Mining Cohorts & Patient Data: Challenges and Solutions for the Pre-Mining, the Mining and the Post-Mining Phases

Myra Spiliopoulou

IEEE Int. Conf. on Data Mining, New Orleans, Nov. 2017

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
- Selective forgetting on the stream
- Incorporating expert knowledge into the learning process

Introduction

Analysis of medical data:

- on cohorts for the identification of risk factors, the development of new diagnostic tests, the understanding of diseases and their evolution
- on social and mobile data for patient support and disease understanding
- on hospital data for clinical decision support

Cohorts [Glenn, 2005]

The term “cohort”

Quoting [Glenn, 2005], page 2: “The term *cohort* originally referred to a group of warriors or soldiers, and the term is now sometimes used in a very general sense to refer to a number of individuals who have some characteristic in common.”

The term “cohort” in “cohort analysis”

Quoting [Glenn, 2005], page 2: “Here and in other literature on cohort analysis, however, the term is used in a more restricted sense to refer to those individuals (human or otherwise) who experienced a particular event during a specified period of time. The kind of cohort most often studied by social scientists is the human *birth cohort*, that is, those persons born during a given year, decade, or other period of time.”

Cohort Analysis [Glenn, 2005]

The term "cohort analysis"

Quoting [Glenn, 2005], page 3: "The term *cohort analysis* is usually reserved for studies in which two or more cohorts are compared with regard to at least one dependent variable measured at two or more points in time."

Purposes of Cohort Analysis [Glenn, 2005], pages 1-2

- "Assessing the effects of aging"
- "Understand[ing] the sources and nature of social, cultural and political change."

Counter-examples – [Glenn, 2005], page 3

- *Cross-sectional study*: Comparison of different groups of individuals with respect to some characteristic/variable – such a study "is conducted with data collected at one point in time, or, more accurately, within a short period of time."
- *Panel study:* Comparison of the the attitudes of a group of individuals at two distinct timepoints – such a study "measures the characteristics of the same individuals at more than one point in time."

Cohorts and their Analysis

Two notes on terminology:

Term "cohort" — from [Glenn, 2005]

a set of individuals that have some characteristic in common

Term "cohort analysis" — from [Zhang et al., 2014]

analysis of data from one cohort

Workflows for Cohort Analysis in the Hospital

→ Workflow for the Medical Experts - Needs and a Solution

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

- **Goal:** get new insights about a population of patients (e.g. all patients of the cardiology unit who have hypertension)
- **Parties involved:** team of physicians + team of technologists
- **Data:** EHR of 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.

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

Expanding the workflow to incorporate . . . [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)

Figure 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)

Figure removed

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)

Figure 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)

Figure 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

Workflows for Cohort Analysis in the Hospital

- ✓ Workflow for the Medical Expert – Needs and a Solution
- → Workflow for a retrospective study – in the Expert's writing style
  - Case study: Identifying risk factors for Charcot Foot from EHRs

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 [1] or because it was a misdiagnosis [2] 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 $p < 0.001$ [3]

[1] unspecific ICD, e.g. ”viral infection, not otherwise specified”
[2] misdiagnosis like ”“gout, not otherwise specified”
[3] 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]

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 [4], esophagean reflux [5]

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

- Some of them preceded Charcot Foot Appropriate for diagnostic test?
  - ⋆ Alkalosis: 100% of the times
  - ⋆ Pulmonary eosinophilia – significantly often

[4] Pulmonary eosinophilia may be treated with steroids; these may affect bone mineral density.
[5] Esophagean reflux might be associated to proton pump inhibitors; -//- -//- -//-

Workflows for Cohort Analysis in the Hospital

- √ Workflow for the Medical Expert – Needs and a Solution
- √ Workflow for a retrospective study – in the Expert's writing style
- → Workflow for a prospective study – in the Expert's writing style
  - Case studies: Intelligent wearables for patients with Diabetic Foot Syndrome (DFS)

Learning Pressure Profiles for Patients with DFS

[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. $\Rightarrow$ Monitor pressure

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

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’ [6] walking trials were recorded”

[6] ”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” [7] 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)

[7] 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

Figure 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 [8] observed in a region r over the S phases [9] 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.

[8] We use average instead of peak pressure.
[9] 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 show similar pressure distributions on 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

Figure removed

and results

#	Sim.	Algorithm	k_{opt}	$Silh_{opt}$
1		k-medoids	4	**0.78**
2	Sim_{rpp}	DBSCAN	1 + 14 Outliers	0.57
3		Hierarchical	2	0.4 (Single L.)
4		k-medoids	4	0.31
5	Sim_{pairs}	DBSCAN	1 + 2 Outliers	0.13
6		Hierarchical	3	0.18 (Avg. L.)
7		k-medoids	4	0.45
8	$Sim_{centers}$	DBSCAN	1	0.12
9		Hierarchical	10	0.18 (Avg. L.)

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

The 4 medoids of the best clustering

Figure removed

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

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

Figure removed – article is open access

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.”

Figure removed – article is open access

Workflows for Cohort Analysis in the Hospital

- ✓ Workflow for the Medical Expert – Needs and a Solution
- ✓ Workflow for a retrospective study – in the Expert's writing style
- ✓ Workflow for a prospective study – in the Expert's writing style
- → Expert's Needs and Words

Placing the expert in the loop

- What does the expert want to say? [Amershi et al., 2014]

What does the expert want to say – and what not?

If the expert is willing to teach a learner:

- ActL: If users are repeatedly asked for labels, they may find this annoying or even "lose track of what they were teaching" [a]
- ReinfL: Users prefer to give positive rather than negative rewards [a]

[a] [Amershi et al., 2014], quoting from cite (Cakmak et al, 11).
[a] [Amershi et al., 2014],citing (Thomas & Braezal, 08) and (Knox & Stone, 12).

The expert also wants to do more: [Amershi et al., 2014]

- provide features, weights, changes in weights etc [a]
- experiment with different model inputs
- query the learner about its decisions [b].

[a] [Amershi et al., 2014], citing the findings of an experiment by (Stumpf et al, 07) on 500 different inputs for the improvement of a classifier.
[b] [Amershi et al., 2014], citing the findings on an interactive prototype by (Kulesza et al, 11).

What does the expert want to say – and what not?

The expert is not necessarily willing to provide labels.

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

Figure removed

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

Using the population-based longitudinal STUDY OF HEALTH IN POMERANIA – SHIP for case studies, e.g.

- Studying the role of smoking in the pathogenesis of thyroid enlargement [Ittermann et al., 2008]
- Building lumbar spine profiles to understand back pain [Klemm et al., 2013, Klemm et al., 2014]
- Learning to separate for the multifactorial disorder hepatis steatosis [Hielscher et al., 2014b]

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

- Formulate a hypothesis,
  e.g. on how smoking affects thyroid enlargement [Ittermann et al., 2008]
- 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

Learning from the study data, the mining way: **Exploit all the variables**

Why?

- Identify variables, the role of which was previously unknown
- Identify subpopulations (also small ones!) that have significantly higher exposure to (previously unknown) risk factors

Challenges:

C1: High-dimensional feature space

$+$ Protocol evolution $\Downarrow$

C2: Systematically incomplete data

[C1:] How to deal with a large feature space?

Prune before Learning

- Reduce the feature space by selecting the most informative variables that are minimally associated to each other e.g. [Hielscher et al., 2014b]
- Build subspaces that contain potentially interesting subpopulations, without revealing the target variable [Niemann et al., 2014a]

Prune after Learning

Perform classification rules discovery and

- drill-down into the rule space Interactive Med Miner v.1 [Niemann et al., 2014b] & v.2 [Schleicher et al., 2017]
- cluster the rules and build *representatives* [Niemann et al., 2017]

Prune during Learning

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

[C2:] How to exploit systematically incomplete timestamped data?

Exploiting systematically incomplete timestamped data

How to incorporate the unlabeled data into the learning process?

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

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

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

Turning sequences of values into new features Figure removed

- 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

Figure removed

stea_seq: important sequence-feature for the male subpopulation

ggt_s_seq: important sequence-feature for the male subpopulation

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

Figure removed

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

Most important evolution features

Figure removed

Exploiting temporal information and gaps:

- Clustering-before-classification
  - groups of similar people
  - groups of people that evolve similarly

  contributes to class separation and to the identification of additional informative variables.

- Feature space selection is an important pre-processing step, before similarities are computed.

Case Study: Understanding how tinnitus symptoms change during the day

Analyzing mHealth data to understand tinnitus symptoms

[Probst et al., 2017a]

Why monitor tinnitus?

- 5.1% to 42.7% of the population experience tinnitus (citing McCormack et al (2016)).
- Some tinnitus patients experience stress, depression, anxiety, fatigue, insomnia, some become even incapable of working.
- Cognitive Behavioural Therapy (CBT) has been shown to reduce the burden of tinnitus
- but patient response to treatment varies – to CBT and, even more, to other forms of treatment.
- One explanation for poor response and inconsistent results is *heterogeneity*:
  - inter-individual heterogeneity: tinnitus varies across patients
  - inter-individual heterogeneity: tinnitus for a patient varies over time

Remembering tinnitus symptoms [Pryss et al., 2017]

Remembering – why?

"The treatment of tinnitus and the early diagnosis of potential comorbidities require assessments on several symptoms, including loudness and variation of the perceived sound(s), distress caused by tinnitus, impact of tinnitus on sleeping behavior, comorbidities, social activity, concentration, and so forth."

Remembering – how well?

"Bratland-Sanda et al (2010) [5] assessed physical activities of patients with eating disorders by retrospective self-reports as well as . . . with an accelerometer. Patients reported significantly less physical activity retrospectively than what was measured prospectively by the accelerometer."

Recording instead of remembering

Ecological Momentary Assessments (EMA): observable variables (e.g. symptoms) are repeatedly *assessed* – citing Trull & Ebner-Priemer (2013) on "ambulatory assessments"

Smartphone-based Recording of Ecological Momentary Assessments

[Probst et al., 2017a, Pryss et al., 2017]

TrackYourTinnitus mobile app:

Registration for tinnitus monitoring

Three questionnaires:

- Mini-TQ-12 on tinnitus-related psychological problems
- TSCHQ (37) on tinnitus sample case history
- Worst Symptom Questionnaire (9)

to be filled once.

EMA on tinnitus

Seven questions on:

- tinnitus loudness
- distress through tinnitus
- valence and arousal

to be answered up to 12 times a day at randomly chosen moments.

- Ambient sounds are captured during each EMA recording.

Analyzing EMA on the TrackYourTinnitus app [Probst et al., 2017a]

Time-of-day dependence of tinnitus loudness and distress:

Materials

- Total assessments: 25,863 Retained: 17,209, after excluding assessments with missing values in any of the target variables and days with less than three assessments.
- 350 participants (253m/94f) with average age 45.4 (over 333, SD=12.1) and median since tinnitus onset 5.4Y (from 0 to 61.8Y)
- Median days per participant 11 (from 1 to 415) with median number of assessments per day 4 (from 3 to 18)

Specifying day and night intervals

- night: 12am–4am
- early morning: 4am–8am
- early evening: 4pm–8pm
- afternoon: 12pm-4pm
- late morning: 8am–12pm
- late evening: 8pm–12am

Analyzing EMA on the TrackYourTinnitus app [Probst et al., 2017a]

Time-of-day dependence of tinnitus loudness and distress:
Selection of findings

- "tinnitus was significantly louder in the late evening compared to the afternoon and early evening."
- "stress-level increased from morning to afternoon, decreased from afternoon to evening, and did not differ compared to the night"
- "Tinnitus was louder and more distressing when the level of stress was higher at a specific time-of- day compared to other times-of-day, when it was higher during a whole day compared to other days, and when it was higher during the whole assessment period for a given participant (compared to other participants)."
- "the effects of time-of-day on tinnitus loudness and tinnitus distress were still significant (i.e., after controlling for the effects of stress)."

Reaching the patients

[Probst et al., 2017b]

Same disease, same population?

Three sources

Source	*Type*	*Sample size*
Tinnitus Center of Univ Hospital Regensburg	outpatient clinic	3786
TrackYourTinnitus	mobile app	867
TinnitusTalk	self-help social platform	5017

Results of comparison

Significant differences in age, gender and time since tinnitus onset $(p < 0.05)$

- Age: TrackYourTinnitus users were younger
- Gender: more female users in TinnitusTalk
- Time since tinnitus onset: users of TrackYourTinnitus & TinnitusTalk had more often acute, resp. subacute tinnitus (less than 3M, resp. 4-6M) or tinnitus for more than 20Y

Understanding the patients

Different subpopulations – different media, different needs

TrackYourTinnitus: Clustering patient evolution [Unnikrishnan, 2017]

Figure removed

Monitoring opinions on treatments in TinnitusTalk [Dandage et al., 2017]

Figure removed

Tinnitus Center Univ Hospital Regensburg [Schneck et al., 2017]

Finding questionnaire entries that capture the loudness/handicap interplay

		Top-10 variables for L_H+
Both	8	THI:{Q10,Q12,Q13,Q16,Q17}, TQ:{Q7,Q10,Q15}
MT_{RF}	2	THI:{Q1,Q23}
LP_{RF}	2	THI:Q21,TQ:Q39

		Top-10 variables for $L+H\_$
Both	7	THI:{Q10,Q12,Q13,Q16,Q17}, TQ:{Q7,Q15}
MT_{RF}	3	THI:{Q1,Q15,Q25}
LP_{RF}	3	THI:{Q7,Q14,Q21}

Closing Remarks

Big and Small Data in Medicine:

- Clinical cohorts built from EHR collections are extracted from millions of records in very heterogeneous data spaces.
- Cohorts have few individuals and large data spaces.
- The cohort construction process requires human expertise and intelligent data access.
- Mining/ML is needed to explore those data, sometimes before and certainly after the cohort is built.

Closing Remarks

There are methods:

- to model the whole sample
- to find subpopulations that are interesting w.r.t. a medical outcome
- to explore the feature space and find best subspace(s)
- to exploit expert knowledge during data exploration and feature space exploration
- to learn from systematically incomplete data

but more work is needed to model the problem

Closing Remarks

New technologies in medicine and healthcare:

New ways for the medical researchers:

- to reach patients
- to understand how diseases evolve and how patients live with them
- to monitor the treatments they developed, as they work in everyday life

New challenges:

- How to monitor in the presence of noise?
- How to discern between patient behaviour as effected by the disease and as effected by less relevant external factors?

Outlook

More methods are needed:

- to help the expert understand a model
- to help the expert select hypotheses worth pursuing further
- to incorporate the expert's knowledge into the model
- to show to the expert that the model is

Acknowledgements

Grants:

- German Research Foundation - grant IMPRINT (2011-2014)
- German Research Foundation - grant OSCAR (2017-2019)
- Otto-von-Guericke-Univ. Magdeburg - innovation fonds

Cooperations:

- ”Predictors of Steatosis Hepatis” & ”Struma_Mining: Studying the Potential of Data Mining for the Analysis of Goitre” with Univ. Medicine Greifswald
- ”Data Mining and Stream Mining In Diabetology” with Faculty of Medicine, OVGU
- ”Tinnitus: Analysis of medical and other health-associated data with mining methods” with Univ. Med. Regensburg, Univ. Ulm, TinnitusTalk
  also under the auspices of TINNET and ESIT networks of excellence

Acknowledgements: The KMD Team

PhD students:

- Uli Niemann & Tommy Hielscher: SHIP cohort analysis
- Uli Niemann: experiment analysis for DFS patients
- Vishnu Unnikrishnan: TrackYourTinnitus (master thesis)

Students of bachelor and master degrees:

- Miro Schleicher: SHIP cohort analysis (Bachelor), TrackYourTinnitus (Master)
- Arne Schneck (Master), Sven Kalle (Bachelor): TinnitusDB
- Sourabh Dandage, Johannes Huber, Atin Janki: TinnitusTalk

Thank you for your Attention!

Bibliography I

[Amershi et al., 2014] Amershi, S., Cakmak, M., Knox, W. B., and Kulesza, T. (2014).
Power to the people: The role of humans in interactive machine learning.
AI Magazine, 35(4):105–120.

[Dandage et al., 2017] Dandage, S., Huber, J., Janki, A., Niemann, U., Pryss, R., Reichert, M., Harrison, S., Vessala, M., Schlee, W., Probst, T., and Spiliopoulou, M. (2017).
Patient empowerment through summarization of discussion threads on treatments in a patient self-help forum.
In *Proc. of Int. Conf. on Biomedical and Health Informatics (ICBHI 2017)*, Thessaloniki, Greece.

[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.

[Glenn, 2005] Glenn, N. D. (2005).
Cohort Analysis.
Quantitative Applications in the Social Sciences. SAGE, 2nd edition.

[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.

Bibliography II

[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.

[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, pages 761–766.

[Klemm et al., 2013] Klemm, P., Lawonn, K., Rak, M., Preim, B., Tönnies, K., Hegenscheid, K., Völzke, H., and Oeltze, S. (2013).
Visualization and Analysis of Lumbar Spine Canal Variability in Cohort Study Data.
In *Proc. of VMV - Vision, Modeling & Visualization*, pages 121–128.

[Klemm et al., 2014] Klemm, P., Oeltze-Jafra, S., Lawonn, K., Hegenscheid, K., Völzke, H., and Preim, B. (2014).
Interactive Visual Analysis of Image-Centric Cohort Study Data.
IEEE Trans. Vis. Graph., 20(12):1673–1682.

Bibliography III

[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., 2017] Niemann, U., Spiliopoulou, M., Preim, B., Ittermann, T., and 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*, Thessaloniki, Greece.

[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*.

[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.

Bibliography IV

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

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

[Probst et al., 2017a] Probst, T., Pryss, R. C., Langguth, B., Rauschecker, J. P., Schobel, J., Reichert, M., Spiliopoulou, M., Schlee, W., and Zimmermann, J. (2017a).
Does tinnitus depend on time-of-day? an ecological momentary assessment study with the "trackyourtinnitus" application.
Frontiers in Aging Neuroscience, 9:253.

[Probst et al., 2017b] Probst, T., Pryss, R. C., Langguth, B., Spiliopoulou, M., Landgrebe, M., Vesala, M., Harrison, S., Schobel, J., Reichert, M., Stach, M., and Schlee, W. (2017b).
Outpatient tinnitus clinic, self-help web platform, or mobile application to recruit tinnitus study samples?
Frontiers in Aging Neuroscience, 9:113.

[Pryss et al., 2017] Pryss, R. C., Probst, T., Schlee, W., Schobel, J., Langguth, B., Neff, P., Spiliopoulou, M., and Reichert, M. (2017).
Mobile crowdsensing for the juxtaposition of realtime assessments and retrospective reporting for neuropsychiatric symptoms.
In *Proc. of IEEE Symposium on Computer-Based Medical Systems (CBMS 2017)*, Thessaloniki, Greece.

Bibliography V

[Schleicher et al., 2017] Schleicher, M., Ittermann, T., Niemann, U., Völzke, H., and Spiliopoulou, M. (2017).
ICE: Interactive Classification Rule Exploration on Epidemiological Data.
In *Proc. of IEEE Symposium on Computer-Based Medical Systems*, Thessaloniki, Greece.

[Schneck et al., 2017] Schneck, A., Kalle, S., Pryss, R., Schlee, W., Probst, T., Langguth, B., Landgrebe, M., Reichert, M., and Spiliopoulou, M. (2017).
Studying the potential of multi-target classification to characterize combinations of classes with skewed distribution.
In *Proc. of IEEE Symposium on Computer-Based Medical Systems*, Thessaloniki, Greece.

[Unnikrishnan, 2017] Unnikrishnan, V. (2017).
Analysis of patient evolution on time series of different lengths.
Faculty of Computer Science, Otto-von-Guericke Univ. Magdeburg.
Master Thesis.

[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.