Expert Systems With Applications 113 (2018) 147–160

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A framework for expert-driven subpopulation discovery and evaluation using subspace clustering for epidemiological data

Tommy Hielscher[a,*], Uli Niemann[a], Bernhard Preim[b], Henry Völzke[c], Till Ittermann[c], Myra Spiliopoulou[a]

[a] Otto-von-Guericke University Magdeburg, Department of Technical and Business Information Systems, Universitätsplatz 2, Magdeburg D-39106, Germany
[b] Otto-von-Guericke University Magdeburg, Department of Simulation and Graphics, Universitätsplatz 2, Magdeburg D-39106, Germany
[c] University Medicine Greifswald, Institute for Community Medicine, Walter Rathenau Str. 48, Greifswald D-17475, Germany

ARTICLE INFO

Article history:
Received 19 September 2017
Revised 11 June 2018
Accepted 2 July 2018
Available online 2 July 2018

Keywords:
Subpopulation discovery framework
Constraint-based subspace clustering
Cohort study data
Hepatic steatosis
Goiter

ABSTRACT

Objective: We propose an intelligent system that assists epidemiology experts in analysing the data of a population-based epidemiological study, in identifying relevant variables for an outcome and subpopulations with increased disease prevalence, and in validating the findings concerning variables and subpopulations in a further, expert-specified cohort. At present, the study of an outcome on a population-based cohort is hypothesis-driven, i.e. the expert must specify the variables to be studied. Our approach rather operates in a data-driven, semi-automated way, enabling the expert to identify variables of relevance and generate hypotheses on them.

Methods: Our system DIVA supports the *D*iscovery, *I*nspection and *VA*lidation of subpopulations with increased prevalence of an outcome, without requiring parameter tuning. DIVA takes as input the cohort of an epidemiological population-based study with *all* variables specified in the study's protocol, as well as inputs from the expert on the similarity of a small number of cohort participants. DIVA uses semi-supervised subspace clustering and subspace construction to identify sets of variables – subspaces – that promote participant similarity with respect to the outcome and with respect to the expert inputs, and then discovers subpopulations with increased outcome prevalence in those subspaces (DIVA module "DRESS"). DIVA uses visual analytics techniques to assist the expert in juxtaposing, filtering and inspecting the characteristics of these subpopulations (web-based DIVA module "D-INSPECTOR"). If the expert has access to a second cohort on a comparable population, DIVA aligns the cohort used for discovery to this second cohort, and then checks whether the subpopulations found in the original cohort are also present in the second one (DIVA module "VALIDATOR").

Results: We applied DIVA to the third wave (SHIP-2) of the SHIP-CORE cohort of the Study of Health in Pomerania (Völzke et al., 2011) for the liver disorder "hepatic steatosis", and on the first wave (TREND-0) of the SHIP-TREND cohort of the same study for the thyroid gland disorder "goitre". We found that most of the subpopulations extracted automatically, and subsequently ranked and filtered by the modules of DIVA, had significantly higher disease prevalence than the general population. We varied the amount of inputs needed from the expert to drive the subpopulation extraction process and found that a very small amount of information, namely the outcome of as few as 4 cohort participants, is sufficient for the identification of several relevant variables and subpopulations. We used a subset of TREND-0 for the validation on goitre and the complete TREND-0 for the validation on hepatic steatosis and found that the significant difference in prevalence for the identified subpopulation also holds in the validation data.

Conclusions: We have shown that DIVA discovers subpopulations and variables of importance with respect to an outcome, while requiring a very small amount of expert inputs. Each combination of variables and each subpopulation corresponds to a hypothesis, the validation of which would have required substantial

* Corresponding author.
E-mail addresses: tommy.hielscher@ovgu.de (T. Hielscher), uli.niemann@ovgu.de (U. Niemann), bernhard@isg.cs.uni-magdeburg.de (B. Preim), voelzke@uni-greifswald.de (H. Völzke), till.ittermann@uni-greifswald.de (T. Ittermann), myra@ovgu.de (M. Spiliopoulou).

https://doi.org/10.1016/j.eswa.2018.07.003

human effort. Thus, DIVA allows for a more effective exploitation of population-based data, not fully automated but driven by the expert and without the need for technical parameter tuning.

A shortcoming of DIVA design is the demand of a specific type of expert inputs, namely "constraints" on the similarity of pairs of participants. Currently, we generate the constraints with a naive utility that is based on random sampling, but we work on the development of an interactive algorithm that would allow the epidemiology expert to inspect a small choice of study participant and give statements on their similarity.

The present version of DIVA considers a single wave of the cohort data, ignoring the evolution of the population during the horizon of the study. Hence, subspace and subpopulation discovery do not take account of changes in the importance of variables. We currently work on the incorporation of algorithms that derive additional variables from the longitudinal data and use them in the Discovery module.

1. Introduction

Researchers in epidemiology collect population-based cross-sectional and longitudinal cohort data, from which they strive to derive insights on pathogenesis, disease pathways and responses to different kinds of treatments (Preim et al., 2016). Similarly to randomized clinical trials, a study on a population-based cohort involves an in/exclusion protocol and a carefully specified set of variables, whose impact on the outcome (e.g. on a disease) is to be investigated. Unlike randomized clinical trials, for which a cohort is prospectively recruited, studies on a population-based cohort are retrospective: the cohort has already been recruited with a protocol that typically involves a substantially larger number of variables. For example, the original protocol of the Study of Health in Pomerania encompassed 8854 variables (Völzke et al., 2011): analysis in such a high-dimensional space is prone to the curse of dimensionality, hence methods for focussed analysis in subspaces are necessary. In this study, we propose DIVA, an intelligent system with which the epidemiologist can semi-automatically, with minimal interaction, identify subsets of variables with potential relevance to a given outcome, can study automatically derived subpopulations that are described by these variables and exhibit considerably higher or lower prevalence of the outcome, and can juxtapose the significance of the prevalence difference in a validation cohort.

The task of supporting epidemiology experts with intelligent IT is being intensively investigated for years. Thew et al. (2009) elaborate on instruments with which epidemiologists can express and share domain knowledge among themselves. Such instruments are designed for hypothesis refinement. However, hypothesis generation, which comes before refinement, calls for decision with respect to (w.r.t.) the selection of the variables to be taken into account.

The selection of a small number of variables for hypothesis formulation from a huge set of variables has been studied by Zhang, Gotz, and Perer (2014) in the context of "cohort specification" from Electronic Health Records. Their system CAVA contains an interactive mechanism with which a clinical expert can select variables manually and study their impact on the outcome, as well as modules for automated management of the data in databases and for machine learning. Further systems in this category include SeekAView (Krause, Dasgupta, Fekete, & Bertini, 2016a), INFUSE (Krause, Perer, & Bertini, 2014) and PROSPECTOR (Krause, Perer, & Ng, 2016b), all of which include utilities for interactive variable selection before machine learning. However, the manual selection among hundreds or thousands of variables seems rather restrictive, since it enforces the expert to concentrate on the variables whose impact on the outcome is known or expected.

In this study, we propose a semi-supervised, self-tunable intelligent system that automates the construction of sets of variables potentially worth exploring, discovers subcohorts characterized by these variables, assists the expert in inspecting them, and validates them automatically in an independent cohort, if any is available. Our system DIVA consists of following modules:

- *Discovery* module: Our core algorithm DRESS + discovers subpopulations by exploiting little background knowledge in the form of pairwise constraints that contain knowledge about the similarity between study participants. Thus, the algorithm avoids the necessity of large quantities of labeled data by utilizing knowledge that can be derived from a limited set of labels or provided by a medical expert.
- *Inspection* module: Our interactive web application D-Inspector provides means to analyze the discovered subpopulations, i.e., juxtapose multiple subpopulations, study the distribution of corresponding variables w.r.t. the medical outcome, and query the set of subpopulations by custom filtering and sorting functionalities.
- *Validation* module: Our Validator checks to what extent the clusters and subspaces found by DRESS + can be reproduced in an independent cohort.

The paper is organized as follows. In the next two sections we discuss first related work and then basic underpinnings of the intelligent, semi-supervised technologies we use. In Section 4 we describe the three components of our approach. In Section 5 we present our results for the disorders fatty liver (hepatic steatosis) and goiter, using data from two cohorts of the Study of Health in Pomerania (SHIP) (Völzke et al., 2011). We close the paper with a discussion and outlook in Section 6.

2. Related work

Relevant literature for our work encompasses advances on intelligent systems for the support of the human expert in the medical domain, as well as advances on the functionalities covered by the modules of our system DIVA. We discuss them hereafter.

2.1. Interactive intelligent systems for cohort analysis

Without providing an integrated workflow to validate and inspect the findings, medical researchers remain skeptical towards the machine learning methods. Cummins (2012) describes that much criticism of the medical community on data mining models is due to the contrast between the sequential process of traditional medical research and the iterative and interactive approach of KDD procedures. To overcome this criticism, scholars should consider (i) involving domain experts into the model generation, (ii) assessing the model's quality by applying it on unseen data, and (iii) calibrating the model for different target populations (Cummins, 2012). Most of the systems described hereafter focus on expert involvement in requirement (i), while quality assessment is incorporated

in the model learning phase. As we explain at the end of this subsection, the *Validation* module of our system DIVA covers the last two requirements.

Intelligent systems that support the expert for interactive cohort analysis include the visual cluster analysis system of Gotz, Sun, Cao, and Ebadollahi (2011), the systems CAVA (Zhang et al., 2014), INFUSE (Krause et al., 2014), PROSPECTOR (Krause et al., 2016b), SeekAView (Krause et al., 2016a) and our earlier system IMM (Niemann, Völzke, Kühn, & Spiliopoulou, 2014) and its extensions (Niemann, Spiliopoulou, Preim, Ittermann, & Völzke, 2017; Schleicher, Ittermann, Niemann, Völzke, & Spiliopoulou, 2017).

Gotz et al. (2011) present a system that takes as input a patient's Electronic Health Record (EHR) as query and identifies subcohorts that are similar to that patient. After an automatic cluster analysis, experts can split and merge found subcohorts at will to refine the generated results. Authors recognize that "Refinement is valuable because cluster analysis algorithms detect statistical patterns, often with little or no a priori semantic knowledge." Instead of requiring the expert to manually refine mediocre clusters in hindsight, DIVA exploits partially labeled data already during the clustering process to obtain useful results right away.

Zhang et al. (2014) propose CAVA, a system for interactive cohort construction and automated learning on EHR. CAVA encompasses tools for interactive refinement of the cohorts and for subcohort customization, including powerful visualization techniques. Machine learning is automated; the expert can also choose among predefined workflows. In contrast, the step of cohort construction, including the choice of variables to be included in the analysis, is the task of the expert. Our proposed system DIVA also encompasses automated analysis of the cohort data, but the main emphasis is on the semi-automated, rather than manual selection of the variables for the analysis. Moreover, unlike the aforementioned two systems, we analyse existing cohorts of population-based studies and not EHR, hence cohort construction from EHR is beyond our scope.

Similarly to DIVA, the systems INFUSE (Krause et al., 2014), PROSPECTOR (Krause et al., 2016b) and SeekAView (Krause et al., 2016a) encompass powerful mechanisms for variable selection as part of the cohort preparation for the analysis. However, the emphasis of these mechanisms is on supporting the expert, who is called to choose the most promising variables. In contrast, DIVA uses subspace discovery for the identification of promising variables, thereby exploiting expert inputs in a semi-automated, rather than manual fashion.

In an earlier work, we introduced the Interactive Rule Miner (Niemann et al., 2014), which encompasses a module for automated classification, as well as an interactive tool for the inspection of classification rules. This system has been designed for the analysis of cohorts and has been used to identify high-risk subpopulations w.r.t. a medical outcome in epidemiological data. The extensions presented in (Niemann et al., 2017; Schleicher et al., 2017) focus on pattern inspection after the automated pattern discovery: Schleicher et al. (2017) propose a workflow for pattern drill-down and the visualization of subpopulations characterized by one or two variables. Niemann et al. (2017) propose mechanisms for clustering the discovered patterns (classification rules), identifying cluster representatives and visualizing them. These systems consider *all* variables of the cohort to be analysed and rely on the learning mechanism for the identification of potentially interesting subpopulations. This has the shortcoming that the learning mechanisms, which are either subgroup discovery Atzmüller (2015); Herrera, Carmona, González, and Del Jesus (2011); van Leeuwen and Knobbe (2012) or conventional classification rules Niemann et al. (2014), produce a very large number of patterns, thus calling for functions that can assess the "interestingness" of the found subpopulations (Atzmüller & Puppe, 2006; Grosskreutz, Rüping, & Wrobel, 2008; van Leeuwen & Knobbe, 2012). The system we propose, DIVA, reduces the input set of variables in a semi-supervised way, and thus allows for a broader choice of algorithms for the machine learning step.

Moreover, epidemiologists are interested in finding generalizable results that hold true also beyond the small fraction of the population they study. Confirming findings on independent cohorts is mandatory to verify their significance. To the best of our knowledge, our framework DIVA is the first to employ automatic validation capabilities when an independent cohort dataset is given by approximating the region of the feature space of discovered subpopulations in a second dataset, and comparing the outcome distribution of the original subpopulation with the approximated one. These subpopulations can be of arbitrary shape through the use of a density-based cluster definition. DIVA neither explicitly (e.g. Niemann et al. (2014) with hyper-rectangular shaped rules), nor implicitly (e.g. Gotz et al. (2011) by scoring the quality of clusters with centric-based measures) assumes that subpopulations are bound to have a specific shape. At last, our framework does not require any inputs from the expert other than instance-level constraints. Parameters at the algorithmic level are calculated with the help of the data density within subspaces of the feature space to gain one reasonable clustering per subspace.

2.2. Supervised feature selection

When large quantities of labeled data are available, feature selection methods could be used to derive relevant dimensions. Classical feature selection approaches are either wrapper-, filter- or embedded methods striving to find subsets of weakly associated features that are highly associated to the target variable (highly relevant and not redundant) (Guyon, Weston, Barnhill, & Vapnik, 2002; Hall, 2000; Kohavi & John, 1997). In DIVA, we aim to minimize the number of labeled instances that are used for feature selection, because these instances must be excluded from the subsequent task of machine learning. We therefore focus on semi-supervised methods on the basis of subspace construction.

2.3. Automated and semi-automated mechanisms for subspace construction

The discovery module of our framework utilizes subspace clustering with constraints to identify subpopulations and relevant features w.r.t. a medical outcome. In subspace clustering the aim is to detect clusters within subsets of the original dimensions (Parsons, Haque, & Liu, 2004; Sim, Gopalkrishnan, Zimek, & Cong, 2013). Kailing et al. present a general subspace search procedure with the algorithm RIS (Kailing, Kriegel, Kroeger, & Wanka, 2003). RIS uses density-based bottom-up subspace clustering and a custom unsupervised quality function based on density properties to score and rank subspaces. In RIS the structure of the subspace is considered during subspace evaluation without taking external knowledge into account. Contrary to RIS, the DIVA framework exploits background knowledge in the form of constraints that provide information regarding the similarity of pairs of study participants.

When using constraints in clustering algorithms the aim is to find clusters such that the different kinds of constraints are satisfied (Ruiz, Spiliopoulou, & Menasalvas, 2007). They reflect limited background knowledge and may guide the clustering algorithm in finding clusters by adjusting the objective function or learn a custom metric over the dataset.

Because constraints can be defined by domain experts, constraint-based methods fit well to medical scenarios. For example, Liu et al. report on an application of constraint-based clustering to remove the negative impact of confounding factors and

thus to find clinically relevant groups of multiple sclerosis patients (Liu, Brodley, Healy, & Chitnis, 2015).

Subspace clustering algorithms that exploit constraints are Constrained K-Means (Wagstaff, Cardie, Rogers, & Schrödl, 2001), SMVC (Günnemann, Färber, Rüdiger, & Seidl, 2014) and DRESS (Hielscher, Spiliopoulou, Völzke, & Kühn, 2016). Constrained K-Means uses instance-level constraints to assign an instance x to the closest cluster so that the assignment decision does not lead to a violation of the constraints related to x. SMVC models multiple clusterings and related dimensions (views) of the data with the help of a Bayesian framework. Because views define different, alternative clusterings of the data, the algorithm does not necessarily drop views that violate many constraints. These methods exhibit some undesired characteristics w.r.t. our application setting, e.g. prior assumptions about number and shape of subpopulations within subspaces or the focus on alternative clusterings. An algorithm we extend within DIVA is Discovery of Relevant Example-constrained SubSpaces (DRESS) (Hielscher et al., 2016). DRESS evaluates subspaces according to the similarity of contained objects and found clusters. Here, subspaces which might be relevant regarding the medical outcome are scored and ranked according to a custom quality function that considers the distance and cluster membership of study participants under different kinds of constraints. However, clusters cannot be further analyzed and extracted. We therefore extend DRESS to return clusters that are associated with relevant subspaces and incorporate it into a fully-fledged framework for further inspection and validation of the discovered subpopulations.

3. Foundations

Our framework is based on principles of subspace clustering and constraint-based clustering to find groups of similar participants w.r.t. the medical outcome. Because cohort study data is often high-dimensional, they may dilute clusters that can only be found in lower-dimensional spaces. Dedicated subspace clustering methods are required to detect the groups of proximal data points in spaces that contribute most to their similarity.

In context of cohort study dataset D with associated feature set F, we denote any subset $S \subseteq F$ as a subspace. For each participant $x \in D$, their respective projections onto S are denoted as $\pi_S(x)$ and the set of participant projections is D_S. Then, a goal of subspace clustering is the discovery of clusters $C \subseteq D_S$ in one or more S.

While useful at first glance, pure subspace clustering is not feasible when looking for potential subpopulations. Without any guidance of the subspace search, any groups of participants that are similar in a number of features are found. This leads to an enormous result space and complexity. For example, groups of participants with similar shoe size and height may be identified regardless of any association with the medical outcome.

To guide the subspace search algorithm w.r.t. the medical outcome our framework also incorporates principles of constraint-based clustering. We adopt two kinds of instance-based constraints for our purpose: must-link (ML) and cannot-link (NL). Such constraints are used to provide knowledge regarding the similarity of objects. In our framework, we define a ML and NL constraint as a set of two participants $\{x, y\} \subseteq D$, with $x \neq y$ and the respective constraints supersets as ML and NL. Further, if a clustering of n clusters $C = \{C_1, \ldots, C_n\}$ is discovered in D, then C satisfies a ML constraint $\{x, y\}$ if x and y are member of the same $C_i \in C$, i.e. $\{x, y\} \subseteq C_i$ holds true. On the other hand, clustering C satisfies a NL constraint $\{x, y\}$ if x and y are not member of the same cluster, i.e. $\neg \exists C_i \in C$: $\{x, y\} \subseteq C_i$ holds true.

Assume a study participant dataset and the target concept diabetes: Here, a ML constraint could be given between a pair of participants with diabetes and a NL constraint between a pair of participants with and without diabetes. Then, the expected result of a constraint-based subspace clustering algorithm is a subspace consisting of informative features and clusters w.r.t. the separation between participants with and without diabetes, that is a subspace consisting of several groups of similar participants which share some key characteristics associated to their diabetes status.

4. Methods

Building on the aforementioned methodologies, the goals of our framework are as follows:

(1) Given cohort dataset D and a set of ML/NL constraints, find groups of participants within subspaces which best describe the concept, as reflected in the constraints, where “best” refers to participant similarity/separation and constraint satisfaction.

(2) Given these groups (subpopulations), provide ways to identify and analyze the most distinct ones w.r.t. to the medical outcome.

(3) Enable experts to investigate whether discovered subpopulations are generalizable or not.

Fig. 1 shows the framework DIVA of our proposed approach to *D*iscover, *I*nspect and *VA*lidate subpopulations in cohort study data leveraging constraint-based subspace clustering methodologies. The complete DIVA framework uses two cohort study datasets as input which share the same feature space: one dataset for the discovery of subpopulations, hereafter denoted as discovery dataset DD, and one dataset for the validation of discovered subpopulations, hereafter denoted as the validation dataset VD. During the preprocessing step an initial matching of DD and VD on covariates is executed, to provide the discovery module (Fig. 1(A)) with the matched discovery dataset and the validation module with the matched validation dataset.Then, within the discovery module the DRESS + algorithm uses constraints to guide a bottom-up subspace search and clustering technique to discover sets of promising clusters representing subpopulations which are potentially associated with a medical outcome. After selecting the clusters with highest quality, the inspection module (Fig. 1(B)) exploits the application D-Inspector that provides mechanisms to investigate and juxtapose the subpopulations, and the validation module (Fig. 1(C)) checks whether subpopulations are existent in independent cohort study data (VD) and similarly associated with the medical outcome.

Note that in the case that a fully independent validation dataset is not available, experts may split their existing dataset in a discovery set and validation set using a suitable sampling strategy or perform stratified cross-validation.

In the next sections we will describe example methods how to generate constraints, the preprocessing step and each module of DIVA.

4.1. Constraints generation

There are many ways experts can generate constraints. Constraints should reflect similarity w.r.t. the medical outcome under study between participants. One of the simplest method to generate such constraints is to ask the expert to label a random number of healthy and ill participants and then automatically generate a ML constraint between each pair of labeled participants with the same medical outcome, and a NL constraint between each pair of labeled participants with different medical outcome. As an alternative, assume that a number of labeled instances already exists. The expert is presented with two instances, x and y of different but known class and a third instance z. Then, the expert is asked to assign z to either x or y, the one that is in his opinion more similar to z. This can be repeated multiple times and from each

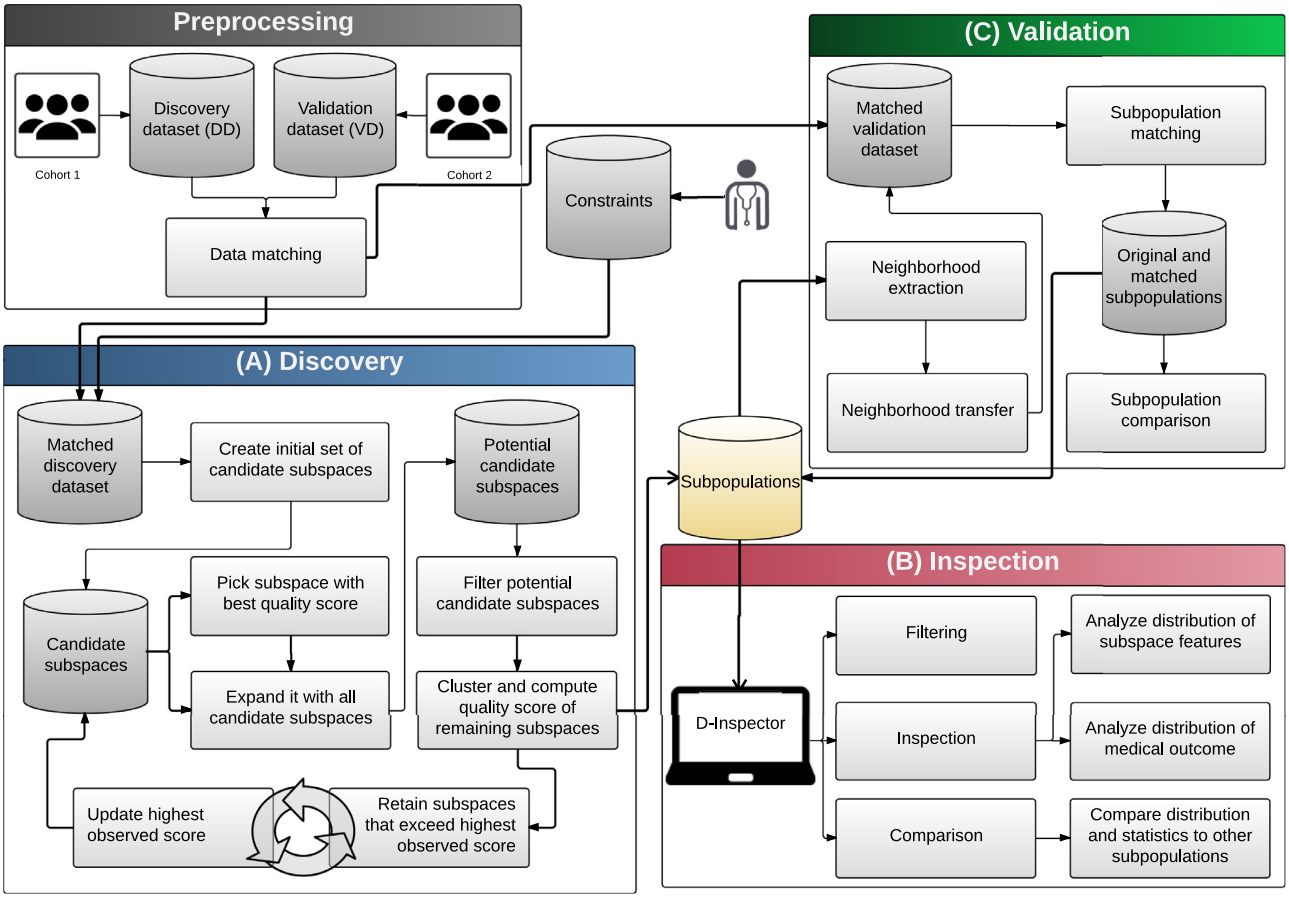

Fig. 1. The DIVA framework: subpopulation discovery, inspection and validation.

expert assignment a ML constraint may be derived, as well as a NL constraint from the preselected instances of different class.

Basically, there is no strict protocol an expert has to follow in order to derive a number of constraints. All that is necessary is the decision of the expert on what he deems to be similar w.r.t. the medical outcome and the study participants.

4.2. DIVA Preprocessing

In the case that a validation dataset *VD* is available to DIVA, the first step of the framework is to match *VD* and *DD* to enable fair comparisons between subpopulations found in *DD* and their counterparts within *VD*. The preprocessing step accounts for significant differences in covariates between *DD* and *VD*, e.g. the participants' age and sex. It reduces any potential bias of the covariates on the validation measure score by employing nearest neighbor propensity score matching (Ho, Imai, King, & Stuart, 2011). First, the original *DD* and *VD* are combined to learn a logistic regression model where the dichotomous target variable indicates whether a participant is member of *DD* or *VD*. For each participant, the model calculates a probability (propensity score) for dataset membership. Each *DD* participant is then *matched* to the *VD* participant with the most similar propensity score yielding two datasets with the same number of participants: First, all participants in *DD* and *VD* are unflagged. Then, in each iteration one unflagged *DD* participant x is selected. If for x no unflagged *VD* participant falls within the maximum propensity score threshold, x is removed from *DD*. Otherwise, x is flagged and the respective *VD* participant is flagged. After each retained *DD* participant has been flagged, each unflagged *VD* participant is removed. Output of the preprocessing step is the set of remaining participants in *DD* (matched discovery dataset) and the set of remaining participants in *VD* (matched validation dataset). The preprocessing step requires the expert to specify a maximum propensity score distance threshold. A higher threshold leads to a larger number of matched participants, whereas the similarity w.r.t. the covariates might be lower than for a lower threshold.

4.3. DIVA Discovery module

The first module of DIVA discovers interesting subpopulations w.r.t. a specific concept (i.e. a medical outcome) without requiring large quantities of labeled data. To do so, we use DRESS + , which extends our constraint-based subspace clustering algorithm DRESS (Discovery of Relevant Example-constrained SubSpaces) (Hielscher et al., 2016).

We assume that if given constraints reflect the separation between the medical outcome of participants rather well, then we can find subpopulations in the data (natural clusters) for whom we can generalize this knowledge as well as the variables (subspaces) that contribute most to this separation. DRESS + translates this assumption by searching for subspaces where ML-constrained participants have high pairwise similarity and NL-constrained participants exhibit high pairwise dissimilarity.

The general workflow of DRESS + is shown in Fig. 1 (A). As input, DRESS + requires the study participant data from which subpopulations should be derived as well as a set of ML and NL constraints between pairs of participants. Constraints can be given by a medical expert, stating whether two individuals are similar or different regarding the medical outcome. For example, if the goal is to find subpopulations of study participants that exhibit hep-

atic steatosis, a medical expert may define ML constraints between participants that are known to have hepatic steatosis and NL constraints between participants with and without hepatic steatosis. As an alternative, constraints may also be derived from small quantities of labeled data (for example between same- and different-class participants).

Using such constraints, DRESS + executes a forward-selection (starting with subspaces of cardinality one) to find subspaces that contain relevant clusters, i.e. finding subspaces that score better than previously visited subspaces according to its underlying quality function. In each iteration the "best" subspace $S_{candidate}$ is chosen and removed from the set of candidate subspaces. Then, DRESS + extends $S_{candidate}$ with the remaining candidates to create the set of potential candidate subspaces (merging step). To reduce complexity, the filtering step drops subspaces that are unlikely to exceed the current quality threshold. Afterwards, the algorithm computes the complete quality scores of the remaining potential candidate subspaces by clustering. If the quality of a subspace exceeds the highest yet observed quality q_{best}, DRESS + retains it as a candidate subspace for further extension, updates q_{best} and stores all contained clusters. Each time a successful subspace merge yields a new candidate subspace, the subspaces used for merging are removed from the candidate set. DRESS + terminates as soon as the candidate subspace set is empty and returns the set of all stored clusters. The quality computation and clustering are adopted from the original DRESS algorithm. They are explained in Section 4.3.1 and Section 4.3.2 based on the work presented in (Hielscher et al., 2016). "Merging" and "Filtering" are extended from the contribution in Hielscher et al. (Hielscher et al., 2016) to allow the storage of clusters and are explained in Section 4.3.3.

4.3.1. Scoring subspace quality

DRESS + evaluates subspaces in a semi-supervised way as it investigates the structure of the complete study data in the subspaces and only requires a small number of instance-level constraints , i.e. background information w.r.t. the medical outcome from a small number of study participants. The quality function used to score (and rank) candidate subspaces in DRESS + is based on the clustering results within the respective subspace and takes the following criteria into account: (1) whether the data exhibits a satisfactory structure within a space w.r.t. the given constraints and

(2) the proximity between constrained participants in the subspace. Criterion (1) is necessary to drop spaces without satisfactory clusters. DRESS + assumes that if the dimensions of a space are associated with one or more relevant subpopulations, a corresponding cluster in this space can be found that separates study participants of this subpopulation from the remaining participants. These clusters should be distinct, meaning that reasonable parameter settings of the clustering algorithm should lead to their detection, i.e. no parameter optimization should be necessary. Because the constraints reflect background knowledge regarding the similarity of the participants' medical outcome, good subspaces should contain clusters with objects under ML-constraints, and separate objects under NL-constrains (putting them in different clusters). For (1), DRESS + calculates how the participant clusters satisfy the provided constraints: Given subspace S, let $ML_{sat}(S)$ be the set of satisfied ML constraints and $NL_{sat}(S)$ be the respective set for the NL constraints. Then

$$q_{cons}(S) = \frac{|ML_{sat}(S)| + |NL_{sat}(S)|}{|ML| + |NL|} \tag{1}$$

defines the constraint satisfaction within S.

By exclusively relying on q_{cons}, DRESS + faces the problem of ignoring the continuous similarity between the constrained participants. Imagine a number of spherical clusters in two different subspaces that satisfy the same constraints, where in one subspace the ML constrained participants have almost identical feature values and in the other subspace these participants lie on opposite border regions of the same cluster. In this scenario, q_{cons} for both subspaces is identical. But in reality the first subspace should be scored better than the latter. This problem is avoided by accounting for criterion (2) and defining a q_{dist}, that incorporates the similarity between participants under constraints in a subspace through distance calculations. We define q_{dist} as the difference in the average distances between NL and ML pairs of participants in the respective subspace: Let $d(S, x, y)$ be the distance between participant x and y in S. DRESS + computes q_{dist} of S as

$$q_{dist}(S) = d_{avg}(S, NL) - d_{avg}(S, ML), \tag{2}$$

$$\text{with } d_{avg}(S, NL) = \sum_{\{x,y\} \in NL} d(S, x, y)/|NL|,$$

$$\text{and } d_{avg}(S, ML) = \sum_{\{x,y\} \in ML} d(S, x, y)/|ML|.$$

Eq. 2 promotes subspaces where ML constrained participants are closer (i.e. more similar) to each other in comparison to NL constrained participants. However, to compute q_{dist} we must deal with the heterogeneity of the cohort study data. This is accomplished by utilizing the Heterogeneous Euclidean Overlap Metric (HEOM) as distance function. HEOM can deal with continuous and nominal features (Wilson & Martinez, 1997): Let $s(x)$, $s(y)$ be the values of feature $s \in S$ for participants $x, y \in D_S$. Then:

$$d(S, x, y) = \sqrt{\sum_{s \in S} \delta(s(x), s(y))^2} \tag{3}$$

where

$$\delta(s(x), s(y)) = \begin{cases} 0 & \text{if } s(x) = s(y) \text{ and } s \text{ is nom.,} \\ s(x) - s(y) & \text{if } s \text{ is continuous,} \\ 1 & \text{otherwise.} \end{cases}$$

Note that when using Eq. 3, continuous features must be normalized into interval [0,1] beforehand. The final quality function of DRESS + for subspace scoring is then given as

$$q(S) = q_{cons}(S) \cdot q_{dist}(S). \tag{4}$$

4.3.2. Clustering

DRESS + uses the density-based clustering algorithm DBSCAN (Ester, Kriegel, Sander, & Xu, 1996). DBSCAN exhibits a number of advantages for the application on cohort study data: outliers which might have special characteristics regarding the medical outcome can be detected. Additionally, the number and size of clusters is not required to be specified as parameter. Further, DBSCAN is not limited to linear boundaries but discovers arbitrarily shaped clusters, so that no assumptions on the variable distributions within a subpopulation have to be made. DBSCAN identifies dense regions in the data space and marks them as clusters. It uses the parameters *eps* and *minPts*. The *eps* parameter defines the neighborhood around a data point/study participant, and *minPts* is the minimum number of neighbors a point must exhibit to be considered a core point. The neighborhoods of each set of inter-connected core points define a dense region and form a cluster. For each (potential) candidate subspace S, DRESS + invokes DBSCAN on D_S. To perform clustering, DRESS + exhibits self-tuning capabilities by automatically calculating a set of "reasonable" parameter settings. The *minPts* parameter is fixed to $minPts = round(ln(|D_S|))$, according to (Ester et al., 1996). Given the fixed *minPts*, the *eps* parameter is calculated separately for each S by using the "knee-point" method (Niemann, Hielscher, Spiliopoulou, Völzke, & Kühn, 2015):

Given $m = minPts$, let $nn(x, m, S)$ denote the m-nearest neighbor of participant x in D_S. First the distance $d(S, x, nn(x, m, S))$ of each participant $x \in D_S$ to its m-nearest neighbor is computed. These distances define the necessary *eps*-value regarding each specific participant to be considered as core point in S. Let $(x_1, \ldots, x_n)$ be the ordered list of participants according to their m-nearest neighbor distance (in ascending order) and $i \in \{1, \ldots, n\}$ denote the position of each participant within the list. DRESS + creates an empty two-dimensional real-valued space and inserts one datapoint per participant, i.e. $\forall i \in \{1, \ldots, n\}$ the algorithm inserts the point $z_i = (i - 1, d(S, x_i, nn(x_i, m, S)))$. The resulting graph is denoted as the m-dist graph. After that, a line $line(z_1, z_n)$ between the points $z_1 = (0, d(S, x_1, nn(x_1, m, S)))$ and $z_n = (n - 1, d(S, x_n, nn(x_n, m, S)))$ in the m-dist graph is computed (the points associated with the first list element x_1 and last element x_n). and $\forall i \in \{1, \ldots, n\}$ the shortest euclidean distance, denoted as $Euclid(line(z_1, z_n), z_i)$, between $line(z_1, z_n)$ and z_i is calculated. Finally, the *eps*-value is set to $eps = d(S, x_i, nn(x_i, m, S))$, correspondent to participant x_i with the highest $Euclid(line(z_1, z_n), z_i)$ (knee-point), i.e. DRESS + chooses the $d(S, x_i, nn(x_i, m, S))$ as the *eps* value that belongs to the participant that maximizes the shortest distance between $line(z_1, z_n)$ and its m-dist graph point.

For the given *minPts*, this heuristic ensures that DBSCAN detects "unique" participants (exhibiting a relative high distance to others) in sparse regions of the data space and flags them as outliers while preserving a number of dense regions (clusters).

4.3.3. Merging and filtering of subspaces for cluster discovery

DRESS + initializes the set of candidate subspaces with all spaces of cardinality one and the empty set of subspace clusters as depicted in Algo. 1. Let F be the set of all features in the original study participant dataset D, the initial candidate set $\mathcal{S}$ is defined as $\mathcal{S} = \{S | S \subseteq F \wedge |S| = 1\}$. During initialization, the quality $q(S)$ of each subspace $S \in \mathcal{S}$ is calculated, stored and all resultant clusters are saved in $\mathcal{C}$. For the general subspace candidate generation and filtering process see Algo. 2: Here, q_{best} is initialized as the highest observed quality among the initial set of candidate subspaces (which is given by Algo. 1). Then, DRESS + iteratively chooses the subspace $S_{candidate} \in \mathcal{S}$ that has the highest quality $q()$ among the subspaces in $\mathcal{S}$ and sets $\mathcal{S} = \mathcal{S} \setminus S_{candidate}$ (cleaning). DRESS + builds new potential candidate spaces $S_{new} = S_{candidate} \cup S^*$ by merging $S_{candidate}$ with all remaining subspaces $S^* \in \mathcal{S}$ in an effort to find subspaces with a $q()$ that beat q_{best}. To limit the number of subspaces that must be fully scored (calculating both q_{cons} and q_{dist}), DRESS + uses a filter condition that prevents the clustering of merged subspaces that probably do not contribute to a improvement w.r.t. the full quality $q()$: For each new potential subspace candidate S_{new} DRESS + picks the subspace $S_{dist} \in \{S^*, S_{candidate}\}$ with higher q_{dist},

$$S_{dist} = \begin{cases} S_{candidate}, & \text{if } q_{dist}(S_{candidate}) \geq q_{dist}(S^*), \\ S^*, & \text{otherwise,} \end{cases}$$

and performs clustering in S_{new} iff the following condition is satisfied:

$$(\Delta(S_{new}, S_{dist}, NL) - \Delta(S_{new}, S_{dist}, ML)) > 0,$$

with $\Delta(S_{new}, S_{dist}, X) = d_{avg}(S_{new}, X) - d_{avg}(S_{dist}, X)$,

This condition checks whether the dissimilarity in the new space between NL constrained participants increases more than for ML constrained participants, i.e. if the space leads to a better separation between them. If a new potential candidate subspace S_{new} satisfies this condition, DRESS + commences with clustering in it, computing its complete quality $q()$ in the process, and stores all found clusters $\mathcal{C}_S$ by setting $\mathcal{C} = \mathcal{C} \cup \mathcal{C}_{D_{S_{new}}}$. After the full quality assessment, if it is found that $S_{new} = S_{candidate} \cup S^*$ satisfies $q(S_{new}) > q_{best}$, DRESS + sets $\mathcal{S} = \mathcal{S} \setminus S^*$ to eliminate the lower quality subspace (preventing further merging with it) and adjusts the current best quality value to $q_{best} = q(S_{new})$. Newly discovered high-quality subspaces are inserted in the candidate set $\mathcal{S}$ for further merging, setting $\mathcal{S} = \mathcal{S} \cup \{S_{new}\}$. DRESS + terminates when the candidate set is empty, i.e. no S_{new} with higher quality than the current q_{best} is found.

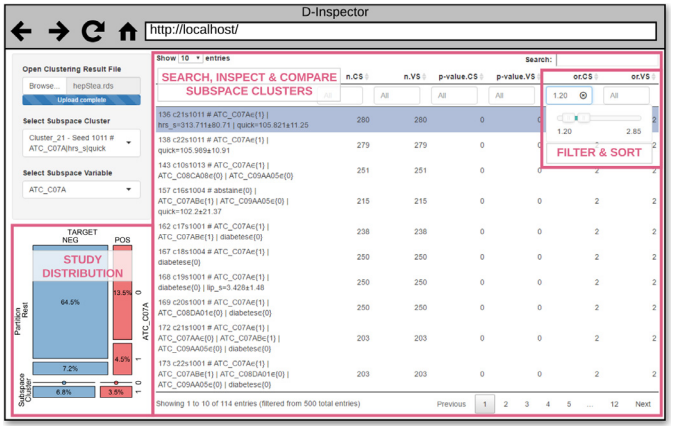

Fig. 2. Screenshot of D-Inspector. Using the sidebar panel on the left, a clustering result file can be loaded. The distribution of a selected subspace cluster variable in comparison with the remaining instances is visualized by a mosaic chart in the bottom left. Users can sort and filter the set subspace clusters by size, χ^2-test p-value, odds ratio or custom queries to search for variables or cutoffs of interest. The results on both *DD* and *VD* are displayed.

4.4. DIVA Inspection module

The inspection part of our framework is used to explore any identified subpopulation and is depicted in Fig. 1 (B). For this we developed the web application D-Inspector that provides various means to get useful insights on the clusters found by DRESS + . Fig. 2 shows the main view of D-Inspector. Here, each subpopulation is described by the number of included participants and the associated feature space (including mean and standard deviation for continuous and distinct values for nominal features). Users can access statistics of the subpopulation w.r.t. the medical outcome which includes size, χ^2-test p-value and odds ratio. In the case of a validated subpopulation, the statistics between the cluster found in *DD* and the matched cluster within *VD* can be compared. Various filter mechanisms allow to reduce the number of shown subpopulations to the most interesting ones. They include filtering according to a subpopulation's desired size, medical outcome statistics and whether the comprising features of a subpopulation are distributed significantly different compared to the remaining participants. As shown in Fig. 2,D-Inspector enables experts to further explore subpopulations by providing methods to analyze and compare the distribution of each comprising feature with boxplots (for continuous features) and mosaic charts (nominal features) to the respective distribution within the complete population.

D-Inspector is build upon the R web application framework Shiny[1] and uses the DT package, an R interface to the JavaScript library Datatables[2] which adds sorting, filtering and server-side processing functionality to HTML tables.

[1] http://shiny.rstudio.com/

[2] https://datatables.net/

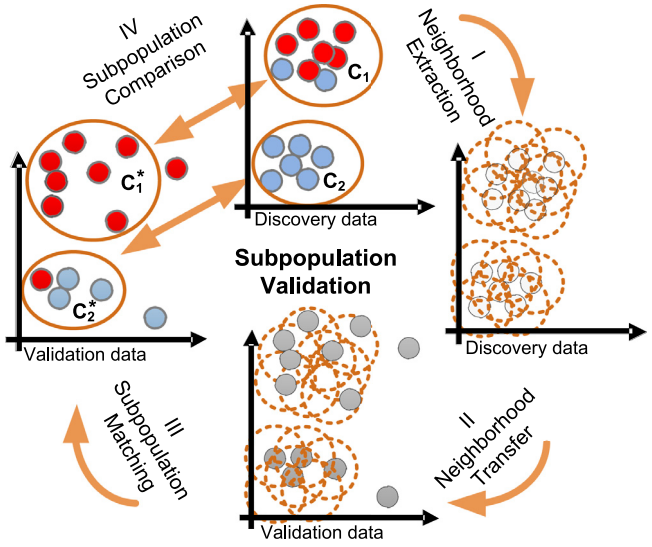

Fig. 3. Tasks involved for validating subpopulations.

4.5. DIVA Validation module

A desirable goal of medical research is that acquired insights are generalizable (Cummins, 2012). The validation module of DIVA investigates whether subpopulations identified by the discovery module are existent in independent cohort study data. DIVA transfers subspace clusters found on one dataset (*DD*) on a second dataset (*VD*), to ultimately assess the agreement w.r.t. class distribution and significance. The general workflow of the involved tasks for validation is depicted in Fig. 3. For each subpopulation C found in *DD*, the following procedure is executed individually: First, the neighborhoods are extracted from the core points of C by storing the position of each core point, the *eps* parameter of the clustering that produced C and the subspace S where C was found. Then, these neighborhoods are transferred by projecting *VD* onto S, thus creating VD_S, and inserting into VD_S one dummy point p on each position of an original core point in C. Following this, the matched subpopulation C^* is created: For each *VD* participant x the module checks if there exists at least one dummy point p where the distance between x and p is less or equal than *eps*, i.e. $d_S(x, p) \leq eps$. All *VD* participants that satisfy this criterion form the cluster counterpart C^* of C. At last, for each C, it's class distribution and significance is compared to C^* and presented by the validation module.

5. Experiments & results

In this section we show findings and evaluate the transferability of our framework's results, i.e. whether identified subpopulations can be discovered in independent datasets. We further evaluate the impact of different DRESS + parameter settings on the results and provide an overview about relevant examples of subpopulations. For this we utilize two real-world cohort study datasets on two different medical conditions.

5.1. Cohorts data

Analyzes are based on data from two independent cohorts of the Study of Health in Pomerania (SHIP), conducted in Northeast Germany (Völzke et al., 2011). In the first cohort (originally called SHIP, later called SHIP-CORE to distinguish from the second cohort), 4308 individuals participated in the baseline examinations (SHIP-0) between 1997 and 2001. In the present analysis we used data from the second follow-up SHIP-2, in which 2333 individuals aged 30–93 years were examined between 2008 and 2012. Parallel to SHIP-2, a second cohort (SHIP-TREND) was established in the same study region where 4420 individuals aged 20–84 participated in the baseline examinations (TREND-0).

Table 1
Characteristics of the cohorts SHIP-2 and TREND-0.

	SHIP-2	TREND-0
Participants	1878	4400
Variables	70	492
Age [years]	58.1 ± 13.5	51.0 ± 14.1
Sex [% female]	53.3%	51.7%
Hep. Stea.	163 pos., 564 neg.	462 pos., 1464 neg.
Goiter	-	1390 pos., 3010 neg.

Table 2
Matching of the hepatic steatosis datasets SHIP-2 and TREND-0 with DIVA preprocessing.

	SHIP-2	TREND-0	p-value
	BEFORE matching		
n	727	1926	-
Age [years]	56.1 ± 12.6	50.1 ± 14.1	< 0.001
Sex [% female]	53.2%	51.7%	0.512
Hep. Stea. [% pos.]	22.4%	24.0%	0.426
	AFTER matching		
n	694	694	-
Age [years]	55.5 ± 12.1	55.5 ± 12.1	1.000
Sex [% female]	53.5%	53.5%	1.000
Hep. Stea. [% pos.]	21.5%	21.5%	1.000

We concentrate on the disorders hepatic steatosis and goiter that serve as case studies to explore the feasibility of our approach. Hepatic steatosis, also known as fatty liver, is a liver disorder which is present in approximately 30% of all adults (Völzke et al., 2011). Although not harmful per se, possible followup diseases like steatohepatitis and liver cirrhosis can cause severe harm (Völzke, 2012). Presence of hepatic steatosis for study participants in both cohorts is indicated through a discretized variable based on the proportion of fat within the liver as measured through Magnetic Resonance Tomography (MRT). Liver-fat MRT results were available for 727 SHIP-2 and 1926 TREND-0 individuals. Catering to the workflows presented in (Niemann et al., 2015) and (Hielscher, Spiliopoulou, Völzke, & Kühn, 2014), a binary variable "H" marks participants which exhibit more than 10% of fat accumulation within their liver as "positive" and the remaining participants as "negative". Table 2 shows the resulting distribution within SHIP-2 and TREND-0.

As second disorder we study goiter, which refers to an enlargement of the thyroid gland that may be defective. The prevalence of goiter is especially high in iodine-deficient regions of the world, reaching up to 80 % (Vanderpump, 2011). In TREND-0, 1390 out of 4400 participants exhibit goiter which is defined by a thyroid gland volume of more than 25 ml for men and 18 ml for women measured by ultrasound, according to (Gutekunst, Becker, Hehrmann, Olbricht, & Pfannenstiel, 1988). Table 1 depicts general characteristics of the cohort datasets SHIP-2 and TREND-0.

5.2. Experimental setup

In the first part of the evaluation, we investigate whether subpopulations that are identified by our framework also exist in independent cohort data. To do so, DIVA matches the datasets according to their distribution in age, sex and medical outcome. Regarding hepatic steatosis, the matched SHIP-2 dataset is used for the identification of subpopulations and the matched TREND-0 dataset for their validation. For goiter, we only utilize TREND-0 which we split into one matched dataset for the identification of subpopula-

tions and one for their validation. The dataset where the subpopulations are identified is the discovery dataset *DD* and the dataset where the subpopulations are validated is the validation dataset *VD*. Then, the discovery module uses DRESS + to identify subpopulations using five ML and five NL constraints that are chosen at random from two constraint pools. For each medical condition, these pools are made up by the set of all participant pairs with the same medical outcome (ML constraints pool) and the set of all participant pairs with different medical outcomes (NL constraints pool) in the respective *DD*. We conduct three measurements for each subpopulation C_{DD} discovered in *DD* and its reconstructed counterpart C_{VD} in *VD*:

- p-value: p-value according to a χ^2-test on the medical outcome distribution within the subpopulations compared to the remaining participants in the respective dataset. We call a subpopulation "significant" if its p-value is below a predefined level $\alpha = 0.05$. Ideally, significant subpopulations found in *DD* should also be significant in *VD*.
- Relative Size Difference (RSD): measurement of the relative difference in size between a subpopulation in *DD* and its reconstructed counterpart in *VD*. The RSD is calculated as

$$RSD(C_{DD}, C_{VD}) = \frac{abs(|C_{DD}| - |C_{VD}|)}{\frac{1}{2} \cdot (|C_{DD}| + |C_{VD}|)}. \quad (5)$$

  Ideally, subpopulations should have similar sizes on both datasets so that the RSD is low.
- Relative Class Distribution Difference (RCDD): measurement of the difference in the relative prevalence of the positive medical outcome of a subpopulation in *DD* that is reconstructed in *VD*. The RCDD is calculated as

$$RCDD(C_{DD}, C_{VD}) = \frac{P(C_{DD})/P(DD)}{P(C_{VD})/P(VD)}, \quad (6)$$

  with $P(C_{DD})$ as the fraction of participants with a positive class label in C_{DD} and $P(DD)$ as the fraction of participants with a positive class label in *DD*. Ideally, subpopulations should have a similar distribution in the medical outcome on both datasets so that the RCDD is low.

Since we do not have real instance-level constraints available, we assess the estimated impact of random choices of constraints on the results. To obtain a sufficiently accurate quality estimate of our subpopulations, we repeat the experiments with 20 different randomized constraint settings. We decided for no less than 20 trials so that we can rely on mean and standard deviation.

In the second part of the evaluation, we vary the number of constraints to analyze how the proportion of significant clusters changes. For this, we calculate the number of subpopulations that are significant ($\alpha = 0.05$) in relation to all found subpopulations, given a fixed number of random constraints. Ideally, more constraints lead to the discovery of more subpopulations that are significant compared to all discovered subpopulations.

5.3. Transferability of identified subpopulations

Table 2 shows the distribution of the datasets before and after matching. Fig. 4 and 6 show the median χ^2-test p-value of the 25 best subpopulations discovered in the *DD* (left part of the figure), and the associated subpopulations in *VD* (right part of the figure), for hepatic steatosis and goiter over 20 runs. Considering hepatic steatosis, up to rank 16 the p-value between a subpopulation found in *DD* and its *VD* counterpart deviates little. Subpopulations up to rank 16 are significant and subpopulations from rank 23 to 25 are not significant in both datasets, which indicates that the results of our framework are transferable. Although, the subpopulations

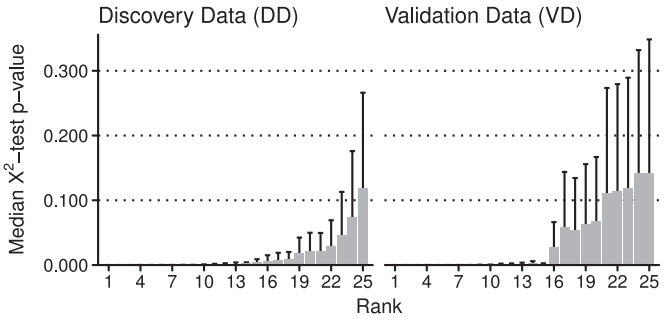

Fig. 4. Ranking of the χ^2-test p-value median over 20 runs for subpopulations found in the hepatic steatosis *DD* compared to their *VD* counterpart. Ordering of the subpopulations in *DD* is mostly preserved in *VD*. The p-values are mostly comparable.

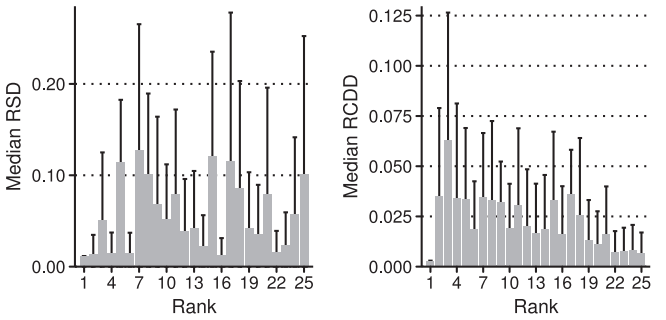

Fig. 5. Median RSD and RCDD for the 25 best ranked subpopulations found in the hepatic steatosis dataset.

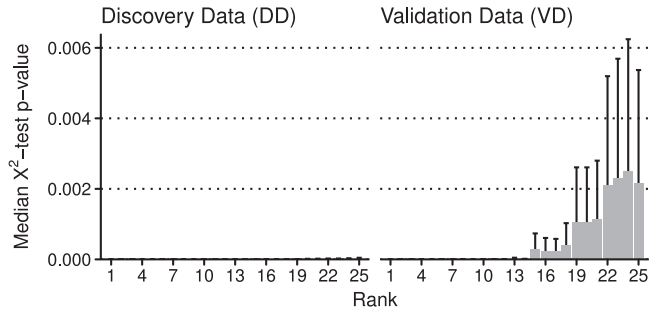

Fig. 6. Ranking of the χ^2-test p-value median over 20 runs for subpopulations found in the goiter *DD* compared to their *VD* counterpart. Ordering of the subpopulations in *DD* is not fully preserved in *VD*. However, p-values are very low and similar for all subpopulations.

from rank 17 to 22 were found significant in *DD* but not significant within *VD*, in many instances these ranks were close to the decision border (i.e., 17, 18, 19, 20). Another favorable result comes from the order of the subpopulations. The ordering is mainly preserved in *VD* with the exception of rank 14 and 18, which indicate stability of our framework. Fig. 5 presents RSD and RCDD on the hepatic steatosis data. RSD shows some variability but is generally low with a median of $\approx$12% in the worst and $\approx$1% in the best cases. For large subpopulations a RSD of 12% corresponds to an approximate size difference of 15 participants between a subpopulation found in *DD* and that is reconstructed in *VD*. RCDD is low across all subpopulations, exhibiting median values between 1.5% and 3.5% most of the time.

For goiter the picture changes slightly. All of the 25 subpopulations are highly significant on both *DD* and *VD*. Although, beginning with rank 15, the p-values increase slightly on *VD*, they remain considerably below 0.01. It is notable that here the ordering is not fully preserved in *VD*, though the absolute deviation in

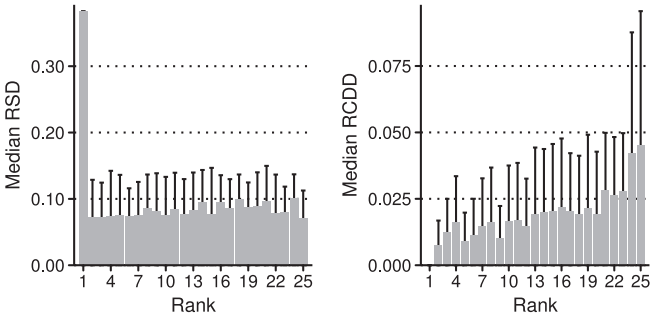

Fig. 7. Median RSD and RCDD for the 25 best ranked subpopulations found in the goiter dataset.

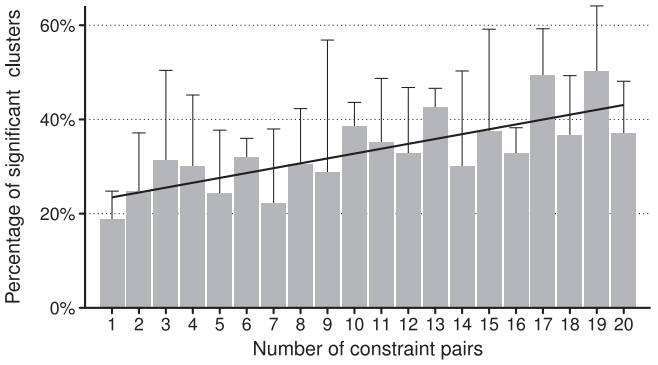

Fig. 8. Percentage of significant subpopulations for different numbers of constraint pairs.

p-value is very low, especially in comparison to the hepatic steatosis data. Median and absolute median deviation of the RSD and RCDD are shown in Fig. 7. The median RSD is low, never exceeding $\approx$10%, except for the best ranked subpopulation. However, there is no medical outcome distribution difference for the rank one subpopulation. In general the RCDD is very low with the median never exceeding $\approx$5% which indicates good transferability of the discovered subpopulations.

5.4. Varying the number of constraints

Fig. 8 shows the relative proportion of significant subpopulations ($\alpha = 0.05$) when varying the number of constraint pairs given to DRESS + on the matched SHIP-2 *DD* (c.f. Tab. 1) . Constraint pairs were chosen at random, with the number of ML and NL constraints being equal, i.e. providing one to 20 ML and NL constraints. For each constraint pair, we performed ten runs and show median and median absolute deviation. Fig. 8 indicates an increase in the relative proportion of significant subpopulations when increasing the number of constraints. While the median for one ML and NL constraint is $\approx$ 20%, it goes as high as $\approx$ 47% (19 pairs) and $\approx$ 38% (20 pairs). The data also shows some variability. Median absolute deviation is moderately high but constant over the constraint pairs. This is due to the random nature of the chosen pairs. The constraint selection is solely based on the medical outcome of participants. It does not take the participants' similarity w.r.t. specific (relevant) characteristics into account. This can lead to the selection of bad constraints, for example must-link constraints between highly dissimilar outlier participants that have the same medical outcome but are not representative for their specific condition. In a real setting, an expert with sound domain knowledge who defines such small number of constraints by hand is unlikely to base constraint decisions exclusively on the outcome.

Table 3
Description and statistics of selected subpopulations.

Hepatic Steatosis				
ID	**Subspace features**	**Size [%]**	**p-value**	**OR**
H#1	age,diabetes	10.0	$1.7\,e^{-09}$	3.01
H#2	female,smoking	11.0	$5.4\,e^{-09}$	0.31
H#3	abstain,physact,smoking	9.9	$2.3\,e^{-09}$	2.06
H#4	ATC_C07A	21.9	$2.7\,e^{-09}$	2.70
Goiter				
ID	**Subspace features**	**Size [%]**	**p-value**	**OR**
G#1	ges_sf12_02,waiidf	33.6	$1.8\,e^{-09}$	0.47
G#2	edyrs,metsyn	26.9	$7.\,e^{-09}$	1.55
G#3	ges_sf12_03,plaque	16.0	$3.4\,e^{-09}$	1.60
G#4	ffs,marit	4.4	$8.2\,e^{-09}$	2.07

5.5. Discussion on the results

In Table 3, we depict four of the subpopulations found to have significantly different prevalence than the overall population for the outcome hepatic steatosis, respectively for the outcome goiter. All eight subpopulations are highly significant with $\alpha = 0.01$ and high odds ratios. For each subpopulation, the distributions of the included features compared to the remaining participants are depicted in Fig. 9 (hepatic steatosis) and Fig. 10 (goiter). A brief description of each feature is provided in Table 4.

Algorithm 1: DRESS + candidate subspace initialization.

Data: Dataset D, original feature set F
Result: Set of subspace clusters $\mathcal{C}$, set of candidate subspaces $\mathcal{S}$, set of subspace quality values Q
Initialize empty set $\mathcal{C}$;
for *each* $f \in F$ **do**
  $\mathcal{C}_f \leftarrow \text{DBSCAN}(D_f)$; `// cluster`
  $\mathcal{C} \leftarrow (\mathcal{C} \cup \mathcal{C}_f)$; `// store initial clusters`
  $\mathcal{S} \leftarrow (\mathcal{S} \cup \{f\})$;
  `// store subspace candidate`
  $Q \leftarrow (Q \cup calcQuality(D_f))$;
  `// store subspace quality`
end

Subpopulations with significant prevalence differences on hepatic steatosis. As can be seen in the upper part of Table 3, H#1 includes participants with diabetes and a rather high relative age that have a much more skewed distribution w.r.t. hepatic steatosis than the complete cohort. These associations are supported in numerous publications such as (Roden, 2006) and (Völzke, Schwarz, Baumeister, & et, 2007). Participants of H#2 exhibit a very low odds ratio, they are all female and current smokers, whereas H#3 has high odds ratio and is made of participants which stated that they were not abstinent to alcohol in the last 12 months, did less than one hour of sports per week (or did not state anything) and are ex-smokers. Correlated features like body mass index or waist circumference were found important in (Hielscher et al., 2014) and (Niemann et al., 2015) regarding fatty liver. The subpopulation H#4 is comprised of all participants under beta-blocker medication. The diagnostic score on hepatic steatosis by Meffert et al. (2014) used this feature on rare occasions.

Subpopulations with significant prevalence differences on goiter. As can be seen in the lower part of Table 3, G#1 consists of mainly non-goiter study participants that have no impairments when performing moderate tasks (`ges_sf12_02` = 3) and without cen-

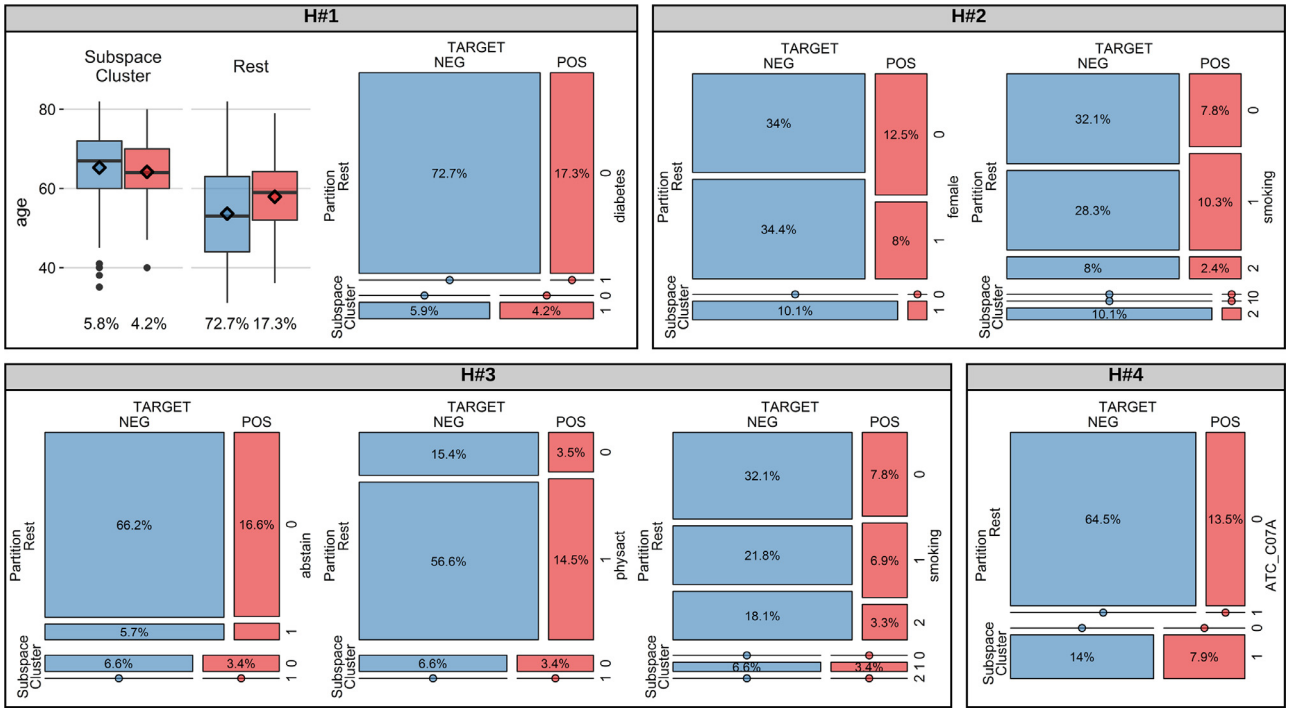

Fig. 9. Distribution of subspace cluster features w.r.t. hepatic steatosis in selected subpopulations that were found by DRESS + . H#2 represents approx. 10% of the population containing considerably older and diabetes-afflicted participants with a higher prevalence of a positive hepatic steatosis outcome. Participants in H#3 are ex-smokers that are less alcohol-abstinent and less physically active than the rest of the population, and hepatic steatosis is also more prevalent for this subpopulation.

tral obesity according to the categorization by the International Diabetes Federation (IDF). The subpopulation G#2 encompasses a higher proportion of participants with (relative) few educational years who suffer from the metabolic syndrome. In literature, studies on independent cohorts found a relationship between somatographic variables and goiter, associations with the presence of metabolic syndrome and borderline significance on associations with the educational level (Zheng, Yan, Kong, Liang, & Mu, 2015), (Rendina, De Filippo, Mossetti, & et, 2012). In G#3 participants with plaque are present. However, associations between goiter and plaque are not yet confirmed. The included `ges_sf12` features in G#1 and G#3 define a specific subgroup of participants, rather

Table 4
Description of selected features.

name	description	values
`abstain`	Abstinence from alcohol (12 months)	0: no 1: yes
`age`	Age of participant	*numeric*
`ATC_C07A`	Intake of beta-blocker	0: no 1: yes
`diabetes`	Suffers from diabetes	0: no 1: yes
`edyrs`	Number of educational years	*numeric*
`ffs`	Food Frequency Score	*numeric*
`ges_sf12_02`	Impaired performing moderate tasks	1: severe limitation 2: slight limitation 3: no limitation
`ges_sf12_03`	Impaired walking multiple stairs	1: severe limitation 2: slight limitation 3: no limitation
`marit`	Marital status	1: single 2: married or relationship 3: separated or divorced 4: widowed
`metsyn`	Suffers from metabolic syndrome	0: no 1: yes
`physact`	Leisure time physical activity	0: $<$ 1h phys. act./week 1: $\geq$ 1h phys. act./week
`plaque`	Plaque	0: no | 1: yes
`smoking`	Smoking status	0: never smoked 1: ex-smoker 2: current smoker
`waiidf`	Waist circumference	0: $<$ 80 cm (men: 94 cm) 1: $\geq$ 80 cm (men: 94 cm)

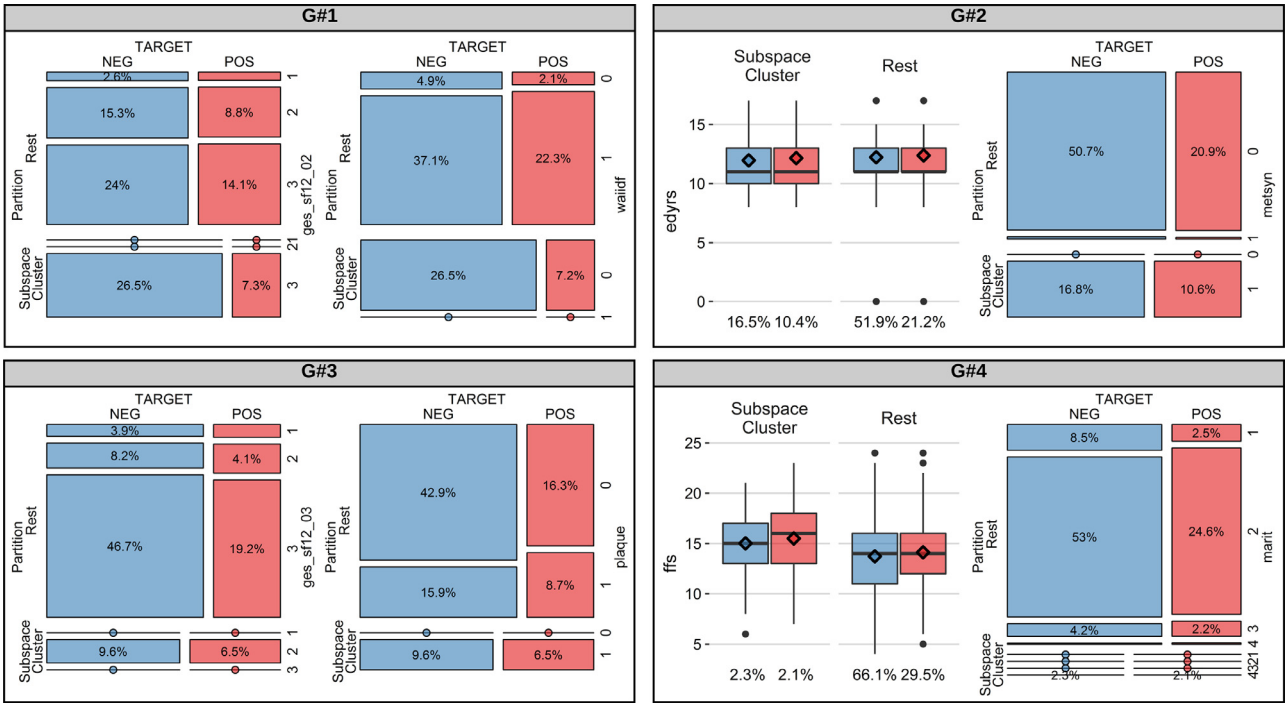

Fig. 10. Distribution of subspace cluster features w.r.t. goiter in selected subpopulations that were found by DRESS + . Participants in G#4, accounting for 4.4% of the study population, are widowed and adhere to a more favorable diet than the rest but are more likely to exhibit goiter. G#1 contains 33.6% of the study population and shows that a preferable waist circumference reduces the probability of having goiter despite the participants' severe limitations while performing moderate activities such as moving a table, vacuuming, bowling.

Algorithm 2: DRESS + subspace processing and cluster generation.

Data: Dataset D, set of subspace clusters $\mathcal{C}$, set of candidate subspaces $\mathcal{S}$, set of subspace quality values Q
Result: Set of subspace clusters $\mathcal{C}$
$S_{candidate} \leftarrow pickBest(\mathcal{S}, Q)$; `// pick best`
$q_{best} \leftarrow q(S_{candidate})$;
`// init. best quality value`
$\mathcal{S} \leftarrow (\mathcal{S} \setminus \{S_{candidate}\})$;
`// clean candidate subspaces`
while $|\mathcal{S}| > 0$ **do**
  for *each* $S^* \in \mathcal{S}$ **do**
    $S_{new} \leftarrow (S^* \cup S_{candidate})$; `// merge`
    if $q_{dist}(S_{candidate}) > q_{dist}(S^*)$ **then** $S_{dist} \leftarrow S_{candidate}$;
    else $S_{dist} \leftarrow S^*$;
    $q_{dist}(S_{new}) \leftarrow calcDistQual(D_{S_{new}})$;
    if $q_{dist}(S_{new}) > q_{dist}(S_{dist})$ **then**
      `// filter criterion`
      $\mathcal{C}_{D_{S_{new}}} \leftarrow$ DBSCAN$(D_{S_{new}})$;
      $\mathcal{C} \leftarrow (\mathcal{C} \cup \mathcal{C}_{D_{S_{new}}})$;
      $Q \leftarrow (Q \cup calcQuality(D_{S_{new}}))$;
      if $q(S_{new}) > q_{best}$ **then**
        $q_{best} \leftarrow q(S_{new})$;
        $\mathcal{S} \leftarrow (\mathcal{S} \cup S_{new})$;
        $\mathcal{S} \leftarrow (\mathcal{S} \setminus \{S^*\})$; `// clean`
      end
    end
  end
  $S_{candidate} \leftarrow pickBest(\mathcal{S}, Q)$;
  $\mathcal{S} \leftarrow (\mathcal{S} \setminus \{S_{candidate}\})$; `// clean`
end

healthy and "fit" (G#1) compared to having mild difficulties ascending multiple stairs (G#3). In the data, these variables are correlated to the somatographic ones, and therefore contain some redundant information.

Limitations on the identified subpopulations. As pointed out by Cummins (2012), calibration of the model for different target populations and validation on unseen data are essential steps towards overcoming the scepticisms of medical experts towards machine learning methods. The Validation module of DIVA has been designed to address this scepticism, but the validation is limited by the availability of comparable populations. In particular, a second independent cohort was only available for hepatic steatosis. For goiter, we rather had to reserve part of TREND-0 for validation, rather than use the SHIP-CORE, because the differences in age and iodine exposition between the two cohorts did not permit matching. This affects the transferability of the identified subpopulations to other populations.

Moreover, the constraints for subspace construction were not delivered by the medical expert. Rather, we generated them, using the class labels as sole criterion. It is likely that a medical expert would consider additional information when defining constraints, and thus may have lead DIVA to the discovery of different subpopulations.

6. Concluding discussion on DIVA

Our system DIVA encompasses a Discovery module that contains semi-supervised and fully automated utilities, an Inspection module for model visualization and a Validation module for cohort matching and model validation. In comparison to existing systems, DIVA is unique in the exploitation of similarity constraints for semi-automated subspace construction, automated learning and validation of the learned models in a second, independent, automatically matched population. The constraint-based subspace dis-

covery method DRESS (Hielscher et al., 2016), on which DIVA builds, has the same subspace construction mechanism but is not a complete system. Our results have shown that DIVA discovers subpopulations that are significantly different than the population with respect to the outcome, although it uses a small amount of expert information.

Subspace construction. The subspaces derived by DIVA cannot be compared to the sets of variables chosen interactively with help of systems like CAVA (Zhang et al., 2014). It must be expected that the human expert will choose variables wiser than a system that exploits only a small fragment of the expert's medical knowledge, especially when the expert is supported by elaborate visualizations as offered in Zhang et al. (2014). We see DIVA as complementary to such interactive systems, since DIVA minimizes the demand on the expert's availability and can select variables that the expert did not think of.

A limitation of this functionality in DIVA is due to the absence of a mechanism that reads and incorporates expert-defined constraints. While designing a rudimentary constraint reader is trivial, the design of an interactive environment that is appropriate for the medical expert is a major challenge. Studies on medical expert - intelligent system interaction, as by Thew et al. (2009), stress different forms of interaction but the specification of similarity constraints between study participants is less of a focus. However, representations of a patient's EHR, as in (Gotz et al., 2011; Zhang et al., 2014), are a good starting point, on which we intend to capitalize in cooperation with the epidemiology expert.

Learning. The machine learning part is fully automated in the Discovery module of DIVA. This is in contrast with the many tuning options offered in a typical machine learning environment and especially in an interactive one. Since DIVA is a self-tuned system, it exhibits a trade-off between execution speed and model robustness, since robustness demands extensive hyperparameter optimization. In the current version of DIVA, self-tuning was kept to a minimum, so efficient workflows for hyperparameter optimization are an urgent future task to ensure robustness.

With respect to diversity of algorithms, the Discovery module of DIVA is richer than our earlier system (Niemann et al., 2017; Niemann et al., 2014; Schleicher et al., 2017). Moreover, the workflow of DIVA allows for the inclusion of further, more elaborate machine learning algorithms, albeit the self-tuning will require substantial extensions.

A current limitation of the Discovery module of DIVA concerns the exploitation of temporal information. Advances on time-series analysis are not applicable in our context of population-based studies, since the cohorts are observed for a few timepoints that constitute tiny sequences rather than multivariate timeseries. Still, methods that incorporate temporal information into the high-dimensional feature space should be exploited. We are currently considering ways of expanding DIVA with modules that exploit temporal data during subspace construction (Hielscher et al., 2014; Niemann et al., 2015) and model learning.

Visualizations. The Inspection module of DIVA is more limited in its functionalities than typical interactive systems in support of the medical expert, including the systems of (Krause et al., 2016a; Krause et al., 2014; Krause et al., 2016b; Zhang et al., 2014). However, this is due to the focus of DIVA on model inspection and comparison rather than cohort construction. With respect to model drill-down, the D-Inspector is less sophisticated than the method of Niemann et al. (2017), but has the advantage of comparing the models delivered by the Validation module in a seamless way.

Automated model validation. The Validation module of DIVA is unique. Obviously, any interactive system for cohort construction, including those cited earlier, can be used to build a second cohort with similar properties than the studied one, and use this second cohort to validate the learned model. However, DIVA takes a cohort and builds a matched subcohort automatically and then automatically validates the learned model learned using this matched subcohort. We expect that this functionality will speed up the strenuous task of model validation, and we plan to investigate this further by comparing the quality of models validated on matched subcohorts to the quality of models validated on manually chosen (sub)cohorts.

References

Atzmüller, M. (2015). Subgroup discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5*, 35–49.

Atzmüller, M., & Puppe, F. (2006). SD-map–a fast algorithm for exhaustive subgroup discovery. Springer. Proc. of European Conference on Principles and Practice of Knowledge Discovery in Databases (KDD), pp. 6–17.

Cummins, M. R. (2012). Nonhypothesis-driven research: Data mining and knowledge discovery. In *Clinical research informatics* (pp. 277–291). Springer London.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of international conference on knowledge discovery and data mining (KDD)* (pp. 226–231).

Gotz, D., Sun, J., Cao, N., & Ebadollahi, S. (2011). Visual cluster analysis in support of clinical decision intelligence. In *Proc. of AMIA* (p. 481).

Grosskreutz, H., Rüping, S., & Wrobel, S. (2008). Tight optimistic estimates for fast subgroup discovery. In *Proc. of european conference on machine learning and principles of knowledge discovery in databases (ECML PKDD)* (pp. 440–456).

Günnemann, S., Färber, I., Rüdiger, M., & Seidl, T. (2014). SMVC: semi-supervised multi-view clustering in subspace projections. In *Proc. of ACM SIGKDD conference on knowledge discovery and data mining (KDD)* (pp. 253–262).

Gutekunst, R., Becker, W., Hehrmann, R., Olbricht, T., & Pfannenstiel, P. (1988). Ultraschalldiagnostik der schilddrüse. *Deutsche Medizinische Wochenschrift (DMW), 113*, 1109–1112.

Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, 46*, 389–422.

Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *Proc. of international conference on machine learning (ICML)* (pp. 359–366).

Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: Foundations and applications. *Knowledge and Information Systems, 29*, 495–525.

Hielscher, T., Spiliopoulou, M., Völzke, H., & Kühn, J. P. (2014). Mining longitudinal epidemiological data to understand a reversible disorder. In *Proc. of international symposium on intelligent data analysis (IDA)* (pp. 120–130).

Hielscher, T., Spiliopoulou, M., Völzke, H., & Kühn, J. P. (2016). Identifying relevant features for a multi-factorial disorder with constraint-based subspace clustering. In *Proc. of IEEE symposium on computer-based medical systems (CBMS)* (pp. 207–212).

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). Matchit: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software, 42*, 1–28.

Kailing, K., Kriegel, H. P., Kroeger, P., & Wanka, S. (2003). Ranking interesting subspaces for clustering high dimensional data. In *Proc. of european conference on principles of data mining and knowledge discovery (ECML PKDD)* (pp. 241–252).

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial intelligence, 97*, 273–324.

Krause, J., Dasgupta, A., Fekete, J. D., & Bertini, E. (2016a). SeekAView: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces. In *Proc. of IEEE symposium on large data analysis and visualization (LDAV)* (pp. 11–19).

Krause, J., Perer, A., & Bertini, E. (2014). Infuse: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics, 20*, 1614–1623.

Krause, J., Perer, A., & Ng, K. (2016b). Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. of conference on human factors in computing systems (CHI)* (pp. 5686–5697).

van Leeuwen, M., & Knobbe, A. (2012). Diverse subgroup set discovery. *Data Mining and Knowledge Discovery (DAMI), 25*, 208–242.

Liu, J., Brodley, C. E., Healy, B. C., & Chitnis, T. (2015). Removing confounding factors via constraint-based clustering: An application to finding homogeneous groups of multiple sclerosis patients. *Artificial Intelligence in Medicine (AIM), 65*, 79–88.

Meffert, P. J., Baumeister, S. E., Lerch, M. M., Mayerle, J., Kratzer, W., & Völzke, H. (2014). Development, external validation, and comparative assessment of a new diagnostic score for hepatic steatosis. *The American Journal of Gastroenterology, 109*, 1404–1414.

Niemann, U., Hielscher, T., Spiliopoulou, M., Völzke, H., & Kühn, J. P. (2015). Can we classify the participants of a longitudinal epidemiological study from their previous evolution? In *Proc. of IEEE symposium on computer-based medical systems (CBMS)* (pp. 121–126).

Niemann, U., Spiliopoulou, M., Preim, B., Ittermann, T., & Völzke, H. (2017). Combining subgroup discovery and clustering to identify diverse subpopulations in cohort study data. In *Proc. of IEEE symposium on computer-based medical systems (CBMS)* (pp. 582–587).

Niemann, U., Völzke, H., Kühn, J. P., & Spiliopoulou, M. (2014). Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis. *Expert Systems with Applications (ESWA), 41*, 5405–5415.

Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: A review. *SIGKDD Exploration Newsletter, 6*, 90–105.

Preim, B., Klemm, P., Hauser, H., Hegenscheid, K., Oeltze, S., Toennies, K., & Völzke, H. (2016). Visual analytics of image-centric cohort studies in epidemiology. In *Visualization in medicine and life sciences III* (pp. 221–248).

Rendina, D., De Filippo, G., Mossetti, G., & et, a. (2012). Relationship between metabolic syndrome and multinodular non-toxic goiter in an inpatient population from a geographic area with moderate iodine deficiency. *Journal of Endocrinological Investigation, 35*, 407–412.

Roden, M. (2006). Mechanisms of disease: Hepatic steatosis in type 2 diabetes–pathogenesis and clinical relevance. *Nature Reviews Endocrinology, 2*, 335–348.

Ruiz, C., Spiliopoulou, M., & Menasalvas, E. (2007). C-DBSCAN: Density-based clustering with constraints. In *Proc. of international workshop on rough sets, fuzzy sets, data mining, and granular-soft computing (RSFDGrc)* (pp. 22–216).

Schleicher, M., Ittermann, T., Niemann, U., Völzke, H., & Spiliopoulou, M. (2017). ICE: Interactive classification rule exploration on epidemiological data. In *Proc. of IEEE symposium on computer-based medical systems (CBMS)* (pp. 606–611).

Sim, K., Gopalkrishnan, V., Zimek, A., & Cong, G. (2013). A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery (DAMI), 26*, 332–397.

Thew, S., Sutcliffe, A., Procter, R., De Bruijn, O., McNaught, J., Venters, C. C., & Buchan, I. (2009). Requirements engineering for e-science: Experiences in epidemiology. *IEEE Software, 26*.

Völzke, H., Schwarz, S., Baumeister, S. E., & et, a. (2007). Menopausal status and hepatic steatosis in a general female population. *Gut, 56*, 594–595.

Vanderpump, M. (2011). The epidemiology of thyroid disease. *British Medical Bulletin, 99*, 39–51.

Völzke, H. (2012). Multicausality in fatty liver disease: Is there a rationale to distinguish between alcoholic and non-alcoholic origin. *World Journal of Gastroenterology, 18*, 501–3492.

Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., et al. (2011). Cohort profile: The study of health in pomerania. *International Journal of Epidemiology, 40*, 294–307.

Wagstaff, K., Cardie, C., Rogers, S., Schrödl, S., et al. (2001). Constrained k-means clustering with background knowledge. In *Proc. of international conference on machine learning (ICML): vol. 1* (pp. 577–584).

Wilson, D. R., & Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligenece Research, 6*, 1–34.

Zhang, Z., Gotz, D., & Perer, A. (2014). Iterative cohort analysis and exploration. *Information Visualization, 14*, 289–307.

Zheng, L., Yan, W., Kong, Y., Liang, P., & Mu, Y. (2015). An epidemiological study of risk factors of thyroid nodule and goiter in chinese women. *International Journal of Environmental Research and Public Health, 12*, 11608–11620.