615.

[369] Wang, B. et al. 'Similarity network fusion for aggregating data types on a genomic scale'. In: *Nature Methods* 11.3 (2014), pp. 333–337.

[370] Wang, F. 'Data Analytics with Electronic Health Records'. In: *AAAI Conference on Artificial Intelligence* (2015).

[371] Wang, F., Hu, J. and Sun, J. 'Medical prognosis based on patient similarity and expert feedback'. In: *Proceedings of the International Conference on Pattern Recognition*. Tsukuba, Japan, 2012, pp. 1799–1802.

[372] Wang, F. and Sun, J. 'PSF: A unified Patient similarity evaluation framework through metric learning with weak supervision'. In: *IEEE Journal of Biomedical and Health Informatics* 19.3 (2015), pp. 1053–1060.

[373] Wang, F., Sun, J. and Ebadollahi, S. 'Integrating distance metrics learned from multiple experts and its application in patient similarity assessment'. In: *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*. Phoenix, Arizona, USA, 2011, pp. 59–70.

[374] Wang, F. et al. 'IMet: Interactive metric learning in healthcare applications'. In: *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*. Phoenix, Arizona, USA, 2011, pp. 944–955.

[375] Wang, H. et al. 'Integrating Omics Data with a Multiplex Network-Based Approach for the Identification of Cancer Subtypes'. In: *IEEE Transactions on Nanobioscience* 15.4 (2016), pp. 335–342.

[376] Wang, Y. et al. 'An Electronic Medical Record System with Treatment Recommendations Based on Patient Similarity'. In: *Journal of Medical Systems* 39.5 (2015), pp. 1–9.

[377] Weinberger, K. Q. and Saul, L. K. 'Distance metric learning for large margin nearest neighbor classification'. In: *Journal of Machine Learning Research* 10 (2009), pp. 207–244.

[378] Wettschereck, D. and Aha, D. W. 'Weighting features'. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 1010 (1995), pp. 347–358.

[379] Wettschereck, D., Aha, D. W. and Mohri, T. 'A Review and Empirical Evaluation of Feature Weighting Methods for a Class of Lazy Learning Algorithms'. In: *Artificial Intelligence Review* 11.1-5 (1997), pp. 273–314.

[380] Wicks, P. et al. 'Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm'. In: *Nature Biotechnology* 29.5 (May 2011), pp. 411–414.

[381] Wiesner, M. and Pfeifer, D. 'Health recommender systems: Concepts, requirements, technical basics and challenges'. In: *International Journal of Environmental Research and Public Health* 11.3 (2014), pp. 2580–2607.

[382] Wilcoxon, F. 'Individual Comparisons by Ranking Methods'. In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83.

[383] Wilson, D. R. and Martinez, T. R. 'Improved heterogeneous distance functions'. In: *Journal of Artificial Intelligence Research* 6 (1997), pp. 1–34.

[384] Witten, I. H. et al. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Boston, Massachusetts, USA: Morgan Kaufmann Publishers, 2011.

[385] Wolfstadt, J. I. et al. 'The effect of computerized physician order entry with clinical decision support on the rates of adverse drug events: A systematic review'. In: *Journal of General Internal Medicine* 23.4 (2008), pp. 451–458.

[386] Wright, A. et al. 'Development and evaluation of a comprehensive clinical decision support taxonomy: Comparison of front-end tools in commercial and internally developed electronic health record systems'. In: *Journal of the American Medical Informatics Association* 18.3 (2011), pp. 232–242.

[387] Xing, E. P. et al. 'Distance Metric Learning with Application to Clustering with Side-Information'. In: *Proceedings of the 15th International Conference on Neural Information Processing Systems, NIPS'02*. Vancouver, Canada, 2003, pp. 521–528.

[388] Xu, R., Nettleton, D. and Nordman, D. J. 'Case-Specific Random Forests'. In: *Journal of Computational and Graphical Statistics* 25.1 (2016), pp. 49–65.

[389] Yang, J. et al. 'Sleep stage recognition using respiration signal'. In: *Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2016*. Orlando, Florida, USA, 2016, pp. 2843–2846.

[390] Yu, L. and Liu, H. 'Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution'. In: *Proceedings, Twentieth International Conference on Machine Learning*. Washington, DC, USA, 2003, pp. 856–863.

[391] Yurtkuran, A., Tok, M. and Emel, E. 'A clinical decision support system for femoral peripheral arterial disease treatment'. In: *Computational and Mathematical Methods in Medicine* 2013 (2013), pp. 1–9.

[392] Zamora, A. et al. 'Theoretical Impact on Coronary Disease of Using a Computerized Clinical Decision Support System in the Prescription of Lipid-lowering Treatment'. In: *Revista Española de Cardiología (English Edition)* 68.1 (2015), pp. 75–78.

[393] Zeng, W. et al. 'Parkinson's disease classification using gait analysis via deterministic learning'. In: *Neuroscience Letters* 633 (2016), pp. 268–278.

[394] Zeng, X. and Martinez, T. R. 'Using a neural network to approximate an ensemble of classifiers'. In: *Neural Processing Letters* 12.3 (2000), pp. 225–237.

[395] Zhang, P. et al. 'Towards personalized medicine: leveraging patient similarity and drug similarity analytics.' In: *Proceedings of the AMIA Joint Summits on Translational Science*. San Francisco, California, USA, 2014, pp. 132–136.

[396] Zhang, Q. et al. 'A framework of hybrid recommender system for personalized clinical prescription'. In: *Proceedings of the The 10th International Conference on Intelligent Systems and Knowledge Engineering, ISKE '15*. Taipei, Taiwan, 2015, pp. 189–195.

[397] Zhang, S. et al. 'Learning from incomplete ratings using non-negative matrix factorization'. In: *Proceedings of the Sixth SIAM International Conference on Data Mining*. Washington, DC, USA, 2006, pp. 549–553.

[398] Zhang, W. et al. 'Predicting potential side effects of drugs by recommender methods and ensemble learning'. In: *Neurocomputing* 173 (2016), pp. 979–987.

Appendix A - Literature Review

Table A.1: Results from a systematic literature review including studies on treatment, therapy, medication or drug decision support or recommender systems. The identified publications were analyzed regarding algorithm (GB - Guideline-based, RB - Rule-based, PB - Probabilistic, DT - Decision Tree, LM - Linear Model, ANN - Artificial Neural Network, CB - Case-based), application, data source, and type of evaluation.

Ref.	**Algor.**	**Application**	**Data**	**Outcome**
[391]	ANN	Femoral periphal arterial disease	-	Matching therapy
[263]	ANN	Acute lymphoblastic leukemia	EHR	Treatment outcome
[309]	BN	Ventilator-associated pneumonia	-	Matching therapy
[334]	BN	Dental treatment	-	Matching therapy
[324]	BN	Cancer	-	Treatment outcome
[84]	BN	Artrial fibrillation	-	no evaluation described
[154]	BN	Diabetes	-	Matching therapy
[208]	BN	Chemotherapy/ADE	-	Matching therapy
[207]	BN	Heart failure	-	Treatment outcome
[88]	CB	Dose planning	EHR	Matching therapy
[323]	DT	Cancer	EHR	no evaluation described
[352]	DT	Hypertension, Benign prostatic hyperplasia	-	Usability
[188]	DT	Subaxial cervical spine injury	-	no evaluation described
[216]	DT	Menopausal treatments	-	no evaluation described
[272]	DT	Medication safety	-	Matching therapy
[167]	DT	Heart failure	EHR	Matching therapy
[340]	DT	Cancer	-	Matching therapy
[304]	GB	Cardivascular disease risk	EHR	Guideline adherence
[101]	GB	Artherosclerotic risk	EHR	no evaluation described
[217]	GB	Regional anesthesia	-	Matching therapy
[130]	GB	Cardiac rehabilitation	EHR	no evaluation described
[98]	GB	Cancer	EHR	Matching therapy
[222]	GB	Chronic pain	EHR	Study not done yet
[290]	GB	Nosocomial infection	LIMS	Treatment outcome
[292]	GB	Depressive disorder	EHR	Treatment outcome
[58]	GB	Hyperlipidemia	-	Treatment outcome
[268]	GB	Hypertension	-	Cost reduction

Ref.	Algor.	Application	Data	Outcome
[80]	GB	Tuberculosis	-	Guideline adherence
[354]	GB	Chronic pain	EHR	no evaluation described
[250]	GB	HIV	EHR	Matching therapy
[392]	GB	Coronary artery disease	-	Treatment outcome
[238]	GB	Medication safety	EHR	Guideline adherence
[8]	LM	Radiotherapy treatment	-	Matching therapy
[245]	LM	Medication safety	-	Matching therapy
[252]	LM	Aneurysm	-	Matching therapy
[137]	LM	Posttraumatic stress disorder	-	no evaluation described
[234]	LM	Acute ischaemic stroke	-	Study not done yet
[260]	LM	Reduced renal function	-	Treatment outcome
[201]	LM/DT	Prostate cancer	EHR	Matching therapy
[31]	RB	ADE	-	Contradicting recommendations
[218]	RB	Cystitis	-	Matching therapy
[185]	RB	Heart failure	EHR	Matching therapy
[269]	RB	Medication safety	-	Cost reduction
[64]	RB	Warfarin treatment	-	no evaluation described
[328]	RB	Medication safety	EHR	Study not done yet
[32]	RB	Medication safety	-	Contradicting recommendations
[183]	RB	ADE	EHR	Matching therapy
[279]	RB	Chronic kidney failure	EHR	no evaluation described
[144]	RB	Osteoporosis	-	Matching therapy
[268]	RB	Hypertension	EHR	Cost reduction
[11]	RB	Diabetes	-	no evaluation described
[244]	RB	Medication safety	EHR	Contradicting recommendations

Appendix B - Data

B.1 Treatment Describing Attributes

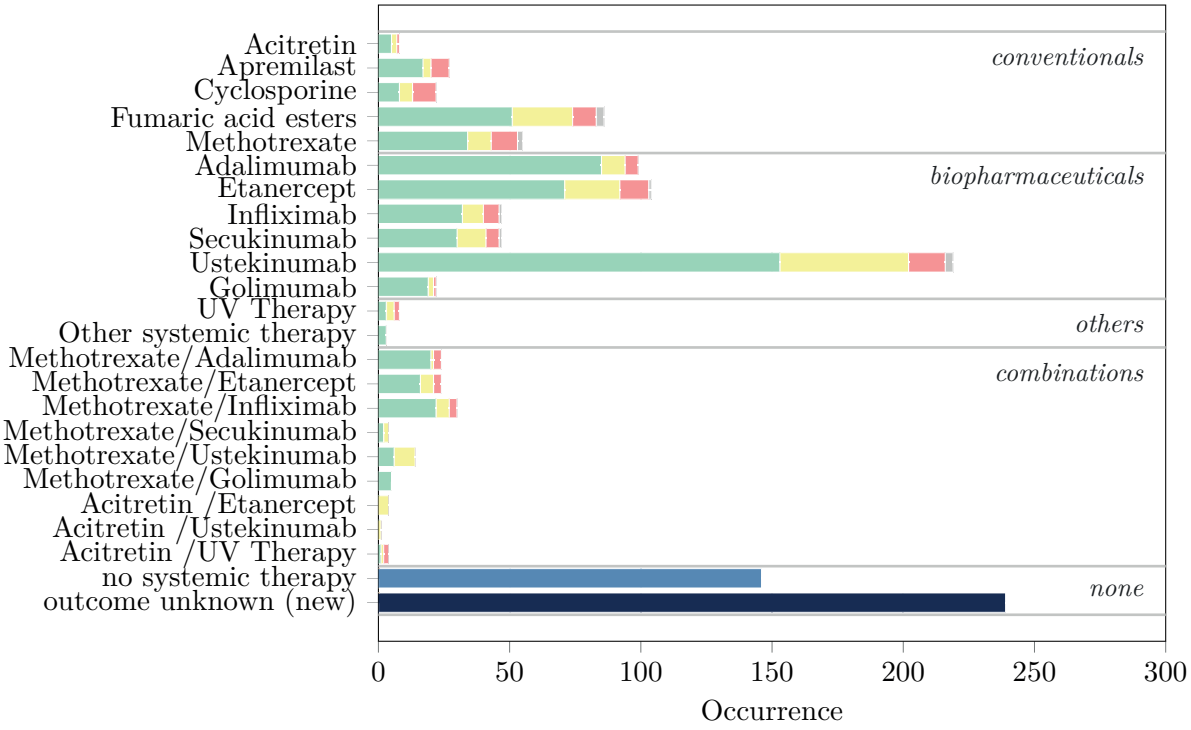


Figure B.1: *Effectiveness* associated with applied therapies classified into good (—), moderate (—), bad (—), or with missing outcome (—).

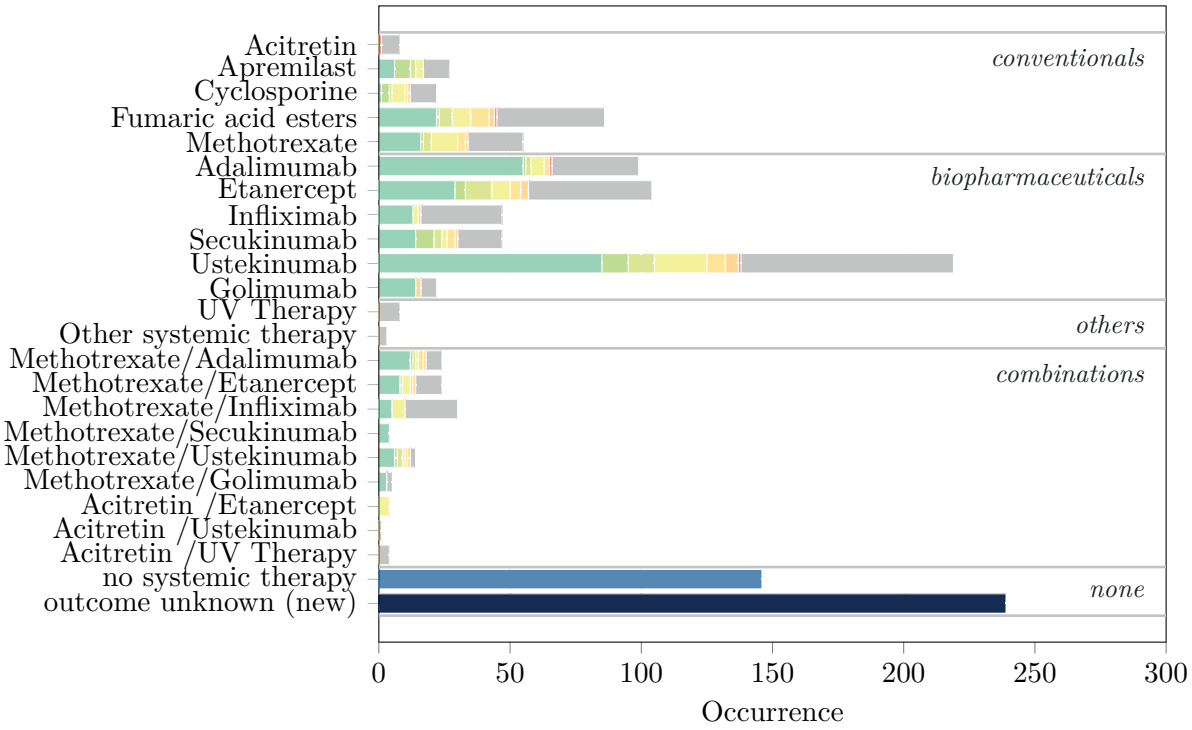


Figure B.2: $\Delta PASI_{rel}$ associated with applied therapies. $\Delta PASI_{rel}$ values range from PASI improvement or controlling the disease (—) to deterioration of the PASI (—), or with missing outcome (—).

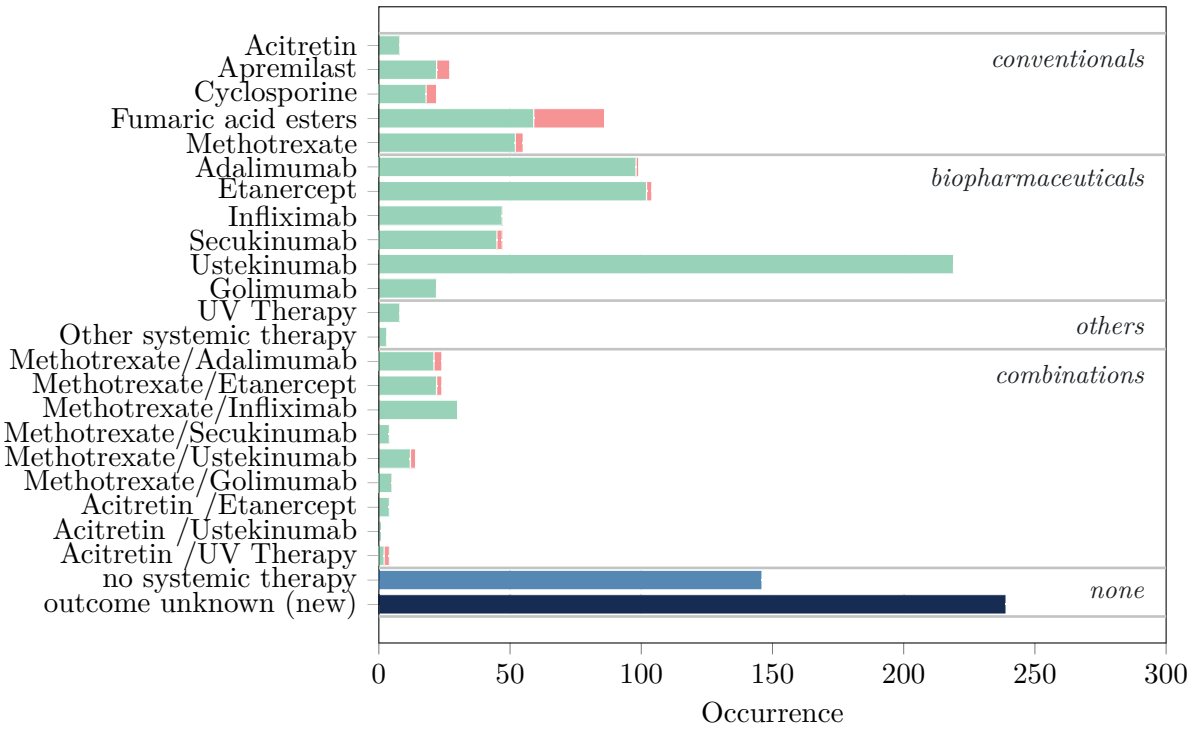


Figure B.3: Observed occurrence (—) and absence (—) of ADEs associated with applied therapies.

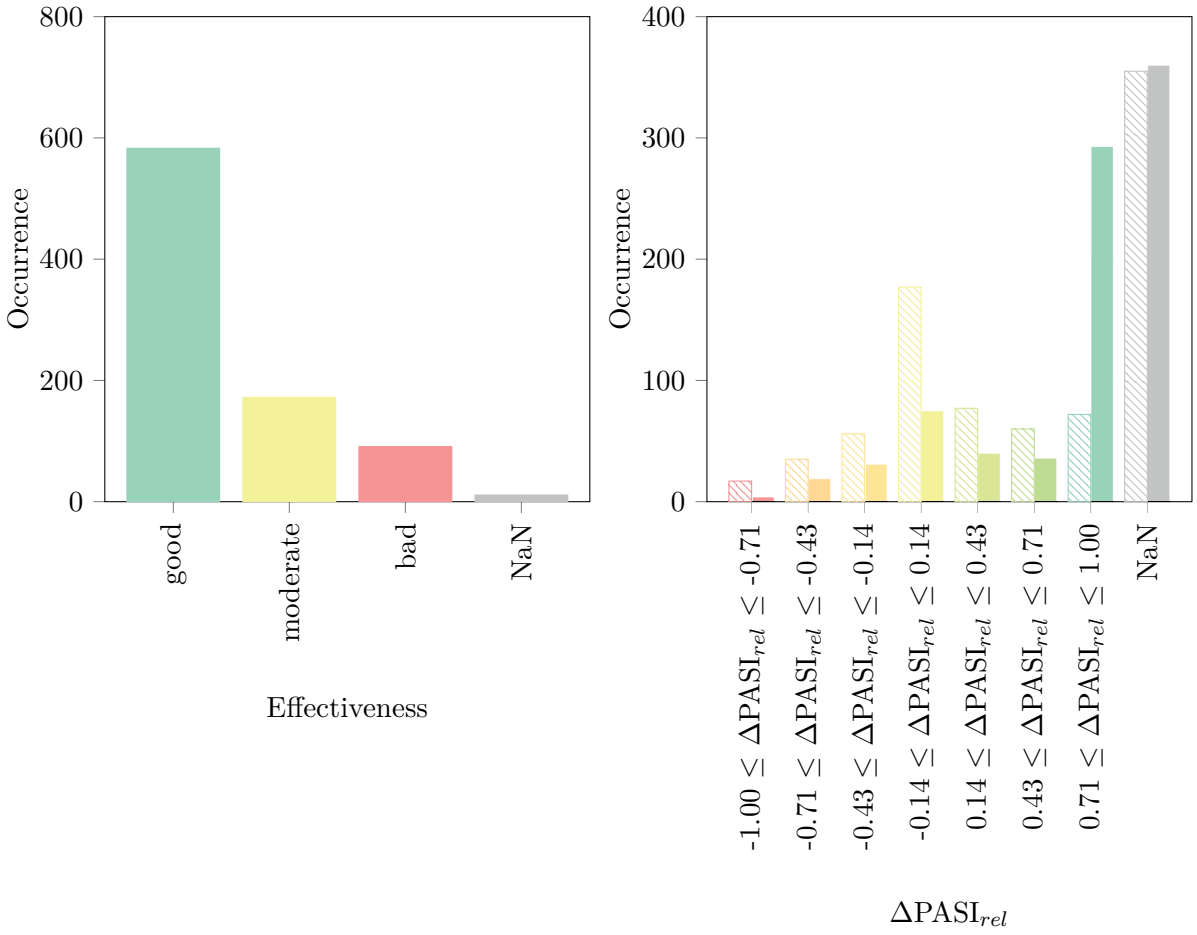


Figure B.4: Distribution of therapy *effectiveness* (a) and $\Delta PASI_{rel}$, i.e. the relative change of the PASI (b) provoked by applied therapies. $\Delta PASI_{rel}$ before (left bars) and after considering controlling therapies as having good outcome (right bars).

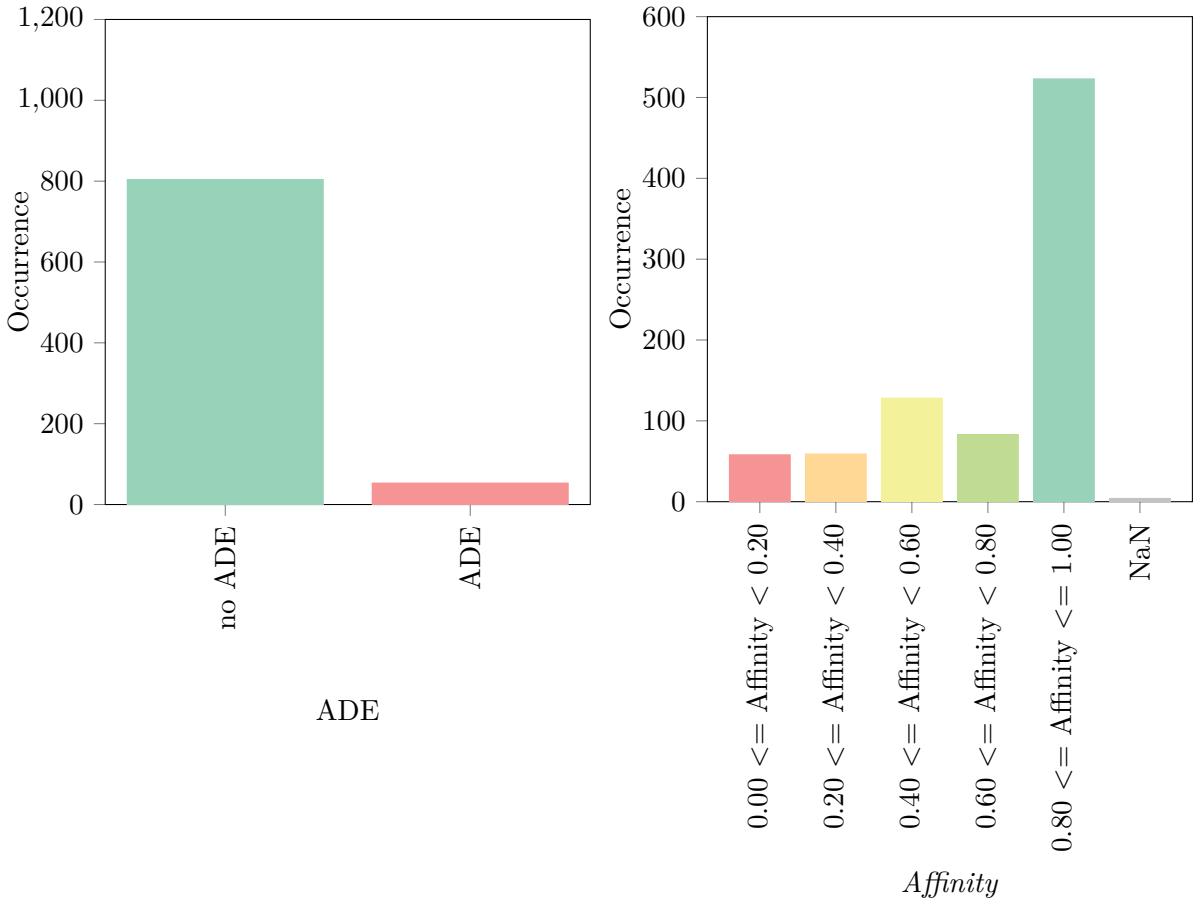


Figure B.5: Distribution of observed ADEs (left) and *affinity* scores computed for all applied therapies (right).

B.2 Treatment History Attributes

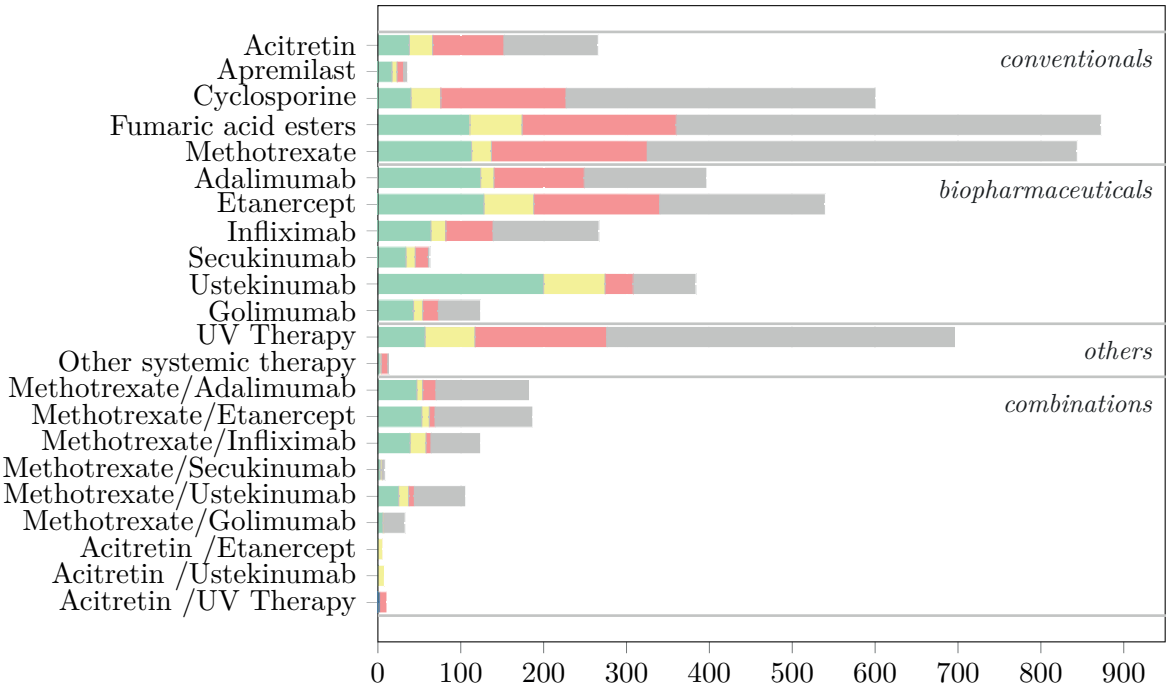


Figure B.6: *Effectiveness* associated with previously applied therapies classified into good (—), moderate (—), bad (—), or with missing outcome (—).

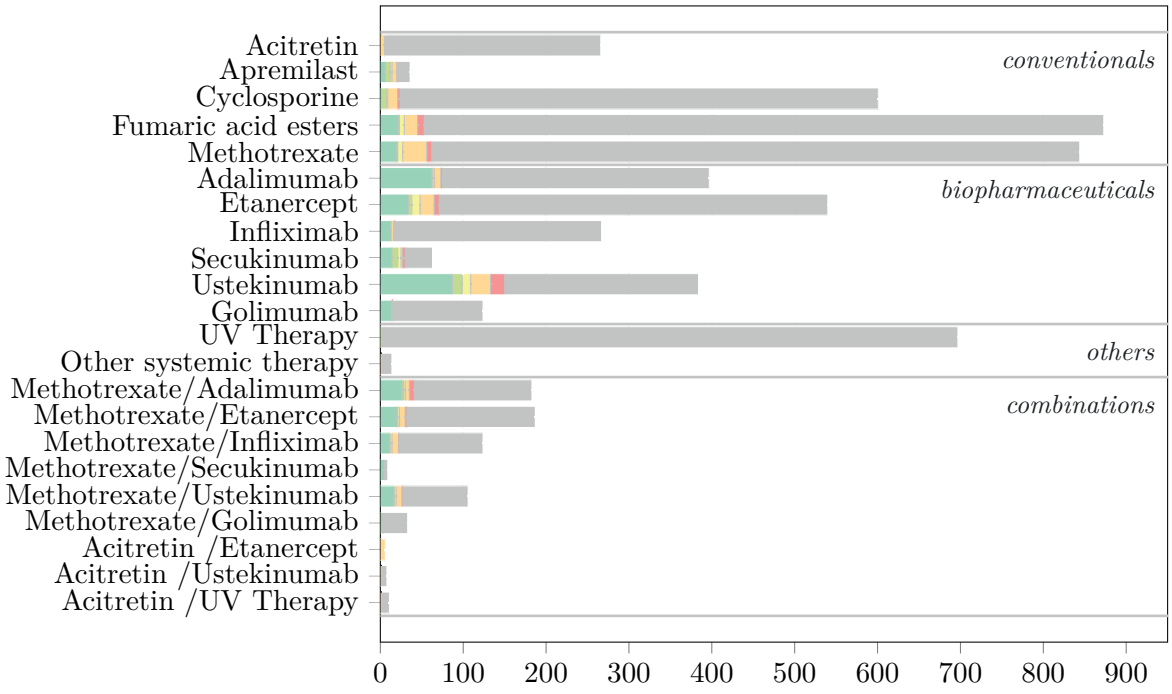


Figure B.7: Relative change of the PASI between two consecutive consultations, i.e. $\Delta PASI_{rel}$, associated with previously applied therapies. $\Delta PASI_{rel}$ values range from PASI improvement or controlling the disease (—) to deterioration of the PASI score (—), or with missing outcome (—).

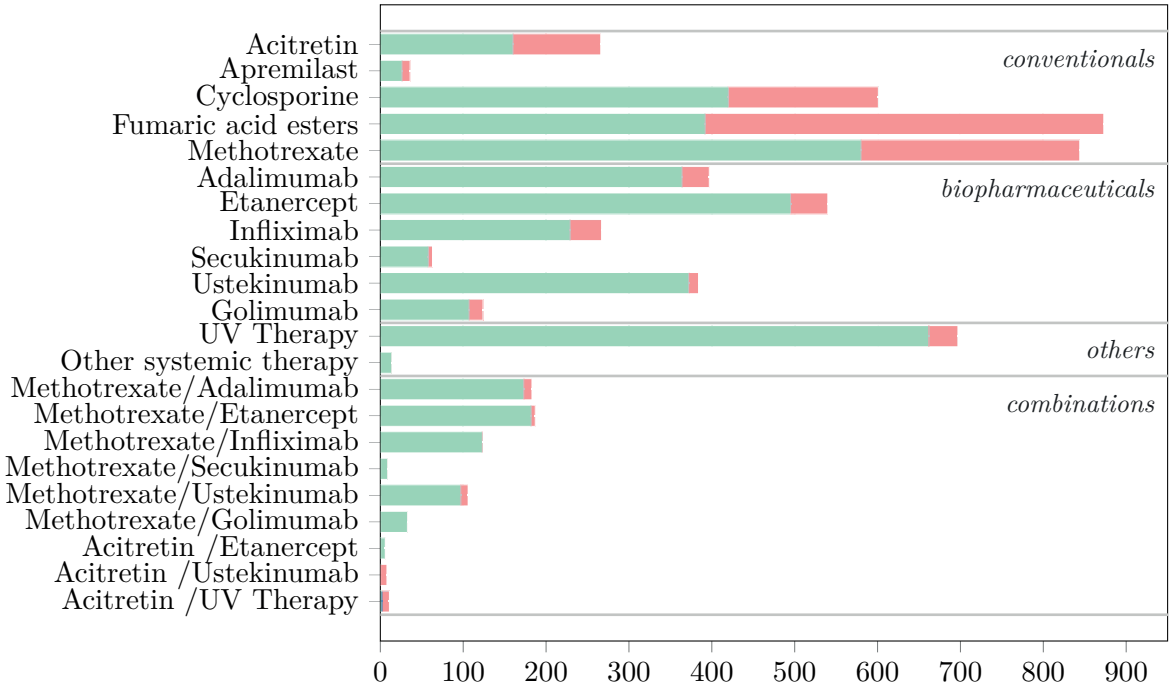


Figure B.8: Observed occurrence (—) and absence (—) of ADEs associated with previously applied therapies.

B.3 Therapy Options

Table B.1: List and categorization of therapy options applied in the provided data.

Drug	**Type**
Topical Therapies	
Vitamine-D-3/Vitamine-D-3-Analogs	-
Topical Glucocorticosteroids (Class 1 or 2)	-
Topical Glucocorticosteroids (Class 3 or 4)	-
Combination Vitamine-D-3/Glucocorticosteroids	-
Topical Retinoids	-
Dithranol	-
Other topical therapy	-
Systemic Therapies	
Acitretin	Conventional
Apremilast	Conventional
Cyclosporine	Conventional
Fumaric acid esters	Conventional
Methotrexate	Conventional
Adalimumab (TNF-α antagonist)	Biopharmaceutical
Etanercept (TNF-α antagonist)	Biopharmaceutical
Golumimumab (TNF-α antagonist)	Biopharmaceutical
Infliximab (TNF-α antagonist)	Biopharmaceutical
Secukinumab (IL-17 antibody)	Biopharmaceutical
Ustekinumab (IL-12/13 antibody)	Biopharmaceutical
Combination	
Acitretin/Etanercept	Conventional/Biopharmaceutical
Acitretin/Ustekinumab	Conventional/Biopharmaceutical
Acitretin/PUVA	Conventional/Biopharmaceutical
Methotrexate/Adalimumab	Conventional/Biopharmaceutical
Methotrexate/Etanercept	Conventional/Biopharmaceutical
Methotrexate/Infliximab	Conventional/Biopharmaceutical
Methotrexate/Secukinumab	Conventional/Biopharmaceutical
Methotrexate/Ustekinumab	Conventional/Biopharmaceutical
Methotrexate/Golumimumab	Conventional/Biopharmaceutical
Other systemic therapy	-
Phototherapies	
PUVA	-
Other UV therapy	-

B.4 Comorbidities

Table B.2: List and categorization of comorbidities recorded in the provided data.

Comorbidity	**Type**
Arterial Hypertension	Cardiovascular diseases
Cerebrovascular disease	Cardiovascular diseases
Cardiac insufficiency	Cardiovascular diseases
Condition after heart attack	Cardiovascular diseases
Condition after stroke	Cardiovascular diseases
Coronary heart disease	Cardiovascular diseases
Lympho-/thrombopenia	Cardiovascular diseases
Diabetes mellitus type 1	Metabolic diseases
Diabetes mellitus type 2	Metabolic diseases
Hyperuricaemia	Metabolic diseases
Hyperlipidemia	Metabolic diseases
Thyroid disease	Metabolic diseases
Elevated transaminases	Hepatic diseases
Hepatopathy	Hepatic diseases
Gastritis/ulcer disease	Gastrointestinal diseases
Morbus Crohn	Gastrointestinal diseases
Colitis ulcerosa	Gastrointestinal diseases
Lactose intolerance	Gastrointestinal diseases
Renal insufficiency	Renal diseases
Chronic bronchitis/COPD	Pulmonary diseases
Latent tuberculosis	Pulmonary diseases
Rheumatoid arthritis	Rheumatic diseases
Depression	Mental diseases/addictions
Smoker	Mental diseases/addictions
EX-smoker	Mental diseases/addictions
Alcohol abuse	Mental diseases/addictions
Other mental disease	Mental diseases/addictions
Asthma bronchiale	Allergic diseases
Urticaria/angioedema	Allergic diseases
Contact allergy	Allergic diseases
Drug allergy	Allergic diseases
Non-melanocytic skin cancer	Cancer
Other malignant tumor	Cancer

B.5 Data Organization

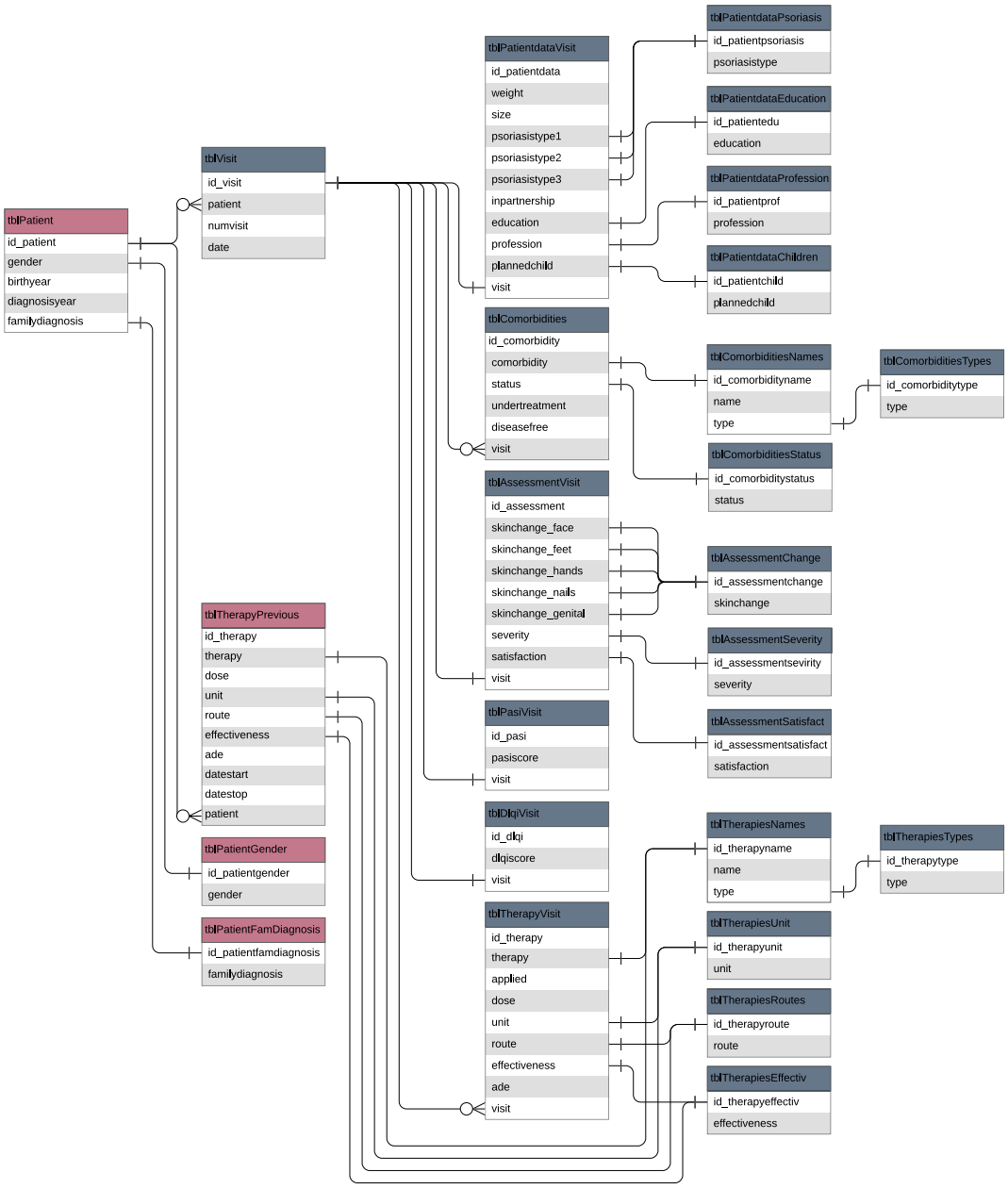


Figure B.9: Psoriasis MariaDB® database structure (ERD)

Appendix C - Dashboard

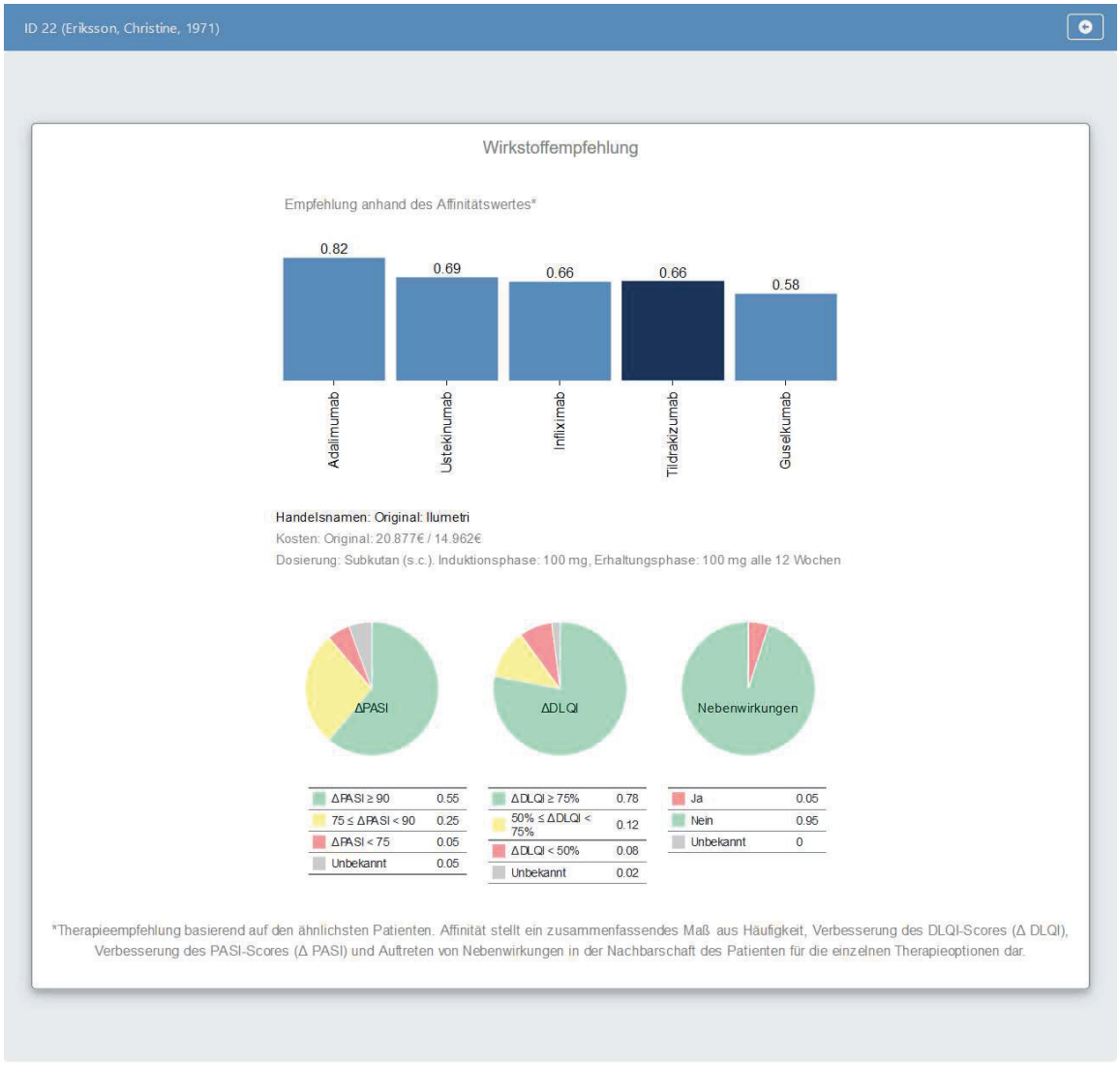


Figure C.1: Psoriasis therapy recommender system GUI: Recommendation dashboard. The predicted *affinity* scores for each therapy option after post filtering is visualized as ordered bar chart. By selecting an option, summary statistics derived from the local neighborhood of the target consultation are shown for each of the outcome indicators. Moreover, brand names of original and biosimilars are shown for the selected medication along with cost and dosage information.

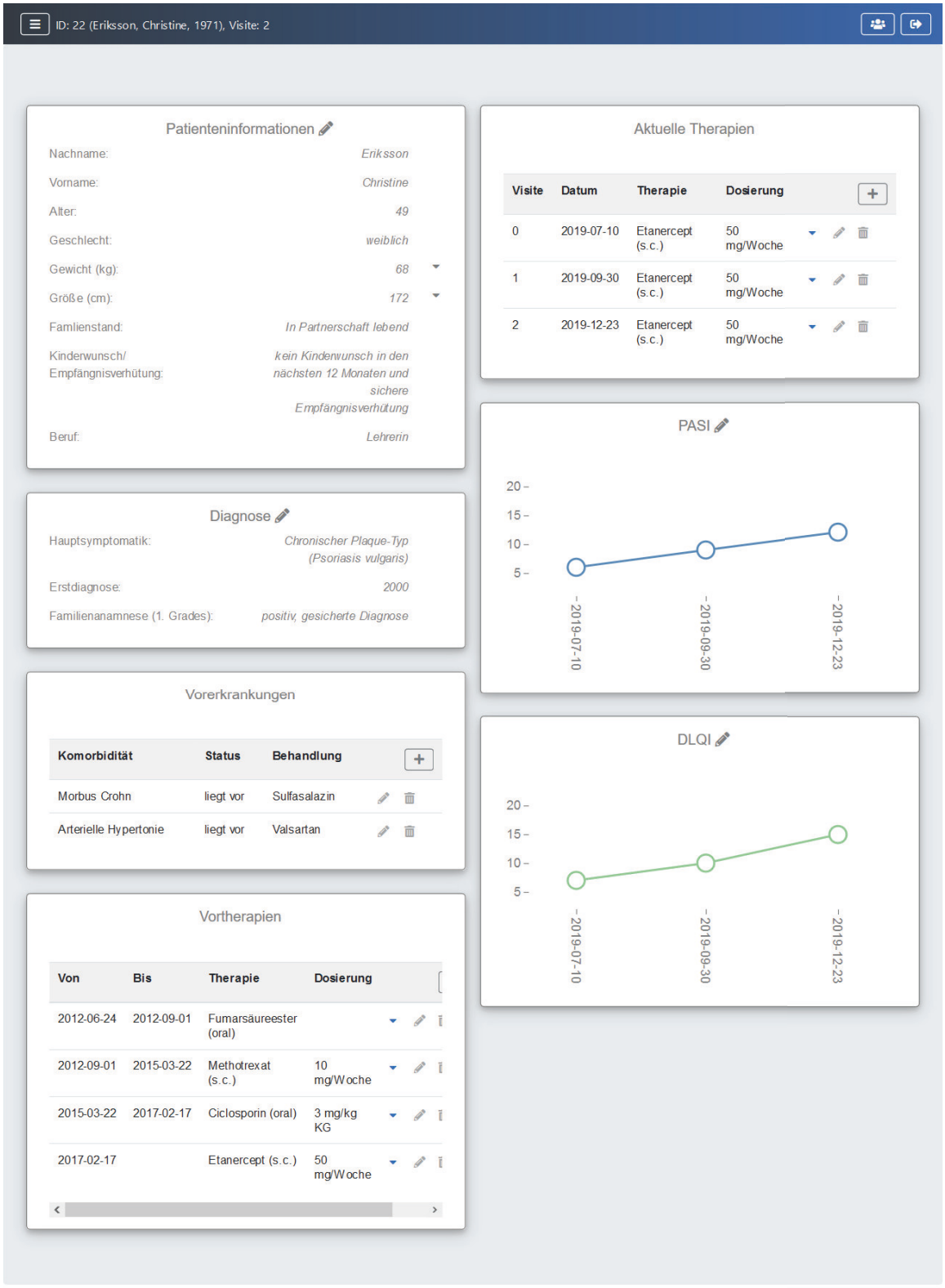


Figure C.2: Psoriasis therapy recommender system GUI: Patient and previous therapy data presentation. Patient data, such as demographic data, diagnosis, comorbidities, and clinical scores, as well as information on previous therapies and outcomes are presented for the selected patient and consultation and are editable.

Appendix D - Algorithm Comparison

Table D.1 summarizes qualitative advantages and disadvantages of the demonstrated algorithms and algorithm variants.

Table D.1: Qualitative comparison of the proposed recommendation algorithms.

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

Method	Advantages	Disadvantages
DR-Rules a (Gower) *DR-Rules b (Gower)* *DR-Rules c (Gower)*	• see *DR (Gower)*	• see *DR (Gower)*
DR-Impute 0 (Gower) *DR-Impute 1 (Gower)*	• see *DR (Gower)*	• see *DR (Gower)*
SLIM	• Optimizes linear coefficients with respect to training data • It is not required to search the entire consultation space during runtime • Embeds attribute selection • Reveals average attribute importance	• Not capable of handling missing values • Depends on sufficient and meaningful training data • Only reveals linear relationships
GBM	• Optimizes model with respect to training data • It is not required to search the entire consultation space during runtime • Capable of handling missing values • Embeds attribute selection • Reveals average attribute importance • Can reveal non-linear relationships	• Depends on sufficient and meaningful training data
Average efficiency *Overall popularity*	• No patient information required • No new patient *cold start* problem	• Not personalized

Appendix E - Further Applications

E.1 SHHS Test und Training Data

Table E.1: Train and test data partitioning.

	Subject ID
Training	0, 1, 2, 3, 4, 5, 7, 8, 11, 12, 13, 14, 17, 20, 21, 22, 23, 26, 27, 28, 29, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 42, 43, 44, 46, 47, 48, 49, 50, 51, 52, 53, 54, 56, 57, 58, 59, 61, 62, 63, 64, 65, 66, 67, 69, 70, 71, 72, 74, 76, 77, 78, 79, 80, 81, 83, 84, 85, 87, 88, 89, 90, 91, 92, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 110, 111, 116, 117, 118, 119, 121, 122, 123, 125, 128, 129, 130, 131, 132, 133, 134, 135, 136, 138, 139, 140, 142, 143, 144, 145, 146, 147, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165, 166, 168, 169, 171, 173, 174, 175, 176, 177, 178, 179, 180, 181, 183, 184, 187, 188, 189, 190, 191, 192, 194, 195, 196, 197, 198, 199, 200, 201, 202, 203, 207, 208, 209, 210, 212, 213, 214, 215, 216, 217, 219, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 236
Test	6, 9, 10, 15, 16, 18, 19, 24, 25, 30, 38, 45, 55, 60, 68, 73, 75, 82, 86, 93, 109, 112, 113, 114, 115, 120, 124, 126, 127, 137, 141, 148, 167, 170, 172, 182, 185, 186, 193, 204, 205, 206, 211, 218, 220, 221, 234, 235

Table E.2: Subject mapping between utilized subject ID (ID) and SHHS subject ID (SHHS).

ID	SHHS	ID	SHHS	ID	SHHS	ID	SHHS	ID	SHHS	ID	SHHS
0	200039	41	201218	82	203150	123	203737	164	204429	205	204846
1	200041	42	201276	83	203155	124	203791	165	204434	206	204861
2	200083	43	201309	84	203158	125	203801	166	204436	207	204881
3	200089	44	201351	85	203171	126	203824	167	204443	208	204894
4	200122	45	201411	86	203179	127	203851	168	204449	209	204895
5	200123	46	201413	87	203184	128	203856	169	204450	210	204910
6	200148	47	201423	88	203213	129	203925	170	204472	211	204912
7	200166	48	201426	89	203251	130	203942	171	204478	212	204916
8	200172	49	201441	90	203255	131	203952	172	204490	213	204922
9	200231	50	201553	91	203259	132	203995	173	204491	214	204927
10	200328	51	201654	92	203274	133	203998	174	204494	215	204929
11	200347	52	201687	93	203292	134	204027	175	204500	216	204937
12	200350	53	201694	94	203311	135	204028	176	204530	217	204952
13	200383	54	201834	95	203324	136	204035	177	204550	218	204960
14	200390	55	201906	96	203330	137	204041	178	204553	219	204963
15	200405	56	202000	97	203333	138	204042	179	204559	220	204985
16	200413	57	202039	98	203350	139	204058	180	204560	221	204995
17	200485	58	202041	99	203375	140	204084	181	204562	222	205011
18	200555	59	202058	100	203384	141	204089	182	204576	223	205060
19	200586	60	202116	101	203386	142	204115	183	204603	224	205069
20	200592	61	202185	102	203442	143	204134	184	204614	225	205083
21	200596	62	202255	103	203457	144	204135	185	204631	226	205116
22	200620	63	202267	104	203476	145	204166	186	204632	227	205146
23	200644	64	202433	105	203478	146	204171	187	204638	228	205152
24	200661	65	202503	106	203483	147	204176	188	204657	229	205177
25	200678	66	202658	107	203488	148	204190	189	204722	230	205232
26	200750	67	202847	108	203512	149	204233	190	204729	231	205276
27	200774	68	202949	109	203523	150	204237	191	204735	232	205285
28	200818	69	202957	110	203533	151	204292	192	204749	233	205295
29	200880	70	202963	111	203554	152	204298	193	204769	234	205311
30	200884	71	202986	112	203561	153	204299	194	204772	235	205328
31	200887	72	202994	113	203599	154	204307	195	204774	236	205383
32	200891	73	203311	114	204326	155	204326	196	204777		
33	200954	74	203324	115	204338	156	204338	197	204778		
34	200955	75	203330	116	204365	157	204365	198	204781		
35	201042	76	203333	117	204368	158	204368	199	204785		
36	201066	77	203350	118	204388	159	204388	200	204789		
37	201067	78	203375	119	204412	160	204412	201	204814		
38	201079	79	203384	120	204413	161	204413	202	204818		
39	201210	80	203386	121	204419	162	204419	203	204831		
40	201218	81	203442	122	204429	163	204429	204	204846		

E.2 Drugs.com and Druglib.com Data Description

Table E.3: Drugs.com and Druglib.com data description. Number of training and test samples, average and standard deviation of review lengths, applied rating thresholds, and label distribution are shown.

Data	Train	Test	Conditions	Drugs	Length	Rating	Label	%
Drugs.com								
Overall Rating	161297	53766	836	3654	458.32 (240.76)	$rating \leq 4$	-1	25
						$4 < rating < 7$	0	9
						$rating \geq 7$	1	66
Side Effects (Annotated)	-	400	141	243	500.385 (209.42)	No Side Effects	0	32
						Mild / Moderate Side Effects	1	28
						Severe / Extremely Severe Side Effects	2	40
Druglib.com								
Overall Rating	3107	1036	1808	541	277.57 (283.21)	$rating \leq 4$	-1	21
						$4 < rating < 7$	0	10
						$rating \geq 7$	1	69
Benefits (Effectiveness)	3107	1036	1808	541	212.87 (198.51)	Ineffective	0	8
						Marginally / Moderately Effective	1	19
						Considerably / Highly Effective	2	73
Side Effects	3107	1036	1808	541	177.36 (197.93)	No Side Effects	0	30
						Mild / Moderate Side Effects	1	53
						Severe / Extremely Severe Side Effects	2	17

Appendix F - Fundamentals

F.1 Decision Trees

F.1.1 Decision Tree Induction

To measure the purity or homogeneity of a collection of training samples $\mathcal{S}$ the *Shannon entropy* $H(\mathcal{S})$ can be employed which is defined as

$$H(\mathcal{S}) = -\sum_{y \in \mathcal{Y}} p(y|\mathcal{S}) \cdot \log_2 p(y|\mathcal{S}) \tag{F.1}$$

where $p(y|\mathcal{S})$ is the proportion of $\mathcal{S}$ belonging to class y. Suchlike, the *information gain* $IG(\mathcal{S}_i, A)$, which origins from information theory, can be calculated as splitting criterion. Information gain measures how well a given attribute $\mathcal{A}$ splits the proportion of training observations $\mathcal{S}_i$ reaching node i, i.e. how well a split reduces entropy and increases homogeneity. Information gain, also denoted as *mutual information*, was initially proposed in [275] as splitting criterion for the ID3 algorithm (Iterative Dichotomizer) and is defined as

$$IG(\mathcal{S}_i, A) = H(\mathcal{S}_i) - \sum_{j \in J} \frac{|\mathcal{S}_i^j|}{|\mathcal{S}_i|} H(\mathcal{S}_i^j) \tag{F.2}$$

with the possible nominal categories j of A yielding the subset $\mathcal{S}_i^j$. Depending on the number of resulting branches, i.e. categories, the algorithm creates a multiway tree which tries to find for each node the attribute yielding the largest information gain. [226, 275]
The ID3 induction process either terminates if all attributes were already selected in a path or if all training examples in a node are member of the same class, i.e. the entropy is zero. The empirical class distribution from the training process are stored for each terminal leaf. Suchlike, for each sample to be classified a probabilistic class membership can be determined. [226, 275]
The information gain described above suffers from the drawback of favoring attributes with a large number of possible categories. To reduce this bias the normalized version of the information gain, the *information gain ratio*, takes number and size of branches into account resulting in

$$IGR(\mathcal{S}_i, A) = IG(\mathcal{S}_i, \mathcal{A}) \cdot \left(-\sum_{j \in J} \frac{|\mathcal{S}_i^j|}{|\mathcal{S}_i|} \cdot \log_2 \frac{|\mathcal{S}_i^j|}{|\mathcal{S}_i|} \right)^{-1} \tag{F.3}$$

Information gain ratio was proposed in [276] which describes the *C4.5* algorithms, the successor of ID3. C4.5 incorporates numerous improvements as described below. One major difference

is that in contrast to its predecessor, C4.5 is capable of not only handling qualitative but also quantitative attributes. During the induction process, a dynamically determined threshold partitions the quantitative attribute into a discrete set of intervals [276]. In contrast to qualitative attributes which are tested at most once on any path in the tree, quantitative attributes may be tested several times with different thresholds.
Finally, the CART algorithm proposed in [37] is a decision tree learning technique that facilitates both, creation of either classification or regression trees, depending on whether the dependent variable is categorical or continuous, respectively. Comparable to C4.5, CART supports both qualitative and quantitative attributes.
For categorical labels CART implements the *Gini index* to measure the impurity or inhomogeneity of a collection of training samples S

$$Gini(\mathcal{S}_i) = 1 - \sum_{y \in \mathcal{Y}} P(y|\mathcal{S}i)^2 \tag{F.4}$$

where $P(y|\mathcal{S})$ is the proportion of $\mathcal{S}$ belonging to class y. The Gini index can be interpreted as the probability to incorrectly label a random sample from the distribution. Consequently, attributes and thresholds which maximize Gini index reduction are selected for the current node. The resulting splitting criterion can be defined as

$$GiniGain(\mathcal{S}i, A) = Gini(\mathcal{S}_i) - \sum_{j \in J} \frac{|\mathcal{S}_i^j|}{|\mathcal{S}_i|} Gini(\mathcal{S}_i^j) \tag{F.5}$$

In contrast to ID3 and C4.5, where qualitative variables with more than two categories lead to multiway splits, the CART algorithm creates binary splits, i.e. each internal node has exactly two outgoing edges. Therefore, multiple categories are divided into two disjoint groups which best improve the splitting criterion. Multiway splits result in broader trees. However, they can suffer from the drawback to fragment the data too quickly which results in insufficient data succeeding the node [149].
As for classification tree induction, also when building a regression tree the objective is to select attributes and thresholds such that a splitting criterion is minimized. However, this criterion must reflect the numeric deviation from the continuous target value. The CART algorithm uses the squared error of each potential split to determine the best splitting option.

To terminate the decision tree induction process, various stop criteria are described. The learning phase can be terminated if all attributes were selected in a path or if all training examples in a node are member of the same class, i.e. the entropy is zero [226, 275]. However, additionally criteria such as the maximum tree depth, i.e. the path length from the root node to a leaf node, the maximum number of decision splits, i.e. branch nodes, the minimum number of observations in a parent node to perform further splits, the minimum number of observations in resulting leafs or a splitting criterion threshold can be defined to further control the specific structure of a grown decision tree.[291]

Finally, the empirical class distribution from the training process are stored for each leaf.

Suchlike, for each sample to be classified a probabilistic class membership can be determined [226, 275]. In case of a regression task, the actual prediction value for each leaf is the weighted mean of the training data stored in the respective leaf.

F.1.2 Decision Tree Missing Values

Both C4.5 and CART are capable of handling missing values in an attribute vector. In case of C4.5, observations reaching an internal node, for which an attribute is unknown, are sent into each branch. However, the observation is weighted, i.e partitioned, with the proportion in which the training observations (for which the attribute was known) were split at this respective node. Determining the information gain during tree induction is also valid with weighted observations. During classification of a new observation the weights are incorporated into the computation of the classification probability at the respective leaf node. [226, 384]
The CART algorithm uses the concept of surrogate splits for handling missing values. During training, a primary splitting attribute and threshold is determined on observations where the respective attribute is not missing. From all training observations, for which the splitting attribute is missing, a list of surrogate attributes and thresholds are determined, which is sorted by their capability to mimic the primary split. The underlying intention is to exploit the correlation between variables to diminish the impact of missing variables. [149].

F.1.3 Decision Tree Pruning

A notable issue with decision trees is their tendency to overfitting. If trees are grown too deeply, branches develop that only represent outliers or noise instead of concepts inherent to the data. In order to tackle this overfitting problem pre- or post-pruning strategies are employed. Pre-pruning, on the one hand, means parameterizing stopping criteria to the problem at hand as prescribed above requiring additional validation data. Post-pruning, on the other hand, does not require any additional data. CART and C4.5 provide a bottom-up post-pruning method replace internal nodes with leaf nodes or to raise subtrees. The decision whether to replace or remove a node is based on the comparison of the expected errors occurring at the respective node and its successors. The assumed true class for each of the internal and leaf nodes is the majority class from the training data reaching that node. Thus, the expected error is computed as the misclassification rate at each node but using the upper-bound of a given confidence interval to derive a more pessimistic error estimate. The expected standard deviation is computed assuming a binomial distribution of the occurring error. [226, 384]

F.1.4 Decision Tree Ensembles

F.1.5 Bagging

One of the earliest methods to generate classifier ensembles introduced in [39] is *bagging* (bootstrap aggregating). To facilitate diversity when training individual classifiers, varying subsets of the training data are employed to build up each single model. Subsets of equal size are boot-

strap sampled. The output of the individual classifiers are aggregated using plurality vote for classification or averaging when predicting numerical outcomes, respectively.

A variation of *bagging*, so called RF introduced in [41], use randomly varying and uncorrelated DTs as base learner to construct an ensemble model (algorithm 2). The underlying intuition was to reduce the variance of the individual decision trees by averaging their outputs [271, 149]. Here, a popular approach is to use bootstrap samples of the dataset and additionally increase diversity and decrease correlation among the individual trees by using random subspaces only, i.e. randomly chosen features, at each node when inducing the base learner. However, any ensemble of DTs each grown with respect to independent and identically distributed random vectors are defined as RF [190, 41].
Bagging of DTs comes along with a noteworthy feature. Already during classifier training an error estimate, the so called *out-of-bag* error, can be determined to assess the classifier quality. Therefore, for each base learner the misclassification rate is computed using all samples not used for training, i.e. the out-of-bag observations. Those error estimates are finally averaged over all base learner and can render a hold-out validation set for hyperparameter tuning unnecessary.

RFs are applied in chapter 7 in the context of sleep stage classification and PD patient classification.

Algorithm 2: Random Forest

Procedure: Training
Input : Training data $\mathbf{S} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ of size N
number of base learner M
fraction f to be drawn from $\mathcal{S}$ at each iteration with replacement
fraction p of attributes to be drawn at each split
Output : Ensemble model $F(\mathbf{x})$

for $m = 1...M$ **do**
randomly draw fraction f from $\mathcal{S}$, i.e. bootstrap sample $\mathcal{S}_m$;
fit base learner to all y from $\mathcal{S}_m$ yielding $h_m(\mathbf{x})$, only use random fraction p of features at each split;
update model: $F(\mathbf{x}) = F_{m-1}(\mathbf{x}) + h_m(\mathbf{x})$;
end

Procedure: Classification
Input : Data sample $\mathbf{x}$
Ensemble model $F(\mathbf{x})$
Output : Prediction $\hat{y}$

for $m = 1...M$ **do**
classify $\mathbf{x}$ using $h_m(\mathbf{x})$ yielding $\hat{y}_m$;
end
do majority voting over all $\hat{y}_m$ to determine $\hat{y}$

F.1.6 Boosting

In [303] it was shown that any *weak learner*, i.e. classification or regression models performing only marginally better than random guessing having an error rate less than 0.5, can be "boosted" to a *strong learner*. Boosting is considered as one of the most powerful classification and regression approaches introduced in the recent years [149]. In case of classification, analogously to *bagging*, an ensemble of models is created by combining base learner, trained on resampled versions of the training data. However, in case of *boosting* the individual classifiers are incrementally added and, in contrast to random samples, trained with the most informative, i.e. most difficult data samples, at each iteration. Suchlike, base learners are combined which compliment each another and are specialized to a specific domain of the data space. Finally, when combining the base learner, each contribution is weighted by its confidence instead of giving equal weight as in case of *bagging*. [190, 271, 149, 384]

One of the most widely used *boosting* algorithms is *AdaBoost* (Adaptive Boosting), which was introduced in [114, 115] and is described in algorithm 3. Here, the individual base learner are generated by drawing samples from an iteratively updated training data distribution. This distribution update increases the likelihood of instances to be included in the training data of a next learner if they were misclassified by the previous one. Suchlike, the algorithm focuses on increasingly difficult instances with each iteration. Also when computing the misclassification rate, the likelihood of the individual samples are taken into account. To ensure the *boosting* effect, only classifiers having a misclassification rate below 0.5 are added to the ensemble and are discarded otherwise. During actual classification of unseen samples a weighted majority voting scheme is applied. Base learner having shown good training performance have more impact, i.e. are given higher weights relative to their classification performance during training. Hyperparamters such as the optimal number of iterations M must be determined using cross validation or a disjunct validation data set.

AdaBoost and related *boosting* algorithms were converted into a generalized, framework [38, 40, 116, 117] and finally denoted as GBM. Here, *boosting* is considered as a numerical optimization problem with the objective to minimize the overall error of the model. Therefore, specialized base learner (*weak learner*) are added incrementally to the ensemble using a GD like approach. The generic framework allows to apply arbitrary differentiable objective functions $L(y, F(\mathbf{x}))$ and any parameterizable classifiers $F(\mathbf{x})$ as base learner. However, regression trees are applied very commonly using squared error $L(y, F(\mathbf{x})) = \frac{1}{2}(y - F(\mathbf{x}))^2$ as objective function. Suchlike, the residuals $\mathbf{r}_m = y - F(\mathbf{x})$ of a model m can be interpreted as negative gradients. However, any objective function derivation can be inserted into the algorithm as *pseudo-residual*. In each iteration the overall model is updated with a base learner $h_m(\mathbf{x})$ fitted to the residuals $\mathbf{r}_m$ yielding $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x})$. Analogously to *AdaBoost*, a weighted majority voting scheme is applied during actual classification or regression. Here, the weighting coefficient γ_m is typically also optimized by solving one-dimensional line search optimization problem. As in case of *AdaBoost* a stopping criterion, such as a data specific fixed number of iterations M or a loss threshold must be determined using cross validation or validation data.

Algorithm 3: AdaBoost

Procedure: Training
Input : Training data $\mathcal{S} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ of size N
number of base learner M
fraction f to be drawn from data distribution at each iteration
Initialize : Distribution $\mathbf{D}_0(n) = \frac{1}{N}$, for $n = 1...N$
Output : Ensemble model $F(\mathbf{x})$

for $m = 1...M$ **do**
  randomly draw fraction f from distribution $\mathbf{D}_m$, i.e. sample $\mathcal{S}_m$;
  fit base learner to all y from $\mathcal{S}_m$ yielding $h_m(\mathbf{x})$;
  calculate error of $h_m(\mathbf{x})$: $\epsilon_m = \sum_{n:h_m(\mathbf{x}_n)\neq y_n} \mathbf{D}_m(n)$ from all $(\mathbf{x}, y) \in \mathcal{S}_m$;
  if $\epsilon_m > 0.5$ **then**
    abort;
  else
    compute coefficient $\beta_m = \frac{\epsilon_m}{1-\epsilon_m}$;
    update distribution $\mathbf{D}_m$: $\mathbf{D}_{m+1}(n) = \frac{\mathbf{D}_m(n)}{|\mathbf{D}_m|} \times \begin{cases} \beta_m & \text{if } h_m(\mathbf{x}_n) = y_i \\ 0 & \text{otherwise} \end{cases}$;
    update model: $F(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \log \frac{1}{\beta_m} h_m(\mathbf{x})$;
  end
end

Procedure: Classification
Input : Data sample $\mathbf{x}$
Ensemble model $F(\mathbf{x})$
Output : Prediction $\hat{y}$

for $m = 1...M$ **do**
  classify $\mathbf{x}$ using $h_m(\mathbf{x})$ yielding $\hat{y}_m$;
end
do majority voting over all $\hat{y}_m$ to determine $\hat{y}$. Weight each classifier contribution using $w_m = \log \frac{1}{\beta_m}$;

Various improvements, especially to tackle overfitting, were added to the initial GBM framework. By weighting the contribution of each base learner using a *shrinkage* parameter or *learning rate* μ, the influence of the individual base learner is reduced and, as a consequence, learning speed decreases and the optimal number of classifiers increases [117]. Also training the base learner on a subsample of the training data only (*Stochastic Gradient Boosting*), which is chosen randomly without replacement at each iteration [118] or other sampling or feature subspace selection schemes, often improve accuracy. Furthermore, considering the numeric values in the leaf nodes of a regression tree as weights, regularization of those parameters by means of L_1- or L_2-norm have proven to additionally improve performance (*Regularized Gradient Boosting*) [62]. This algorithm is applied in the model-based CF setting in section 5.5 to predict treatment outcome and derive recommendations.

Algorithm 4: Gradient Boosting Machine (GBM)

Procedure: Training

Input : Training data $\mathcal{S} = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$ of size N
number of base learner M
differentiable objective function $L(y, F(\mathbf{x}))$

Output : Ensemble model $F(\mathbf{x})$

Initialize : Initial model with constant value γ: $F_0(\mathbf{x}) = \arg\min_{\gamma} \sum_{n=1}^{N} L(y_n, \gamma)$

for $m = 1...M$ **do**

compute *pseudo-residuals*: $r_{nm} = \left[\frac{\partial L(y_n, F(\mathbf{x}_n))}{\partial F(\mathbf{x}_n)}\right]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}$, for $n = 1...N$;

fit base learner to all *pseudo-residuals* $\mathbf{r}_m$, i.e. train on all $(\mathbf{X}, \mathbf{r}_m)$ yielding $h_m(\mathbf{x})$;

compute coefficient γ_m: $\gamma_m = \arg\min_{\gamma} \sum_{n=1}^{N} L(y_n, F_{m-1}(\mathbf{x}_n) + \gamma h_m(\mathbf{x}_n))$;

update model: $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \gamma_m h_m(\mathbf{x})$;

end

Procedure: Classification

Input : Data sample $\mathbf{x}$
Ensemble model $F(\mathbf{x})$

Output : Prediction $\hat{y}$

for $m = 1...M$ **do**

classify $\mathbf{x}$ using $h_m(\mathbf{x})$ yielding $\hat{y}_m$;

end

do majority voting over all $\hat{y}_m$ to determine $\hat{y}$. Weight each classifier contribution using $w_m = \gamma_m$;

F.1.7 Decision Tree Ensemble Interpretability

As stated in 3.3.2, single DTs are characterized by good interpretability. The generated model can be visualized by a two-dimensional graphic or described as a set of if-then rules, however, often suffer from overfitting and inaccuracy. Linear combination strategies of trees as *bagging* and *boosting* are capable of significantly improve classification accuracy, however, at the expense of some analyzing and interpretation capabilities.

Nevertheless, one insight into the decision making process, which can be derived, are estimates of average attribute importance. The cumulated improvements of the split-criterion at each split can be regarded as measure of relevance associated to the splitting variable. Those values can finally be averaged over all trees yielding reliable overall attribute importance estimates. However, two properties of this importance measure must be kept in mind. In case of correlated attributes, usually only one attribute is rewarded with high importance whereas the correlated attribute will end up with low scores. Additionally, the attribute importance of categorical attributes is typically biased towards variables with many categories.

Furthermore, ensembles of classification models typically not only return a majority class only but also the support for a given class. This numeric value can be interpreted as some probability

measure and used for ranking of different classifier outputs. Hence, those models are practicable for RS applications which provide ranked lists of recommendations.
Various approaches are proposed in the literature which aim at mimicking an overall better performing but complex model, e.g. neural networks [394, 46, 72, 348] or ensembles of DTs [91], by training a more compact and less complex *surrogate* model, e.g. linear models or DTs. Suchlike, not only the reduction of processing time and space requirements can be facilitated, but also the level of interpretability can be increased. All of those approaches have in common that no information about the inner workings of the underlying complex model is required (*model-agnostic interpretation methods* [230]). Furthermore, all algorithms use the labels predicted by the complex ensemble model (*teacher*), to train a simpler surrogate model (*student model*).
Such surrogate models can be global, but also local interpretable surrogate model approaches ar proposed in the literature which aim at explaining the complex models predictions of individual target instances (LIME [287], MAPLE [270]). To do so, less complex and interpretable surrogate models, which approximate the complex model, are trained on data from the neighborhood of a target instance only.

F.2 Matrix Factorization

The basic ideas concerning MF algorithms derive from Latent Semantic Indexing (LSI) for document comparison [83, 300, 299]. By transforming a document representations to a lower rank approximation, also denoted as *word embedding*, LSI was intended to reveal latent concepts contained in documents.
In the CF setting, the same approach can be applied to the n rank user-item feedback matrix $\mathbf{R}$. Here, MF approaches aim at mapping the user representations in $\mathbf{R}$, i.e. the explicit or implicit feedback of user u on items i, to a joint lower-dimensional space by capturing their most important latent factors [182]. As $\mathbf{R}$ typically comprises a large number of correlations among users and items, i.e. inherent redundancy, a good approximation $\hat{\mathbf{R}}$ can be achieved with rank $k << n$, i.e. by only retaining the most important dimensions [288, 4].
Such latent factor models are closely related to Singular Value Decomposition (SVD) using the *Eigendecomposition* which factorizes $\mathbf{R}$ into the three matrices

$$\mathbf{R} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{F.6}$$

where $\mathbf{U}$ is the $|\mathcal{U}| \times n$ matrix of left singular vectors, $\mathbf{V}$ is the $|\mathcal{I}| \times n$ matrix of right singular vectors and $\mathbf{\Sigma}$ is the $n \times n$ diagonal matrix of ordered singular values of $\mathbf{R}$. Here, the left singular vectors can be interpreted as latent user features whereas the right singular vectors can be interpreted as latent item features. [300, 299]
If only a subset of the k largest singular values of $\mathbf{\Sigma}$ are used, a lower-dimensional approximation $\hat{\mathbf{R}} = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ of $\mathbf{R}$ using only the largest singular values in $\mathbf{\Sigma}_k$ along with the corresponding singular vectors $\mathbf{U}_k$ and $\mathbf{V}_k$ can be obtained.

However, when factorizing the user-item feedback matrix $\mathbf{R}$ in a CF application, sparsity raises difficulties as conventional SVD is undefined when the matrix is incomplete. One solution is to assign default values, e.g. the mean of the corresponding row or column to missing entries [169, 300]. However, this approach significantly increases the computational load and is likely to introduce a considerable bias to the data due to inappropriate imputation.

Several works suggest to model a factorization $\mathbf{R} = \mathbf{P}\mathbf{Q}$ by only using the observed entries of $\mathbf{R}$ directly [27, 52, 181, 259, 347]. Due to the inherent redundancies in the data, the fully specified low-rank approximation assumed to be determined even with a small subset of the entries in the original matrix only [4]. The factor vectors are typically learned from previously observed implicit or explicit feedback by minimizing an objective function $L(\mathbf{P}, \mathbf{Q})$ such as the squared entry-wise L_2-norm (*Frobenius-norm*) of the approximation error on the known feedback

$$L(\mathbf{P}, \mathbf{Q}) = ||\mathbf{R} - \mathbf{P}\mathbf{Q}^T||_2^2 = \sum_{u,i}(r_{ui} - \mathbf{p}_u\mathbf{q}_i)^2 \tag{F.7}$$

Several optimization algorithms are proposed in the literature for minimizing the objective function. Typically, GD or SGD are applied. However, there are also more specialized optimization methods as Alternating Least Squares (ALS). As the defined objective function $L(\cdot)$ is convex for $\mathbf{P}$ or $\mathbf{Q}$, only but not in both variables together, ALS optimizes $\mathbf{p}_u$ and $\mathbf{q}_i$ in an alternating

fashion to transform the overall optimization problem into a convex and hence optimally solvable problem.
Regularization, which penalizes large coefficients in $\mathbf{P}$ and $\mathbf{Q}$, has been shown to be essential to address overfitting and improve generalization capabilities [4, 182, 288]. Here, typically the entry-wise L_2-norms of $\mathbf{P}$ and $\mathbf{Q}$ are added to the objective function.

$$L(\mathbf{P}, \mathbf{Q}) = \sum_{u,i} (r_{ui} - \mathbf{p}_u \mathbf{q}_i)^2 + \lambda(||\mathbf{p}_u||_2^2 + ||\mathbf{q}_i||_2^2) \tag{F.8}$$

which are controlled with the regularization parameter λ [182, 288]. Numerous MF algorithms have been developed within the recent years. Even though latent vectors represent dominant correlation patterns in $\mathbf{R}$, the individual latent features typically become hardly interpretable. One specialized factorization methods intended to cope with this issue is Non-negative Matrix Factorization (NMF) [397]. This approach is especially useful in applications where feedback comprises positive values only. Therefore, an additional constraint is induced to the optimization function in equation F.8 which forces the factors in $\mathbf{P}$ and $\mathbf{Q}$ to be non-negative: $\mathbf{P} \geq 0$, $\mathbf{Q} \geq 0$. Suchlike, no subtractions are involved when computing $\hat{\mathbf{R}}$ by multiplying the resulting matrices $\mathbf{P}$ and $\mathbf{Q}$.

Based on MF, similarity can be computed between user representations $\mathbf{p}_u$ comparable to the traditional memory-based CF. Using those fully specified vectors representing users $\mathbf{p}_u$, more robust and, due to the lower dimensionality of the vectors, more efficient similarity computations are yielded [288]. A further approach to model the feedback prediction r_{ui} a user u gives on an item i is the even more efficient way to compute the inner products of user representations $\mathbf{p}_u$ and item representations $\mathbf{q}_i$ in the transformed space [182, 345].

Comparing both, traditional memory-based CF methods with techniques employing MF, the first methods are more intuitive and explainable. However, factorization techniques improve scalability and efficiency. Moreover, as the resulting latent factors can be considered as the extent of interest a user has in an item and to which an item possesses the latent factors, MF approaches bear the additional potential to reveal meaningful relations between pairs of users or items which are otherwise unknown [182].

F.3 Missing Value Imputation

(a) *Complete-case analysis*, also denoted as *listwise deletion*: Instances containing missing values are simply discarded. Assuming the data to be MCAR, suchlike no data errors are introduced and data is kept reliable. Though, reducing the dataset is subject to an overall loss of information which affects the statistical power and generalizability of further analyses [262].

(b) *Single value imputation*: Imputation of single value estimates at places of missing data in order to make most of the available data. Depending on the proportion of the missing values and the further analyses, this approach is prone to introduce noise into the data and cause bias or false classification results if not applied appropriately [241, 262].
A straightforward statistical approach for missing categorical attributes is imputation of the most frequently occurring, i.e. the mode of the available values. This approach is especially justifiable in case of very unbalanced distributions. However, also values based on domain knowledge can be imputed. In case of numeric attributes, mean or median imputation are basic methods, however, at the expense of under-representation of data variability and ignored correlation between attributes. [173]
These imputation approaches can be further extended to the imputation of mode, mean, median or domain knowledge values conditional to other attributes or the class of a data instance.

(c) *K nearest neighbors imputation*: The K most similar instances based on the known attributes are identified and mean, mode or median is computed from this neighborhood only. The applied distance measure and the selected number of neighbors K are key issues determining this imputation method. Moreover, this method requires sufficient data for reliable attribute modeling. [22, 266]

(d) *Hot deck imputation*: Considers only the most similar neighbor, a random valid value from the local neighborhood, or a single value based on another criterion [241, 173]. Hot deck imputation is a simple and widely used approach which is also proposed by [266] for patient data. However, this approach may suffer from the shortcoming that global properties are ignored when determining the imputation value [241].

(e) *Multiple imputation*: Missing values are replaced with a set of plausible but different versions of that attribute in order to maintain the variability of the missing values. In the following, either the average value is used or multiple versions of the dataset are available for further analysis. Suchlike, in contrast to single value imputation, uncertainty accompanied with data imputation is accounted for. Multiple imputation is recognized as the standard method to deal with missing data in many areas of research. [164, 294, 262]

(f) *Last Observation Carried Forward (LOCF)*: In case of sequential or time-series data, every missing value is replaced with the last observed value from the same subject. This method strongly assumes the values to remain unchanged and may underestimate the variability of the data. [164]

(g) *Interpolation or Curve fitting*: Also interpolation or curve fitting methods can be applied to estimate missing values in a sequence of continuous attributes. Curve fitting also has, to a limited extend, the potential to even extrapolate missing values to previous or future time points.

Adjacency-based imputation proposed by [10] for patient data sequences use linear interpolation to fill missing values between two known values and prior or posterior values for extrapolation. Those approaches require sufficient data points preceding or succeeding a missing entry.

(h) *Missing indicator*: An additional missing category for qualitative attributes and a fixed dummy value for quantitative attributes is imputed. This method is popular as it obtains the full dataset. However, even with few missing values and under the MCAR assumption this method is assumed to be subject to bias [262].

There are various more sophisticated imputation approaches proposed to estimate missing values, such as modeling the multivariate attribute distribution, exploiting correlations among attributes by training a regression model, or employing further machine learning-based procedures to build predictive models for value estimation. However, all of them rely on sufficient training data for attribute modeling. [173, 369, 359, 266, 241, 24]
Missing value analysis, i.e. the problem of imputing missing values in an incompletely specified data matrix is closely related to methods from RS research such as CF detailed in chapter 3. Many methods proposed in the related literature can also be applied for RSs and RS methods, such as latent factor models, were studied in the context of missing value analysis prior to application in RS [4].