Learning from Cohort Data

Myra Spiliopoulou

September 19 – ECML PKDD 2016

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INF

FACULTY OF
COMPUTER SCIENCE

KMD

The Knowledge Management & Discovery Lab Magdeburg

Research Buzzwords in the KMD Lab

- Mining ratings, opinions, texts, cohorts
- Streams of high-dimensional data to model / to predict: evolution of preferences, evolution of individuals
- Systematically incomplete data
- Incorporating expert knowledge into the learning process with constraint-based learning, semi-supervised learning, active learning

1. Cohorts in Population-based Studies
    - Building Cohorts in a Population-based Study
    - Using the cohort data for learning, the traditional way
    - Exploiting a cohort's feature space in a supervised way
    - Exploring a cohort's feature space in a semi-supervised way
    - Learning from timestamped, systematically incomplete cohort data
2. Building Clinical Cohorts from Electronic Health Records
    - Building and exploring a cohort with association rules
    - Building and exploring a cohort with Visual Analytics
3. Experiments with Clinical Cohorts
    - DFS Example I: Learning Profiles of Pressure Load during Walking
    - DFS Example II: Learning Profiles of Pressure Load during Standing
4. Closing Remarks

Example: The population-based longitudinal STUDY OF HEALTH IN POMERANIA – SHIP [Völzke et al., 2011]

PICTURE REMOVED: the region of the study

Example: The population-based longitudinal STUDY OF HEALTH IN POMERANIA – SHIP

SHIP cohort profile [Völzke et al., 2011]

- Selection criteria: main residence in Pomerania (Germany), age 20-79
- Cohorts and numbers
  - SHIP (SHIP-Core)
    - SHIP-0: n=4338, 1997-2001
    - SHIP-1: n=3300, 2002-2006
    - SHIP-2: n=2333, 2008-2012
    - SHIP-3: . . .
  - SHIP-TREND
    - SHIP-TREND-0: n=4420, 2008-2012
    - SHIP-TREND-1: . . .
- Recordings
  - sociodemographics
  - somatographic tests, medical/lab tests
  - ultrasound & MRT

Learning from the population-based study data, the traditional way

- Formulate a hypothesis,
  e.g. on how smoking affects thyroid enlargement
- Select the data appropriate for this hypothesis
  - Which cohort waves?
  - Which population strata?
  - Which variables?
- Perform a retrospective study on those data
- Perform also a prospective study for validation

Effect of smoking on thyroid enlargement [Ittermann et al., 2008]

Goal: Study the effect of smoking on thyroid volume progression [1] and on goiter in a region with improved iodine supply for different age strata.

Motivation of the study

- There are regions with iodine deficiency and regions with iodine sufficiency.
- Smoking has been associated with thyroid volume enlargement in some studies (in regions with/without iodine deficiency), but other studies found no association.
- In some regions, measures have been taken to improve iodine supply.

Pomerania was a region of iodine deficiency and high goiter prevalence before the 90s.
”The improved supply of iodine salt into food productions and individual salt consumption during the 1990s in the study region of Northeast Germany led us to the paradoxical situation of high goiter prevalence in a region of improved iodine supply (14)” quoting [Ittermann et al., 2008].

[1] If thyroid volume increases, it usually does not decrease again.

Effect of smoking on thyroid enlargement [Ittermann et al., 2008]

Study subjects and assessments in SHIP-0 and SHIP-1

- 1647 participants (856 m/791 w)
- Iodine concentration: measured by urine iodine excretion
- Thyroid volume: assessed with thyroid ultrasonography
- Thyroid volume progression:= difference between thyroid volume in SHIP-1 vs SHIP-0
- Goiter:= thyroid volume $>$ 18 ml (w), resp. 25 ml (m)

Smoking status and dependent variable

Smoking status

[1] never smoker
[2] smoker at SHIP-0 and at SHIP-1
[3] smoker at SHIP-1 but not at SHIP-0
[4] smoker at SHIP-0 but not at SHIP-1
[5] non-smoker at SHIP-$\{$0,1$\}$ but former smoker (no smoking for $\geq$ 12 months before SHIP-0/SHIP-1 examination)

Goiter status

[1] presence of goiter at SHIP-0
[2] absence of goiter at both SHIP-0 and SHIP-1
[3] presence of goiter at SHIP-1 but not at SHIP-0

Effect of smoking on thyroid enlargement [Ittermann et al., 2008]

The procedure for statistical analysis (page 763)

"Data on quantitative characteristics are expressed as median and inter-quartile range. Data on qualitative characteristics are expressed as percent values or absolute numbers, as indicated. The study population was divided into three groups according to the presence or absence of goiter at baseline and follow-up (1, presence of goiter at baseline; 2, absence of goiter at baseline and follow-up; 3, absence of goiter at baseline but presence at follow-up). Comparisons between groups were made usingχ^2 test (qualitative data) or Wilcoxon test (quantitative data). Wilcoxon's signed rank test was used for paired data. Determinants of thyroid volume change and incident goiter were analyzed by linear and logistic regression respectively. All models were adjusted for age, gender, and body mass index. In the first step, both analyses were performed separately for three different age strata (20–39, 40–59, and 60–79 years). In the second step, analyses were performed for the whole population, and interactions between the smoking variables and age were tested. Interactions were kept in the models for P values $<$ 0.1. . . . From linear regression models, the β and its 95% confidence interval (95% CI) and from logistic regression, odds ratio, and its 95% CI are given. A value of P$<$0.05 was considered statistically significant. . . . "

Effect of smoking on thyroid enlargement [Ittermann et al., 2008]

The main findings

- Elder subjects (stratum of 60-79 years) who were smokers at both SHIP-0 and SHIP-1 were at higher risk for thyroid volume progression. Some of them had already diagnosed goiter in SHIP-0. After exclusing them, the association of smoking and thyroid enlargement disappeared.
- Young subjects (stratum of 20-39 years) who were smokers at both SHIP-0 and SHIP-1 were at lower risk of goiter incidence.

An explanation for the findings on young participants (page 765)
”The goitrogenous effect of cigarette smoking can be partly explained by elevated plasma cyanate (CN^-) concentrations in smokers (26). Univalent anions with sizes similar to iodide, such as CN^-, are able to competitively inhibit the transport of iodide into the thyroid gland.”

Messages to take away from [Ittermann et al., 2008]

- The medical question stands at the beginning of the study.
- The (few!) features of possible relevance are selected judiciously.
- The machine learning method (if any) is a small step in the workflow. The correctness of its results must be guaranteed!
- The statistical analysis is a larger step in the workflow. The purpose is to guarantee the correctness of the final results.
- The medical discussion is the most important part.

When we look for subpopulations that have higher prevalence of the disease than the general population:

? how to exploit *all* features, since we do not know their importance in advance?

How to deal with a large feature space?

▶ Reduce the feature space by selecting the most informative variables that are minimally associated to each other [Hielscher et al., 2014b]

× Perform classification rules discovery and deliver statistics and navigation aids for the exploration of the rule space [Niemann et al., 2014a, Niemann et al., 2014c]

× Build subspaces that contain potentially interesting subpopulations, without revealing the target variable [Niemann et al., 2014b]

▶ Discover feature subspaces that contain interesting subpopulations in a semi-supervised way [Hielscher et al., 2016]

Learning to separate the cohort data [Hielscher et al., 2014b]

Setting up the dataset for classification – example: "hepatis steatosis"

- Random sample of 578 SHIP-2 participants (314 F, 264 M)
- Target variable derived from a numerical variable showing the fat accumulation in the liver as percentage (mrt_liverfat_s2)
- A - liver fat $\leq$10%, B - liver fat $\in$ (10%,25%], C - liver fat $>$25%

PICTURE REMOVED: picture showing the class distribution for women and for men

Requirements: good model, understandable, explainable

+ Excellent recordings, perfectly sanitized, no missing information
+ Many informative variables
− Many variables, few entities
− Systematically incomplete data

Learning to separate the cohort data [Hielscher et al., 2014b]

Mining Tasks

- Supervised discretization of all dimensions using information gain
- Construction of a subspace of dimensions that maximizes merit
- Specification of a similarity function for high-dimensional data
- kNN classification
- Mosaic graphs for the presentation of important features

PICTURE REMOVED: picture showing the workflow from [Hielscher et al., 2014b]

Learning to separate the cohort data [Hielscher et al., 2014b]

Dimensionality reduction with merit maximization: Find a maximal subset of features that are informative towards the target and are not redundant towards each other.

Merit M_S of a set of n features S [Hall, 2000]

$$M_S = \frac{n\overline{r_{cf}}}{\sqrt{n + n(n-1)\overline{r_{ff}}}}$$

$\overline{r_{cf}}$: mean feature-class dependency, $\overline{r_{ff}}$: mean feature-feature dependency of S, i.e.

$$\overline{r_{ff}} = \frac{\sum_{i=1}^{n-1}\left(\sum_{j=i+1}^{n} SUC(a_i, a_j; I)\right)}{\frac{n}{2}\cdot(n-1)} \quad , \quad \overline{r_{cf}} = \frac{\sum_{i=1}^{n} SUC(c, a_i; I)}{n}$$

using the Symmetrical Uncertainty Coefficient [Press et al., 1992]:

$$SUC(a_i, a_j; I) = 2\cdot\frac{IG(a_i, a_j; I)}{H(a_i, I) + H(a_j, I)} = 2\cdot\frac{H(a_i, I) - H(a_i, I|a_j, I)}{H(a_i, I) + H(a_j, I)}$$

where $H(a_i, I)$ and $H(a_j, I)$ are the feature entropy values for dataset I and features a_i, a_j. $IG(a_i, a_j; I)$ specifies the entropy reduction on a_i and I given a_j.

Learning to separate the cohort data [Hielscher et al., 2014b]

Mining Tasks

- Supervised discretization of all dimensions using information gain
- Construction of a subspace of dimensions that maximizes merit

Computation of a merit-maximizing subset of features using Correlation-based Feature Selection (CFS) [Hall, 2000]

Starting with an empty set of features S from the feature space F:

- add to S the feature $a \in F$ that leads to the highest new merit-value $M_{S \cup a}$.
- remove a from F

until F is empty, or adding any feature from F to S decreases the merit.

- Specification of a similarity function for high-dimensional data
- kNN classification
- Mosaic graphs for the presentation of important features

Learning to separate the cohort data [Hielscher et al., 2014b]

Mining Tasks

- Supervised discretization of all dimensions using information gain
- Construction of a subspace of dimensions that maximizes merit
- Specification of a similarity function for high-dimensional data
- kNN classification

Accuracy plots for different similarity functions

PICTURE REMOVED: accuracy plots from [Hielscher et al., 2014b]

- Mosaic graphs for the presentation of important features

Learning to separate the cohort data [Hielscher et al., 2014b]

Example of a two-dimensional mosaic graph

PICTURE REMOVED: one mosaic graph example [Hielscher et al., 2014b]

Learning to separate the cohort data [Hielscher et al., 2014b]

PICTURE REMOVED: mosaic graphs of important features for the male subpopulation, from [Hielscher et al., 2014b]

Learning to separate the cohort data [Hielscher et al., 2014b]

PICTURE REMOVED: mosaic graphs of important features for the female subpopulation, from [Hielscher et al., 2014b]

Messages to take away from [Hielscher et al., 2014b]

- √ Many features are correlated: *merit-based feature selection* allows us to focus on those that are informative and not heavily dependent on each other.
- √ Mosaic graphs allow us to show one-dimensional subpopulations.

When we look for subpopulations that have higher prevalence of the disease than the general population:

- ? how to make sure that no feature of *possible* importance is projected away?
- ? how to involve the medical expert, instead of removing features automatically?

Constraint-based clustering meets subspace clustering

Clustering with instance-based constraints

For a set of clusters ζ and two distinct instances x, y:

- A *Must-Link constraint* on x, y is satisfied by ζ if there is a $C \in \zeta$ so that $x, y \in C$.
- A *Cannot-Link constraint* on x, y is satisfied by ζ if there are $C_1, C_2 \in \zeta$ so that $x \in C_1, y \in C_2$ and $C_1 \cap C_2 = \emptyset$.

DRESS – Discovery of Relevant Example-constrained SubspaceS [Hielscher et al., 2016]

Given a dataset D and a set of ML and NL constraints,
find the "best" subspace S of the feature space F:

- The clustering in S is of best quality.
- The clustering in S satisfies the constraints.

DRESS [Hielscher et al., 2016]

Quality of a subspace S

Quality wrt constraint satisfaction

$$q_{constraints}(S) = \frac{|ML(S)| + |NL(S)|}{|ML| + |NL|}$$

Cluster stretching with respect to constraints

$$q_{dist}(S) = \frac{\sum_{(x,y)\in NL} d_S(x,y)}{|NL|} - \frac{\sum_{(x,y)\in ML} d_S(x,y)}{|ML|}$$

Overall subspace quality

$$q(S) = q_{constraints}(S) \cdot q_{dist}(S)$$

DRESS workflow [Hielscher et al., 2016]

PICTURE REMOVED: picture showing the workflow from [Hielscher et al., 2016]

DRESS evaluation [Hielscher et al., 2016]

Alternatives for feature selection

- No feature selection: all features used for learning
- Correlation-based Feature Selection [Hall, 2000], using m% of the labeled instances
- DRESS, using n% of the labeled instances

Impact of the feature selection on the performance of a classifier

TABLE REMOVED: table showing the performance of the classifiers, from[Hielscher et al., 2016]

DRESS [Hielscher et al., 2016]

Subpopulations found with DRESS

PICTURE REMOVED: picture showing mosaic graphs of important features, from [Hielscher et al., 2016]

Exploring Data & Feature Space with Semi-Supervised Subspace Clustering

Messages to take away from [Hielscher et al., 2016]

- Cluster discovery must take place only in conjunction to a medical outcome.
- Knowledge on the medical outcome can be used to inform the clustering process. But how to do so while keeping some data for validation?
- Constraint-based Subspace Clustering contributes to the discovery of interesting feature combinations and subpopulations ← Few constraints suffice.

Open Issues:

- How to get the constraints? ? Visual Analytics
- How to explain the clusters?

See examples in [Deschamps et al., 2013, Niemann et al., 2016b]

Discovering interesting features and subpopulations while exploring time

Modeling change and event occurrence: Approaches

- **Core methodology:**
  J. D. Singer, J. B. Willett (2003) *Applied Longitudinal Data Analysis - Modeling Change and Event Occurrence*. OXFORD UNIVERSITY PRESS
- **One advanced approach:** State Space Models, see e.g.
  - A.C. Smith (2015) ”State space modeling for analysis of behavior in learning experiments”, Chapter 10 in *Advanced State Space Methods for Neural and Clinical Data*, p. 231–253
  - F. Krüger, M. Nyolt, K. Yordanova, A. Hein, T. Kirste (2014) ”Computational State Space Models for Activity and Intention Recognition. A Feasibility Study”, PLOSone 9(11), Nov. 2014

What if there is systematic data loss over time?

Discovering interesting features and subpopulations while exploring time

Dealing with data loss over time

In longitudinal population-based studies, we experience a *systematic loss* of data over time because:

- Some participants exit the study.
- The protocol may change:
  - New medical technologies emerge that allow for better testing of some conditions and even for better diagnostics. The use of such technologies is gradually taken over in study protocols. Example: MRT
  - New scientific questions arise: new diseases, conditions become of interest; new types of medical assessments are performed.
  - The recording of some tests or assessments is discontinued.

Case Study "Fatty liver": Classification with exploitation of systematically incomplete data [Niemann et al., 2015]

Long term impact of some disorders

PICTURE REMOVED: title and text from an article by Söderberg et al. on NAFLD, published in *Hepatology*, 2010

Exploiting systematically incomplete timestamped data

How to incorporate the unlabeled data into the learning process?

- **Key idea 1:** Exploit people similarity during learning
  $\Downarrow$
  Clustering-andThen-classification
- **Key idea 2:** Use similarity as a feature
  $\Downarrow$
  ClusterIDs as features
- **Key idea 3:** Model people similarity across the time axis

$$\Downarrow$$

- cohort member $:=$ vector of value-sequences [Hielscher et al., 2014a]
- cohort member $:=$ member of an evolving cluster [Niemann et al., 2015]

Exploiting similarity of incomplete value-sequences for learning [Hielscher et al., 2014a]

Original workflow on the static data [Hielscher et al., 2014b]

PICTURE REMOVED

Learning from incomplete value-sequences [Hielscher et al., 2014a]

Extended workflow on the historical data

PICTURE REMOVED: picture showing the workflow from [Hielscher et al., 2014a]

Learning from incomplete value-sequences [Hielscher et al., 2014a]

Turning sequences of values into new features

PICTURE REMOVED: example of creating sequence-features from [Hielscher et al., 2014a]

- Discretization: stepwise partitioning of the continuous range of values into segments, so that gain is maximized
- Within-feature density-based clustering of the value-sequences
- Deriving sequence-features to exploit the cross-wave similarity of participants for each feature

Learning from incomplete value-sequences [Hielscher et al., 2014a]

Most important sequence-features

stea_seq: most important sequence-feature for the female subpopulation

stea_seq: important sequence-feature for the male subpopulation

ggt_s_seq: important sequence-feature for the male subpopulation

PICTURE REMOVED: picture showing mosaic graphs, from [Hielscher et al., 2014a]

Exploiting patient evolution for learning [Niemann et al., 2015]

PICTURE REMOVED: outline of the approach in [Niemann et al., 2015]

Learning from evolving clusters [Niemann et al., 2015]

Workflow and variants for learning over the incomplete wave data

PICTURE REMOVED

Learning from evolving clusters [Niemann et al., 2015]

How does the workflow affect classification quality?

PICTURE REMOVED: performance plots from [Niemann et al., 2015]

Sensitivity, Specificity and F-Measure scores of different classifiers (baselines - in grey) and by their workflow-enhanced counterparts (colored lines), when varying the number k of neighbours to a cohort member

Learning from evolving clusters [Niemann et al., 2015]

Most important evolution features

PICTURE REMOVED: picture showing boxplots from [Niemann et al., 2015]

Exploiting systematically incomplete timestamped data

Messages to take away from [Hielscher et al., 2014a, Niemann et al., 2015]:

- The systematically incomplete waves can be reasonably exploited for learning:
  - *Key idea:* model and exploit the similarity of cohort participants over time
  - *Similarity of sequences:* Each feature of the original feature space is translated into a sequence-feature with as many values as are there in the individual waves.
  - *Finding:* The value-sequences for some features contribute to separation and are more informative than the individual values.

$\Downarrow$

- Similarly evolving cohort participants constitute subpopulations with accentuated characteristics.

$\Uparrow$

- Dimensionality reduction is a mission-critical part of the workflow. Without it, similarity cannot be exploited in a reliable way.

Disorders Associated with Charcot Foot [Munson et al., 2014]

Charcot Foot is a rare disease: the bones/joints get brittle and disintegrate.

- Charcot Foot usually follows a bone injury.
- It often appears as followup of diabetes.
- Some risk factors are known, but the pathogenesis is not completely understood.

Goal of the study is to identify novel associations between Charcot Foot and other disorders/diseases,
paying particular emphasis on the temporal relationship in such associations.

The chase for Charcot Foot cases

- *Site of the study:* University of Michigan Health System (UMHS), encompassing three hospitals with six speciality centers (including a diabetes center with a podiatric clinic)
- *Complete dataset:* 1.6 million patients with 41.2 million ICD-9 codes (timestamped)
- *Candidates for Charcot Foot diagnosis:* “arthropathy associated with a neurological disorder” (ICD-9 code 713.5), amounting to 388 patients.

Diagnoses Associated with Charcot Foot [Munson et al., 2014]

Method

- Reviewing by Experts to separate among (1) well-known associations, (2) associations that were less known / had the potential to be novel, (3) uninformative associations – either because the ICD was unspecific [2] or because it was a misdiagnosis [3] that was later followed by the correct one, namely ”Charcot Foot”
- Investigation of the role of diabetes by separating between patients with Charcot Foot and diabetes (n=282), and those with Charcoot Foot but without diabetes (n=106) and investigating the dominant associations
- Ranking of the associations on p-values and odds ratio
- Testing the significance of the temporal relationship, i.e. when another diagnosis precedes the 713.5 diagnosis, using binomial test and $p <$0.001 [4]

[2] unspecific ICD, e.g. ”viral infection, not otherwise specified”

[3] misdiagnosis like ””gout, not otherwise specified”

[4] The test was on whether the one ICD-9 code preceded the other in a non-random way.

Diagnoses Associated with Charcot Foot [Munson et al., 2014]

Main findings 1:

676 (of 710) associations with p-value $<$ 0.001; 603 with odds ratio $>$5.0

- Some were not reportedly linked to Charcot Foot but can be associated to it on the basis of existing etiology models. (e.g. bladder disorder; diseases/disorders associated with neurotrophic influences)
- Some diagnoses could be explained by diabetes, e.g. obesity, peripheral neuropathy.
- Associations that did not fit to etiology models but had very high odds ratio were: alkalosis, pulmonary eosinophilia [5], esophagean reflux [6]

Main findings 2:

111 ICD-9 codes with significant temporal relationship to Charcot Foot

- Four of them followed Charcot Foot (327.23 ”obstructive sleep apnea”, 786.7 ”abnormal chest sounds”, 353.6 ”phantom limb syndrome”, 786.9 ”nonspecific symptoms involving the chest and respiratory system”)
- Alkalosis preceded Charcot Foot 100% of the times; pulmonary eosinophilia also preceded it (significantly).

[5] Pulmonary eosinophilia may be treated with steroids; these may affect bone mineral density.

[6] Esophagean reflux might be associated to proton pump inhibitors; -//- -//- -//-

Diagnoses Associated with Charcot Foot [Munson et al., 2014]

Restrictions of the study – reported

- No correction for multiple testing
- Associations refer to single diagnostic codes, but many codes are correlated.
- All associations are made available (as supplementary material) but only some are presented in the main paper.
- Temporal relationship refers to one diagnosis preceding the other, but this does not imply causality.

Restrictions of the study – indicated

- Long elapsed time between entry to UMHS and Charcot Foot diagnosis: mean: 6.6 years; median: 6.2 years Diabetes since when?
- Many distinct ICD-9 codes per patient: mean: 84.8; median: 68.5 Is this typical for Charcot Foot?
- Some of the reported diagnoses are associated with age

? Support of the identified associations ?

Diagnoses Associated with Charcot Foot [Munson et al., 2014]

Messages to take away from [Munson et al., 2014]

Medical researchers use EHRs to build clinical cohorts for retrospective studies.

- The cohorts are usually small, even for non-rare diseases.
- Stratification, especially w.r.t. variables of known influence, might further decrease the size of a cohort.
- The data are noisy.
- The records are systematically incomplete.

- All results must be tested on significance.
- For frequent itemset discovery, correction for multiple testing is necessary.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Cohort Analysis on Electronic Health Records:

Parties involved: team of physicians + team of technologists

Goal: get new insights about a population of patients (e.g. all patients of the cardiology unit who have hypertension)

Data: EHR for all hospital patients (timeseries of patient recordings)

Conventional workflow – from [Zhang et al., 2014] with extensions

At the beginning, there is a question/observation – a concrete phenomenon that must be explained (cf. use cases in [Zhang et al., 2014]).

1. The (team of) physician(s) devise one or more hypotheses.
2. The physicians specify the cohort needed for the study of each hypothesis, possibly in interaction with a data analyst or DB expert.
3. The DB expert writes scripts to create the cohort and extract the data.
4. Data analysts build models according to the instructions of the physicians, e.g. on age and gender adjustment.
5. Physicians become a presentation/visualization of the model(s) and check whether their hypothesis is supported.
6. If necessary, GOTO 2.

Cohort Exploration with Help of Visual Analytics

Visual Analytics in the medical domain [Zhang et al., 2014]:

- *Interaction:* Move away from the one-way paradigm of data analysis [7], by letting the user intervene in model learning
- *Visualization:* Deliver intuitive data representations that a domain-expert can understand, explore and manipulate
- *Provenance:* Monitor the progress of user-system interaction, record states, enable the sharing of results - of visual explorations

Example:

- System CAVA [Zhang et al., 2014] for Cohort Analysis with Visual Analytics designed for retrospective cohort studies on electronic health records

[7] The traditional virtuous circle of data mining closes after the model is inspected.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

How to increase interaction in cohort analysis? [Zhang et al., 2014]

- *Early cohort definition:* The physicians must be able to define a cohort themselves in an ad hoc way, whenever they see fit (cf. steps 2 and 3 of the conventional workflow).
- *Flexible visualization:* The physicians must be able to inspect the cohort in different ways, without having to ask the technologists.
- *Flexible analysis:* The physicians must be able to invoke analytics modules and use them to perform analytics tasks without having to ask the technologists.
- *Cohort refinement and expansion:* The physicians must be able to modify themselves the cohort, i.e. the choice of patients and the choice of variables on them (cf. steps 6 and 1 of the conventional workflow).
- *Iterative analysis:* Cohort definition, visualization, analysis, refinement and expansion may need to be performed repeatedly, on the results of the previous iterations.

i.e. foster interaction between physician and system in a complete workflow, taking the technologists out of the workflow.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

The elements of CAVA:

- **Cohorts:** Data construct

A cohort is a choice of individuals with their properties (feature space)

Inner feature space: set of properties shared by all cohort members
Outer feature space: set of all properties of the cohort members

- **Views:** Visualization components (library)

A view is a visualization component that

- presents a cohort to a user, and
- allows the user to modify the cohort interactively.

- **Analytics:** Computational elements (library)

An analytics component is

a piece of software that creates or modifies a cohort.

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

High-level architecture of CAVA (fig. 3, page 9)

PICTURE REMOVED

Data provenance

- *Population database:* contains all information about all individuals in the population; is expanded by new information (derived via analytics or views)
- *Cohort database:* contains the description of each cohort (as defined by the user) and the IDs of the cohort members

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Placing the CAVA elements into a workflow (fig. 5, page 11)

PICTURE REMOVED: picture showing the workflow from [Hielscher et al., 2014b]

Analytics components in CAVA

- *Batch analytics modules*, including a ”demographics module” and a ”risk stratification module”
- *On-demand analytics modules*, including a ”patient similarity component” (published in AMIA 2010), a ”utilization analysis component” (published in AMIA 2012) and a ”heart failure risk assessment component” (published in AMIA 2012)

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

CAVA Example 1: Building a cohort iteratively (fig. 6, page 14)

PICTURE REMOVED

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

CAVA Example 2: Analyzing a cohort interactively to find cardiac patients with high risk of re-hospitalization (fig. 7, page 15)

PICTURE REMOVED

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Evaluation by a domain expert - a very experienced emergency room physician, also having long experience in hospital management

Usability and design

- Ease-of-use and speed in comparison to the typical procedure: only a couple of days would be needed to build a cohort, in comparison to at least two weeks for answering basic questions
- More statistics are needed, next to the graphical views e.g. to conclude whether there were enough patients (in support of some finding)

Applicability to the challenges of healthcare

- Appropriate for quick and easy experimentation on patient groups
- Patient similarity function is a very promising aid:
  - + for finding similar patients, if the cohort being built is too small
  - + in combination with on-demand-analytics, which can show trends of interest to the physicians
- CAVA workflow agrees with the way things are being done
- Limited amount of patient detail, as physicians usually need also unstructured information (e.g. discharge summaries) and not only tables

Iterative Cohort Analysis and Exploration [Zhang et al., 2014]

Messages to take away from [Zhang et al., 2014]

- EHR are used to build cohorts. This must be done interactively.
- The workflow of cohort construction and analysis is central. It must be built around use-cases.
- Expert involvement is crucial. It must take place early on:
  - → Does the expert understand the environment?
  - → Can the expert use the environment in their everyday work?
  - → Can the expert's everyday work incorporate this environment?
- The physicians need a lot of information, which cannot be always predefined.
- Statistics are imperative but what statistics?

Learning Pressure Profiles for Patients with Diabetic Foot Syndrome

[Deschamps et al., 2013, Niemann et al., 2016a, Niemann et al., 2016b]

Why monitor DFS?

- Likelihood of foot amputation among patients with diabetic foot syndrom is up to 40 times higher than among non-diabetics.
- Increased foot temperature may indicate the onset of an ulceration.

How to monitor DFS?

- Monitor temperature
  - Detect increases of temperature – across days
  - Detect discrepancies between the temperature of the right and the left foot
- Pressure modulates temperature. ⇒ Monitor pressure

Pressure modulates temperature – Visualization

[Grützner et al., 2015]

PICTURES REMOVED: sequence of pictures showing the effect of plantar pressure on foot temperature, from [Grützner et al., 2015]

Learning Pressure Profiles for Patients with DFS

[Deschamps et al., 2013, Niemann et al., 2016a, Niemann et al., 2016b]

Why profiles?

Two possible learning tasks:

1. Understand what makes patients different from healthy people.
2. Find subpopulations of patients which are different from healthy ones and check what makes them different. Profile learning

On learning profiles:

- The method of choice is clustering.
- The target variable is hidden during learning.
- The target variable is revealed for the validation of the model.
- Variables known to influence the outcome must be considered when building the cohort.
- Variables that are known to modulate the outcome may not be used for learning.
- Clinical variables that are expected to be helpful in *explaining* the profiles must be collected for the cohort but may not be used for learning.

Classification of forefoot plantar pressure distribution in persons with diabetes [Deschamps et al., 2013]

Goal of the study

- Find groups of participants with similar ”forefoot loading” gait patterns
- Check whether there are groups of diabetics who can be separated (”isolated”) from healthy participants – i.e. who have different forefoot loading patterns

Gait analysis on patients with diabetes and on healthy subjects

- *Instrumentation:* a passive 3D motion analysis system with a 10 m walkway, with a plantar pressure platform & two force plates on it allowing for ”detection of specific gait events as well as a continuous calibration of the pressure plate with the AMTI force plate . . . ”.
- *Protocol:* Individuals walked barefoot at their own speed ”until five ‘representative’ [8] walking trials were recorded”

[8] ”A trial was considered representative if the participants made clear pedobarograph contact with good inter-trial consistency, judged by visual inspection of an experienced researcher.”

Clustering on forefoot loading [Deschamps et al., 2013]

Study subjects: 97 diabetics & 33 controls (45-70 Y, BMI 20-40)

- *patients:* no walking aids, no orthopaedic lower limb surgery, oedema score < 2, no active foot ulcer, no amputation, no Charcot neuroarthropathy
- *controls:* no orthopaedic lower limb surgery nor injury, no (known) neurological nor systemic disease

Clustering on forefoot loading [Deschamps et al., 2013]

Analysis

- K-Means clustering on the ”Relative regional Impulses” [9] of the hallux and of the 5 metatarsal regions of each foot
  - Euclidean distance of RrI after conversion into z-scores
  - 10 runs per K; the best run is chosen
  - best K is chosen by using silhouette coefficient

  for patients (best: K=4), for controls (best: K=3), for all participants together (best: K=4)
- Statistical analysis to determine ”statistical” (significant) differences between clusters
- Juxtaposition of the clusters with the characteristics the participants (including age, BMI and assessments)

[9] RrI is an aggregated signal, derived from the pressure recorded in the different regions.

Clustering on forefoot loading [Deschamps et al., 2013]

Main findings

- Distinct clusters that correspond to different forefoot loading profiles
- One cluster that consists only of diabetic feet and ”illustrates the poor contribution of the medial column of the forefoot to the overall weight bearing function of the forefoot”
- Most clusters in agreement with earlier studies that performed K-Means for pressure-based profiles

concluding that
”There seems to emerge a new era in diabetic foot medicine which embraces the classification of diabetic patients according to their biomechanical profile. Classification of the plantar pressure distribution has the potential to provide a means to determine mechanical interventions for the prevention and/or treatment of the diabetic foot.” (quoting from the Abstract)

Pressure Profiles for DFS-Patients [Niemann et al., 2016a]

Goal of the study: Understand how DFS-patients apply plantar pressure when they are standing.

- **Study subjects:** 20 patients (5F/15M, age 66.2 $+-$ 8.4 years)
  - diabetes duration: 16.2 $+-$ 11.7 years), type 1 or type 2 diabetes
  - sensomotoric peripheral polyneuropathy
  - Vibration threshold not exceeding 2/8 in the Rydel/Seiffer tuning fork test
- **Protocol:** Interchange of standing and resting phases
  - R-phase: resting, seated, for 5 min
  - S-phase: standing and applying pressure actively

  as follows:
  - *sequence:* SRSRS: S (5 min) – R – S (10 min) – R – S (20 min)
  - *trial*: SRSRS – R – SRSRS

Learning Pressure Profiles for DFS-Patients [Niemann et al., 2016a]

Pressure recordings in the different foot regions

PICTURE REMOVED

When do two participants apply plantar pressure the same way?

Pressure Profiles for DFS-Patients [Niemann et al., 2016a]

When do two participants apply plantar pressure the same way?
Two feet are similar, if they show similar pressure distributions on all regions.

Basis of computations: Relative Plantar Pressure

$$RPP = \frac{observedPlantarPressure - MIN}{MAX - MIN}$$

where MIN and MAX are computed over all S-phases of all sensors.

1. Distance defined over the average RPP [10] observed in a region r over the S phases [11] of all trials:

$$d_{RPP}(i,j) = \sqrt{\sum_{r=1}^{|R|} (\mu(i,r) - \mu(j,r))^2}$$

where $\mu(i,r)$ is the average RPP recorded for foot i in region r.

[10] We use average instead of peak pressure.

[11] We later concentrated on one S-phase only.

Pressure Profiles for DFS-Patients [Niemann et al., 2016a]

When do two participants apply plantar pressure the same way?
Two feet are similar, if they experience similar pressure distributions across all regions.

2. Distance defined over the pressure distribution in pairs of regions of each foot:
   Two feet are similar if the slopes of most of the $\binom{8}{2}$ regression lines are similar, whereby the goodness of fit of each line is taken into account.
3. Distance defined over the *centers of pressure* in the regions of each foot:
   - For each region r, cluster the average RPPs observed in it, producing a set of clusters $\xi(r)$.
   - the distance between two feet i,j for region r is the distance of the centers of the clusters to which the feet belong for this region.
   - Aggregate over all regions.

Pressure Profiles for DFS-Patients [Niemann et al., 2016a]

Workflow

PICTURE REMOVED

and results

TABLE REMOVED: table showing the performance of different algorithms for different similarity functions, from [Niemann et al., 2016a]

Pressure Profiles for DFS-Patients [Niemann et al., 2016a]

The 4 medoids of the best clustering

PICTURE REMOVED

Pressure Profiles of Patients vs Controls [Niemann et al., 2016b]

Going from a CS publication to a medical research publication:

- One clustering algorithm with one of the distance functions has the best performance. ⇒ Keep only the winner.
- Show that the findings are associated with diabetes.
  - Compare with a population of controls

Pressure Profiles of Patients vs Controls [Niemann et al., 2016b]

Q1: Why FOUR clusters?

Fig 2. ”Quality Assessment of k-medoids clustering using the Silhouette coefficient.”

PICTURE REMOVED

Pressure Profiles of Patients vs Controls [Niemann et al., 2016b]

Q2: How do we know that the clusters of the patients are different from those of the controls?

Fig 3. ”Summary of the clusters’ relative plantar pressure distribution.”

PICTURE REMOVED

Pressure Profiles of Patients vs Controls [Niemann et al., 2016b]

Q3: Do we see those clusters just because the controls are too different from the patients?
Q4: How do we know that those clusters are not *trivially explained*?

TABLE REMOVED

Table 1. ”Cluster description and composition, separated by DiabGr and ContrGr. There were no significant inter-cluster differences except for clusters 2 and 4 (height, weight and BMI) and clusters 3 and 4 (height); $\alpha = 0.05$.”

Pressure Profiles of Patients vs Controls [Niemann et al., 2016b]

Going from a CS publication to a medical research publication:

- √ Keep only the clustering algorithm + distance function that delivers the best results.
- ▶ Show that the findings are associated with diabetes.
  - √ Compare with a population of controls
  - √ Show that the diabetics' profiles are different from / the same as those of the controls.
- √ Show that each cluster is different from the others.
- √ Explain the clusters through variables that have not been used in clustering.

Closing Remarks

Big and Small Medical Data

- Cohorts have few individuals and large data spaces.
- Cohorts (esp. clinical ones) are often built from EHR collections with millions of records.
- Mining/ML is needed to explore those data, sometimes before and certainly after the cohort is built.

Closing Remarks

Methods that learn profiles from medical data

- Methods that find subpopulations that are interesting w.r.t. a medical outcome.
- Methods that explore the feature space and find best subspace(s).
- Methods that exploit expert knowledge during data exploration and feature space exploration.

Solutions for learning on systematically incomplete data

KEEP IN MIND: A man is not an average woman.

- Exploit similarity among patients w.r.t. some variables only
- Learn from systematically incomplete waves (see e.g. [Niemann et al., 2015]

Outlook

More methods are needed:

- Methods that help the expert understand and explore the model.
- Solutions that help us demonstrate that the model is.

The medical expert demands a *comprehensible* evidence that the model is correct and best for the data.

Research is needed on associating the models we produce with such evidence.

Thank you for your Attention!

Bibliography I

[Deschamps et al., 2013] Deschamps, K., Matricali, G. A., Roosen, P., Desloovere, K., Bruyninckx, H., Spaepen, P., Nobels, F., Tits, J., Flour, M., and Staes, F. (2013).
Classification of forefoot plantar pressure distribution in persons with diabetes: A novel perspective for the mechanical management of diabetic foot?
PLOS ONE, 8(11):e79924.

[Grützner et al., 2015] Grützner, J., Szczepanski, T., Kellersmann, J., Malanowski, J., Klose, S., and Mertens, P. R. (2015).
Smart diabetic insole - towards home feet health monitoring in order to prevent diabetic foot ulcer.
In *BMT*.

[Hall, 2000] Hall, M. A. (2000).
Correlation-based feature selection for discrete and numeric class machine learning.
In *Proc. of 17th Int. Conf. on Machine Learning*, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann.

[Hielscher et al., 2014a] Hielscher, T., Spiliopoulou, M., Völzke, H., and Kühn, J.-P. (2014a).
Mining longitudinal epidemiological data to understand a reversible disorder.
In *Proc. of Symposium on Intelligent Data Analysis*, pages 120–130.

[Hielscher et al., 2014b] Hielscher, T., Spiliopoulou, M., Völzke, H., and Kühn, J.-P. (2014b).
Using participant similarity for the classification of epidemiological data on hepatic steatosis.
In *Proc. of IEEE Symposium on Computer-Based Medical Systems*, pages 1–7.

Bibliography II

[Hielscher et al., 2016] Hielscher, T., Spiliopoulou, M., Völzke, H., and 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*.

[Ittermann et al., 2008] Ittermann, T., Schmidt, C. O., Kramer, A., Below, H., John, U., Thamm, M., Wallaschofski, H., and Völzke, H. (2008).
Smoking as a risk factor for thyroid volume progression and incident goiter in a region with improved iodine supply.
Europ. J. of Endocrinology, (159):761–766.

[Munson et al., 2014] Munson, M. E., Wrobel, J. S., Holmes, C. M., and Hanauer, D. A. (2014).
Data mining for identifying novel associations and temporal relationships with charcot foot.
Journal of Diabetes Research, 2014.

[Niemann et al., 2015] Niemann, U., Hielscher, T., Spiliopoulou, M., Völzke, H., and Kühn, J. (2015).
Can We Classify the Participants of a Longitudinal Epidemiological Study from Their Previous Evolution?
In *IEEE Symposium on Computer-Based Medical Systems*, pages 121–126.

[Niemann et al., 2016a] Niemann, U., Spiliopoulou, M., Samland, F., Szczepanski, T., Grützner, J., Ming, A., Kellersmann, J., Malanowski, J., Klose, S., and Mertens, P. R. (2016a).
Learning pressure patterns for patients with diabetic foot syndrome.
In *Proc. of IEEE Symposium on Computer-Based Medical Systems*.

Bibliography III

[Niemann et al., 2016b] Niemann, U., Spiliopoulou, M., Szczepanski, T., Samland, F., Grützner, J., Senk, D., Ming, A., Kellersmann, J., Malanowski, J., Klose, S., and Mertens, P. R. (2016b).
Comparative clustering of plantar pressure distributions in diabetics with polyneuropathy may be applied to reveal inappropriate biomechanical stress.
PLOS ONE.
accepted in August 2016.

[Niemann et al., 2014a] Niemann, U., Spiliopoulou, M., Völzke, H., and Kühn, J.-P. (2014a).
Interactive medical miner: Interactively exploring subpopulations in epidemiological datasets.
In *Demo Track of ECML PKDD 2014*, pages 460–463. Springer.

[Niemann et al., 2014b] Niemann, U., Spiliopoulou, M., Völzke, H., and Kühn, J.-P. (2014b).
Subpopulation Discovery in Epidemiological Data with Subspace Clustering.
Foundations of Computing and Decision Sciences (FCDS), 39(4):271–300.

[Niemann et al., 2014c] Niemann, U., Völzke, H., Kühn, J.-P., and Spiliopoulou, M. (2014c).
Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis.
Expert Systems with Applications, 41(11):5405–5415.

[Press et al., 1992] Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992).
Numerical Recipes in C.
Cambridge University Press, Cambridge, UK.

Bibliography IV

[Völzke et al., 2011] Völzke, H., Alte, D., Schmidt, C. O., Radke, D., Lorbeer, R., Friedrich, N., Aumann, N., Lau, K., Piontek, M., Born, G., et al. (2011).
Cohort profile: the Study of Health In Pomerania.
Int. J. of Epidemiology, 40(2):294–307.

[Zhang et al., 2014] Zhang, Z., Gotz, D., and Perer, A. (2014).
Iterative cohort analysis and exploration.
Information Visualization (Info Vis), pages 1–19.