PART 1: KNOWLEDGE DISCOVERY FROM EPIDEMIOLOGICAL DATA

Myra Spiliopoulou
Knowledge Management & Discovery Lab
Faculty of Computer Science
Otto-von-Guericke University Magdeburg, Germany

@ECML PKDD 2014 – Nancy, France

Who am I?

- Business Informatics Professor in Univ. Magdeburg
- doing research in Data Mining
- focussing on evolution and change
- studying how diseases progress and patients evolve

SETTING THE SCENE

Famous success stories of mathematical epidemiology

1760: Bernoulli shows that variolation (against smallpox) could contribute to increasing life expectancy in France.

1854: John Snow analyzes the cholera outbreak in London and identifies a well of infected water as the epicenter of cholera spread.

Early example of **real-time** epidemiology

1911: Sir Ronald Ross finds that malaria is spread by the Anopheles mosquitos and builds a spatial model for the spread of malaria.

M. Marathe and A.K.S. Vullikanti (2013) "Computational Epidemiology", *CACM* 56(7), pp. 88-96, 07/2013, DOI:10.1145/2483852.2483871

Computational Epidemiology

- is an interdisciplinary area
- setting its sights on developing and using computer models
- to understand and control the
- spatiotemporal diffusion of disease through populations.

M. Marathe and A.K.S. Vullikanti (2013) "Computational Epidemiology", *CACM* 56(7), pp. 88-96, 07/2013, DOI:10.1145/2483852.2483871

Science in support of real-time epidemiology

- [assessing] pandemic risk
- [identifying] vulnerable populations
- [evaluating] available interventions
- [assessing] implementation possibilities
- [learning from] pitfalls & [promoting] public understanding

Fineberg and Wilson,
editorial from Science (2009)
on the role of (other) science(s) in policymaking,
in support of real-time epidemiology

M. Marathe and A.K.S. Vullikanti (2013) "Computational Epidemiology",
CACM 56(7), pp. 88-96, 07/2013, DOI:10.1145/2483852.2483871

Part of content removed
(copyright considerations)

M. Marathe and A.K.S. Vullikanti (2013) "Computational Epidemiology", *CACM* 56(7), pp. 88-96, 07/2013, DOI:10.1145/2483852.2483871

Epidemiology covers
more than
the spatiotemporal diffusion of diseases.

Diseases, Disorders, Impairments

Alzheimer's: degenerative disease of the brain;
progression cannot be stopped

Mild Cognitive Impairment: often precedes dementia

Glaucoma: degenerative disease of the eye;
progression can be stopped

Traumatic brain injury:
non-degenerative; can be healed (only partially?)

Hepatic steatosis: disorder of the liver;
progression (fat accumulation) can be stopped;
favors diseases that can be only partially healed

Epidemiology is ...

a scientific discipline
that provides reliable knowledge for clinical medicine focusing on prevention, diagnosis and treatment of diseases [15].

Research in epidemiology aims at

- characterizing *risk factors* for the outbreak of diseases
- evaluating the efficiency of certain treatment strategies

[15]
R.H. Fletcher and S.W. Fletcher (2011). *Clinical Epidemiology.* Lippincott Williams & Wilkins

B. Preim, P. Klemm, H. Hauser, K. Hegenscheid, S. Oeltze, K. Toennies and H. Voelzke (2014). "Visual analytics of image-centric cohort studies in epidemiology", *Visualization in medicine and life sciences III*, Springer.

What do epidemiologists want to find out?

- Risk factors and protective factors:
  - What factors (lifestyle, genetic vars) favour the impairment?
  - What factors are protective against it?
- Interventions:
  - How does the intervention affect a patient's health state?
  - How does the intervention affect disease progression?
- Progression:
  - How does the disease progress?
  - What affects the progression of the disease?
  - What affects the health state of a patient?

AGENDA

- Understanding the data
  - WHAT data are there?
  - WHY were they collected?
  - HOW were they collected?
  - Data reliability issues
- Specifying the learning tasks - Examples
  - Predicting a patient's health state
  - Understanding disease progression
  - Understanding the impact of an intervention
- Closing remarks
  - DOs and DONTs in epidemiological mining
  - Are the epidemiological data BIG?

THE DATA

- WHAT data are there?
- HOW were they collected?
- WHY were thery collected?
- HOW MUCH to rely on WHICH data?

What data are there?

Cohort studies

- Population-based
- Cross-sectional
- Randomized controlled trials

Sample types

- Random – reflects the prevalence of the outcome in the population
- Balanced (patients & controls) – for juxtaposition of patients to controls

Traumatic Brain Injury

- WHAT: Cross-sectional
- WHAT: longitudinal for patients but not for controls
- WHY: study patient evolution given intervention
- HOW: right box
- RELIABILITY CHECK on correlations between pre- & post-recordings, specification of target

- 15 TBI patients
  (from a Rehabilitation Centre where they underwent a neurorehabilitation)
  - age: 18-51 years (m=32.13)
  - education: 8-18 years (m=13.7)
  - time since injury at begin of study: 2-6 months (m=3.8)
  - duration of neurorehabilitation: 7-12 months (m=9.4)
- 14 controls matched for age (31.93), years of education (15.57) and gender
- MEG, neuropsychological assessments
  - Patients: pre-/post-neurorehabilitation
  - Controls: once

N.P. Castellanos, N. Paul, V.E. Ordonez, O. Demuynck, R. Bajo, P. Campo, A. Bilbao, T. Ortiz, F. del-Pozo and F. Maestu (2010) "Reorganization of functional connectivity as a correlate of cognitive recovery in acquired brain injury", *BRAIN* (133), 2365–2381, DOI: 10.1093/brain/awq174

DCE-MR Images on Breast Cancer

- WHAT: Cross-sectional
- WHY: study the DCE-MRI potential for tumor malginancy classification
- HOW: right box, [18]
- RELIABILITY CHECK on target variable & correlations among records

- 68 DCE-MRI
  (Dynamic Contrast-Enhanced Magnetic Resonance Images)
- 50 patients (age: 36-73, m=55)
- BENIGN: 31, MALIGNANT: 37
  confirmation carried out via
  - histopathologic evaluation or
  - follow-up studies after 6 to 9 months
- only lesions detected in MRI
- 1.0 T open MRI scanner

[18] U. Preim, S. Glaßer, B. Preim, F. Fischbach and J. Ricke (2012) "Computer-aided diagnosis in breast DCE-MRI – Quantification of the heterogeneity of breast lesions", *Europ. Journal of Radiology*, 81(7):1532–1538.

S. Glaßer, U. Niemann, P. Preim and M. Spiliopoulou (2013) "Can we distinguish between benign and malignant breast tumors in DCE-MRI by studying a tumor's most suspect region only?" In Proc. of 26th IEEE Int. Symposium on Computer-Based Medical Systems (CBMS'13)

Study of Health in Pomerania

- WHAT: longitudinal population-based study
- HOW: right box, citation

Two independent cohorts

Selection criteria:
- main residence in Pomerania (Germany)
- age 20-79

- Cohort SHIP
  - SHIP-0 (1997-2001): 4308
  - SHIP-1 (2002-2006): 3300
  - SHIP-2 (2008-2012): 2333
- Cohort SHIP-TREND
  - SHIP-TREND-0 (2008-2012): 4420

Recordings:
- sociodemographics
- somatographic tests
- medical/lab tests
- ultrasound & MRT

H. Voezke, D. Alte, ..., U. John and W. Hoffmann (2011) “Cohort profile: the Study of Health In Pomerania,” *Int. J. of Epidemiology* 40(2), 294–307

Hepatic Steatosis

- WHAT: Random sample
- WHY: study the potential of data mining for classification – outcome "hepatic steatosis"
- HOW: right box
- RELIABILITY CHECK on target variable and on correlations, treatment of NULL values

578 SHIP-2 participants (314 F, 264 M)

- NEGATIVE: 438, POSITIVE: 108+32
- derived from the fat accumulation in the liver (mrt_liverfat_s2) as:
  - A (negative): <= 10
  - B (positive): (10, 25]
  - C (positive): > 25

U. Niemann, H. Voelzke, J.-P. Kuehn and M. Spiliopoulou (2014) “Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis,” *J. of Expert Systems with Applications,* 41(11), 5405–5415

What data are there?

Cohort studies

- **Population-based**: "Cohort studies measure variables of interest at some early time point and follow the subjects to observe who develops the disease."
- **Cross-sectional**: "Case-control or cross-sectional studies identify odds ratios for the variable (or exposure) while controlling for confounders to estimate the relative risk."
- **Randomized controlled trials**: "Randomized controlled trials are the gold standard for determining relative risks of single interventions on single outcomes."

J.C. Weiss, S. Natarajan, P.L. Peissig, C.A. McCarty, and D.Page (2012) "Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records", AI Magazine, 33-45, Winter 2012, ISSN 0738-4602

What data are there?

Cohort studies
- Population-based
- Cross-sectional

Sample types
- Random – reflects the prevalence of the outcome in the population
- Balanced (patients & controls) – for juxtaposition of patients to controls

Reliability Check
- Veracity of variables, esp. target variable
- Correlations between variables
- Correlations between recordings
- Correlations between values of variables

+ Subpopulation-dictated NULL values

Caution:

- Epidemiological mining is supervised.

- This does not imply that there is a target variable in the data.

AGENDA

✔ Understanding the data
- WHAT data are there?
- WHY were they collected?
- HOW were they collected?
- Data reliability issues

❖ Specifying the learning tasks - Examples
- Predicting a patient's health state
- Understanding disease progression
- Understanding the impact of an intervention

❖ Closing remarks
- DOs and DONTs in epidemiological mining
- Are the epidemiological data BIG?

THE LEARNING TASKS

- ➜ Predicting a patient's health state
- ➜ Understanding disease progression
- ➜ Understanding the impact of an intervention

Predicting a patient's health state

Prediction for Traumatic Brain Injury

- Recovery after traumatic brain injury
  - P.J.Andrews, D.H.Sleeman, P.F.Statham, A.McQuatt, V.Corruble, P.A.Jones, et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. *J. of Neurosurgery*, 97:326–336, 2002.

- Outcome after traumatic brain injury
  - A. Brown, J. Malec, R. McClelland, N. Diehl, J. Englander, and D. Cifu. Clinical elements that predict outcome after traumatic brain injury: a prospective multicenter recursive partitioning (decision-tree) analysis. *J. of Neurotrauma*, 22:1040–1051, 2005.
  - A. Rovlias and S. Kotsou. Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables. *J. of Neurotrauma*, 21:886–893, 2004.
  - A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Head injury dataset

- WHAT: sample from NTDB after filtering and cleaning
  NTDB is not population-based!
- WHY: test the potenntial of an ANN to predict in-hospital death
- HOW: right box, citation (includes explanations on the test data)
- Reliability Check on data records, distribution of target variable in test set

Records from NTDB 6.2: positive head CT only

- 11 input variables:
  - age, sex
  - On-Scene: total GCS score and individual components
    Emergency Dept: total GCS score and individual components, first systolic blood, pressure
- Training set: 7769 records
  - 72% male, mean age of 39.1 years
  - mean total os-GCS =f 8.3 (eye=2.3, verbal=2.4, motor=3.6)
  - mean total ED-GCS = 8.5 (eye=2.4, verbal=2.4, motor=3.8)
- Test: 100 records, records with GCS=15 removed
  - 74% male, mean age of 37.1 years
  - mean total os-GCS = 7.8 (eye=2.2, verbal=2.3, motor=3.4)
  - mean total ED-GCS = 7.6 (eye=2.1, verbal=2.2, motor=3.3)
- Classes: in-hospital survival (75%), in-hospital death

National Trauma Data Bank:

- a national registry maintained by the American College of Surgeons
- ca. 3,000,000 records assembled from 712 hospitals, 2002 - 2007
- data points collected by the individual reporting hospitals
- data points verified by the American College of Surgeons for logical consistency and completeness but not for accuracy

A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Predicting the outcome of head injury

INPUT:
7769 records for training
100 records for testing

METHOD: ANN with "informative sampling"

OUTPUT:
- 30 ANN models

Split the training set in subsets of size p, $D_1,...,D_n$
Initialize a dedicated training set X
REPEAT
  FOR i=1...n
    1) Train 30 ANN models on subset Di
    2) Informative sampling
      - Compute difference between survival and death predictions per record
      - Add to X the record causing most disagreement
    3) Train the ANNs (with mutation) on X
  ENDFOR
UNTIL a plateau is reached

A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Predicting the outcome of head injury

INPUT:
7769 records for training
100 records for testing

METHOD: ANN with "informative sampling"

OUTPUT:
- 30 ANN models

EVALUATION:
- ensemble of top-5 models
- comparison to clinicians

Clinicians:

- 5 neurosurgery residents
- 4 neurosurgery staff physicians

Performance computation:

- Table of the 100 test patients:
  - one row per patient
  - row contains the 11 clinical variables
- Clinical predictions were made:
  - at one sitting
  - marked on the table
  - with no real-time feedback on performance

A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Part of content removed
(copyright considerations)

INPUT:
7769 records for training
100 records for testing

METHOD: ANN with "informative sampling"

OUTPUT:
- 30 ANN models

EVALUATION:
- ensemble of top-5 models
- comparison to clinicians

A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Part of content removed
(copyright considerations)

Limitations ?

INPUT:
7769 records for training
100 records for testing

METHOD: ANN with "informative sampling"

OUTPUT:
- 30 ANN models

EVALUATION:
- ensemble of top-5 models
- comparison to clinicians

A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Part of content removed
(copyright considerations)

Limitations ?

INPUT:
7769 records for training
100 records for testing

METHOD: ANN with "informative sampling"

OUTPUT:
- 30 ANN models

EVALUATION:
- ensemble of top-5 models
- comparison to clinicians

A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.

Understanding disease progression

Model learning from historical data

- Objective: to model
  - the progression of a disease, and eventually
  - the disease stages (e.g. at discrete timepoints of the observation horizon)

- Questions to ask in advance:
  - Do we know whether the disease is degenerative?
  - Are there treatments that can cure or slow down the progression of the disease?
  - Do we know whether some participants were subjected to treatment?

Learning disease progression with **no** temporal data

INPUT: cross-sectional data

METHOD: Pseudotemporal Bootstrap

OUTPUT:
- pseudo-timeseries model
- HMM built upon it

Given a labeled cross-section dataset of size T
and the corresponding TxT distance matrix:

1) initialize k pseudo-timeseries,
   each starting with a healthy entry
   and ending with a diseased entry

start & end entries: chosen randomly with replacement

2) build the shortest path between the two endpoints of each timeseries
3) derive a pseudo-timeseries model
4) for (h=classes+1, h++)
   train a HMM with h hidden states
   until the HMM captures disease features of interest

A. Tucker, and D. Garway-Heath (2010) "The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data", *IEEE Trans. on Inf. Tech. in Biomedicine*, 14(1), 79– 85.
Y. Li, S. Swift and A. Tucker (2013) "Modelling and analysing the dynamics of disease progression from cross-sectional studies", *J. of Biomedical Informatics*, 46(2), 266-274.

Pseudotemporal bootstrap for glaucoma prediction

INPUT: cross-sectional data

METHOD: Pseudotemporal Bootstrap

OUTPUT:

- pseudo-timeseries model

Visual Fields Dataset 1 – for learning:

- 162 participants
  - HEALTHY: 84, GLAUCOMATOUS: 78

Visual Fields Dataset 2 – for validation:

- 23 out of 255 patients with ocular hypertension
  - volunteers of a placebo-controlled trial (treatment for prevention of glaucoma onset)
  - clinical visits every ca. 6 months
  - reproducible VF loss (observed within a period of 6 years – median)
  - HEALTHY: 358, GLAUCOMATOUS: 229

Confirmation according to [5]

[5] AGIS, "Advanced Glaucoma Intervention Study. 2, visual field test scoring and reliability", *Opthalmology*, 101(8), 1445-1455, 1994.

A. Tucker, and D. Garway-Heath (2010) "The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data", *IEEE Trans. on Inf. Tech. in Biomedicine*, 14(1), 79– 85.

Part of content removed
(copyright considerations)

Experimental results

Y. Li, S. Swift and A. Tucker (2013) "Modelling and analysing the dynamics of disease progression from cross-sectional studies", *J. of Biomedical Informatics*, 46(2), 266-274.

Part of content removed (copyright considerations)

Further exploration through clustering:

- Computation of the expected values of the variables associated with each state
- Computation of the clustering values of the variables discovered using k-means
- Comparison to the mean values for normal and glaucomatous data

Y. Li, S. Swift and A. Tucker (2013) "Modelling and analysing the dynamics of disease progression from cross-sectional studies", *J. of Biomedical Informatics*, 46(2), 266-274.

Understanding the impact of an intervention

Intervention after Traumatic Brain Injury

- How does the intervention affect the observable?
  - A. Marcano-Cedeno, P. Chausa, A. Garcıa, C. Caceres, J.M. Tormos, and E.J. Gomez. "Data mining applied to the cognitive rehabilitation of patients with acquired brain injury", *J. of Expert Systems with Applications*, 40:1054–1060, 2013.

- To what extend does the intervention bring patients close to controls?
  - Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

RECALL

Traumatic Brain Injury

- WHAT: Cross-sectional
- WHAT: longitudinal for patients but not for controls
- WHY: study patient evolution given intervention
- HOW: right box
- RELIABILITY CHECK on correlations between pre- & post-recordings, specification of target

- 15 TBI patients
  (from a Rehabilitation Centre where they underwent a neurorehabilitation)
  - age: 18-51 years (m=32.13)
  - education: 8-18 years (m=13.7)
  - time since injury at begin of study: 2-6 months (m=3.8)
  - duration of neurorehabilitation: 7-12 months (m=9.4)
- 14 controls matched for age (31.93), years of education (15.57) and gender
- MEG, neuropsychological assessments
  - Patients: pre-/post-neurorehabilitation
  - Controls: once

N.P. Castellanos, N. Paul, V.E. Ordonez, O. Demuynck, R. Bajo, P. Campo, A. Bilbao, T. Ortiz, F. del-Pozo and F. Maestu (2010) "Reorganization of functional connectivity as a correlate of cognitive recovery in acquired brain injury", *BRAIN* (133), 2365–2381, DOI: 10.1093/brain/awq174

Part of content removed (copyright considerations)

Controls vs

Patients before and after treatment

Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

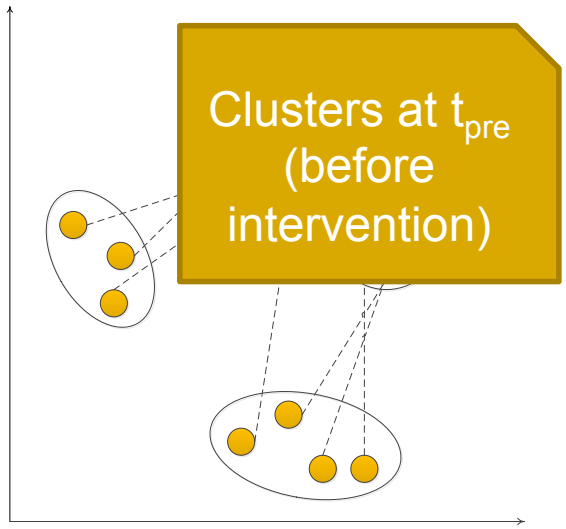
Improvements after TBI treatment

INPUT: TBI dataset

METHOD: EvolPredictor

OUTPUT: projected states of the patients

Workflow on Patient data



Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

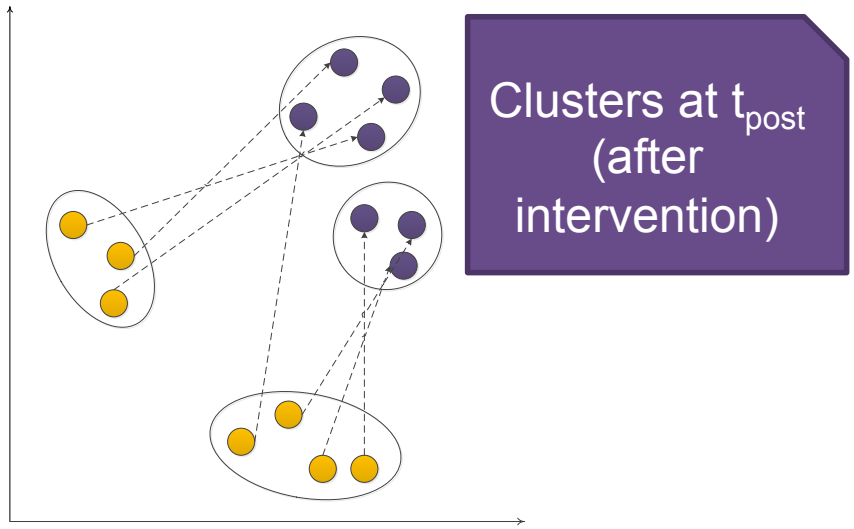
Improvements after TBI treatment

INPUT: TBI dataset

METHOD: EvolPredictor

OUTPUT: projected states of the patients

Workflow on Patient data



Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

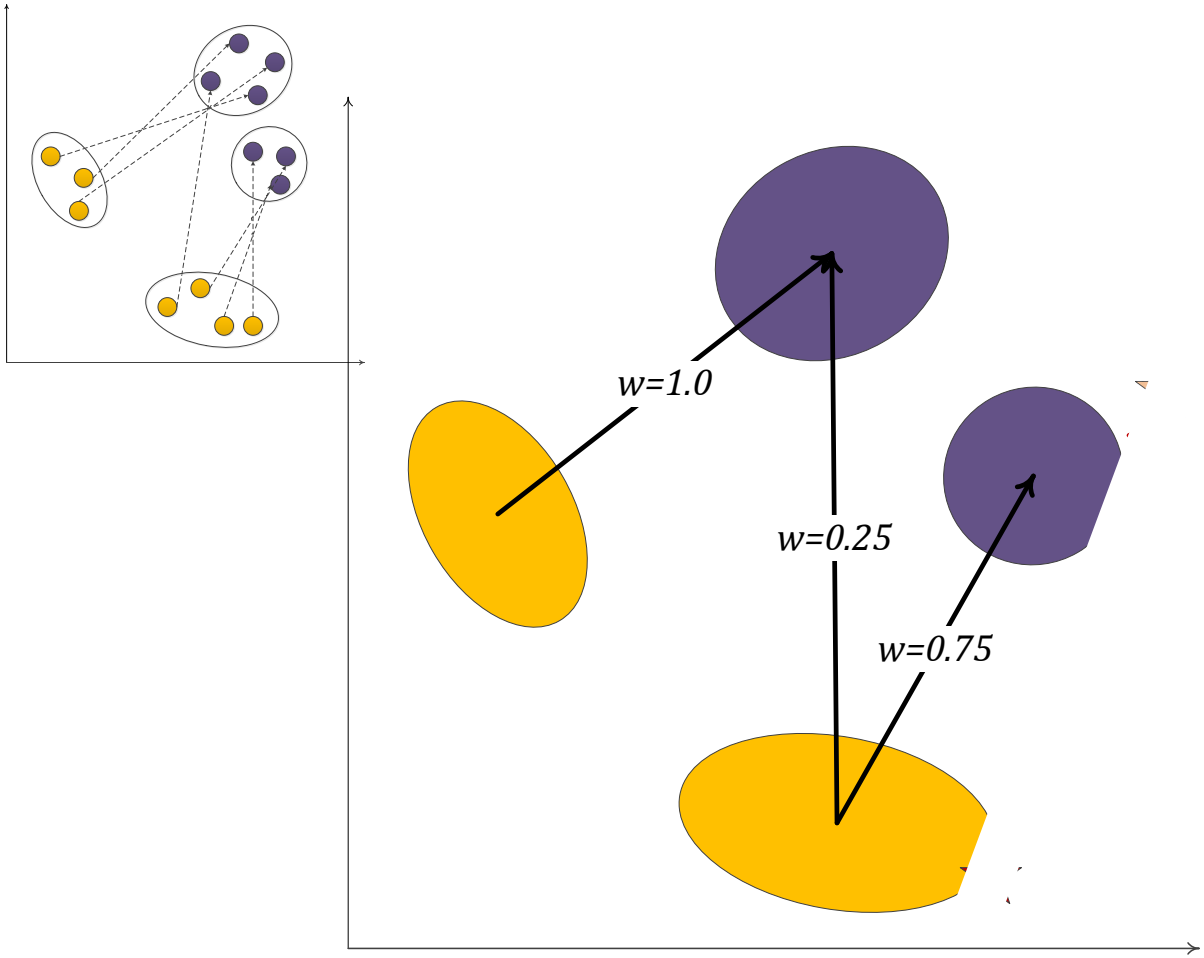
Improvements after TBI treatment

INPUT: TBI dataset

METHOD: EvolPredictor

OUTPUT: projected states of the patients

Workflow on Patient data



Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

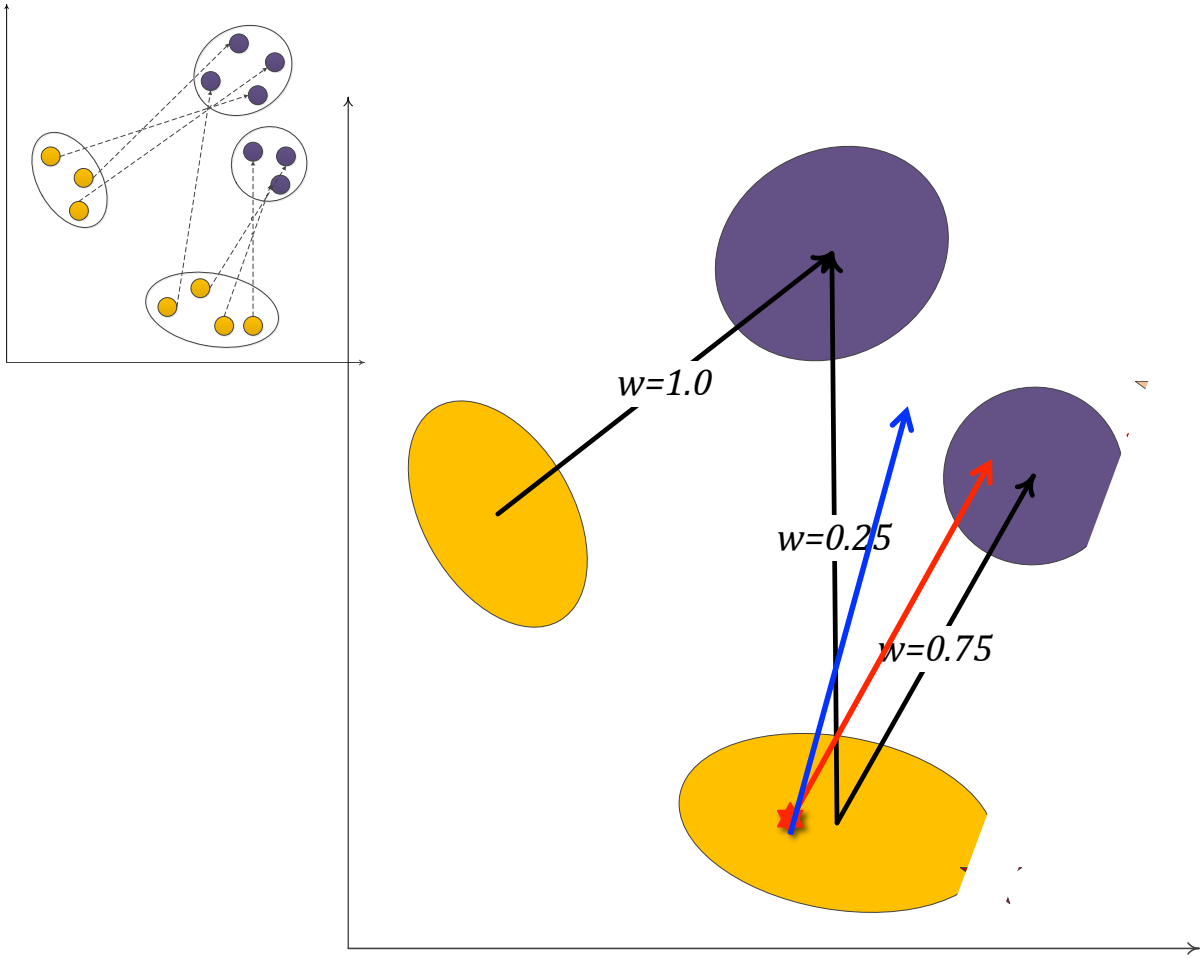
Improvements after TBI treatment

INPUT: TBI dataset

METHOD: EvolPredictor

OUTPUT: projected states of the patients

Learning for Prediction



Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

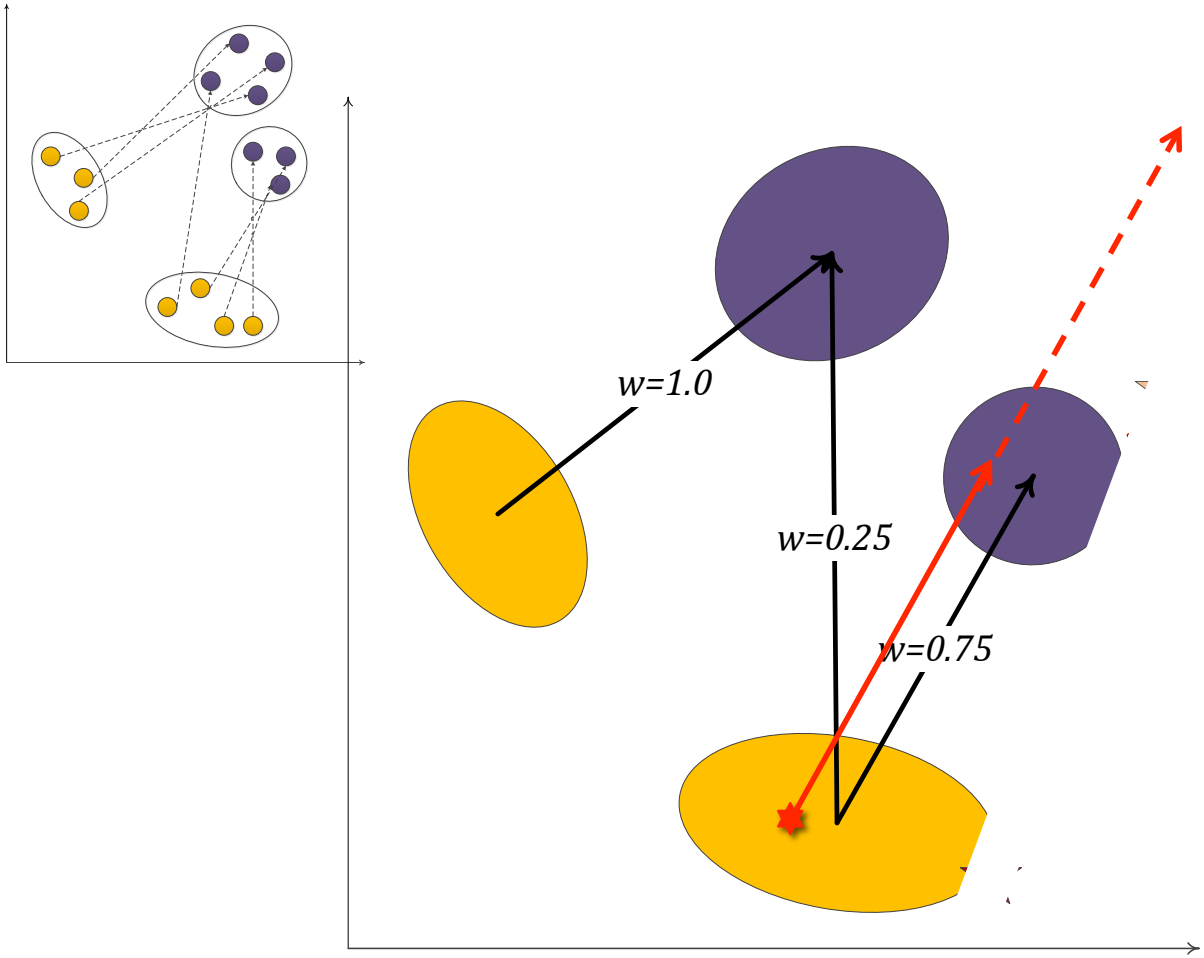
Improvements after TBI treatment

INPUT: TBI dataset

METHOD: EvolPredictor

OUTPUT: projected states of the patients

Prediction as projection



Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

Part of content removed
(copyright considerations)

Experimental results

Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

Improvements after TBI treatment

INPUT: TBI dataset

METHOD: EvolPredictor

OUTPUT: projected states of the patients

TODO:

- Is the effect of the intervention additive?
- What are the cluster semantics?
- Does the moment of intervention play a role? The duration?

➜ How to refine the model although the sample is so small?

Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

AGENDA

✔Understanding the data
- WHAT data are there?
- WHY were they collected?
- HOW were they collected?
- Data reliability issues

✔Specifying the learning tasks - Examples
- Predicting a patient's health state
- Understanding disease progression
- Understanding the impact of an intervention

❖Closing remarks
- DOs and DONTs in epidemiological mining
- Are the epidemiological data BIG?

DO'S & DONT'S IN EPIDEMIOLOGICAL MINING

ON Clustering

- Don't do clustering.

ON: Clustering for Personalized Medicine

Goal: deliver "personalized medicine" to each single patient

Problem: transfering insights from conventional models
(learned on population-based data)
to very small subgroups of people

EXAMPLE:
"... a 50-year-old man who runs every day may paradoxically have high levels of both good high-density lipoprotein (HDL) cholesterol, which helps to clear the arteries —high amounts of exercise can elevate it— and of bad low-density lipoprotein (LDL) cholesterol, which is a risk factor for coronary disease. Following conventional medical wisdom, the man's physician may want to prescribe medication to lower the LDL levels without actually knowing if it is necessary, because there is not a current capability to pull population-wide data on such a relatively small cohort. "

G. Groth (2012) *Analyzing Medical Data.* CACM 55(6), pp. 13-15, 06/2012, DOI:10.1145/2184319.2184324

ON: Clustering for Personalized Medicine

Goal: deliver "personalized medicine" to each single patient

Problem: transfering insights from conventional models
(learned on population-based data)
to very small subgroups of people

Approach: detect previously unspecified subpopulations of
people that share common determinants
(i.e. factors associated with an outcome)

ON: Clustering for Personalized Medicine

Goal: deliver "personalized medicine" to each single patient

Problem: transfering insights from conventional models (learned on population-based data) to very small subgroups of people

Approach: detect previously unspecified subpopulations of people that share common determinants (i.e. factors associated with an outcome)

➔ **DO** clustering only in the context of the target variable !

BIG
EPIDEMIOLOGICAL DATA?

BIG epidemiological data ?

- **Volume:**
  - Small sample size
  - BIG sample dimensionality
- **Variety:**
  - Almost all thinkable data types
- **Variability:**
  - Value range of each variable depends on recording protocol and
  - on hardware specifications
- **Velocity:**
  - Low for longitudinal studies
  - High for sensor recordings
  - Necessary for studies where evolution is relevant
- **Value:**
  - Indespensable for the advancement of medical research

ACKNOWLEDGEMENTS

Funding and Cooperations

- PROJECT at **German Research Foundation**

IMPRINT "Incremental Mining for Perennial Objects" (2011– 2014)

- GRANT from **Innovation Fonds of the OVGU**

- COOPERATIONS

StreaMED "Data Mining and Stream Mining for Epidemiological Studies on the Human Brain"
with the Center of Biomedical Technology (CTB), Madrid

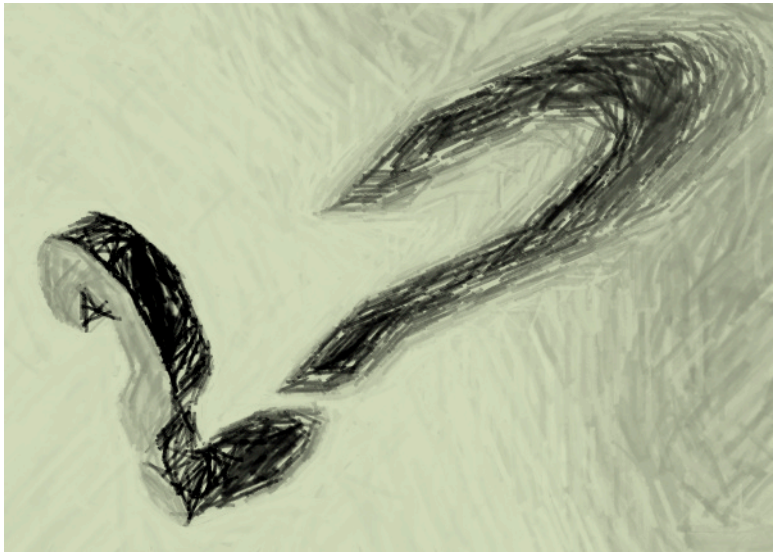
SHIP/2012/06/D "Predictors of Steatosis Hepatis"
with the University Medicine Greifswald

People

- KMD Team
  - Uli Niemann & Tommy Hielscher
  - Zaigham Faraz Siddiqui
  - Pawel Matuszyk & Georg Krempl
- Univ. Greifswald (Germany)
  - Henry Völzke
  - Jens-Peter Kühn
- Madrid
  - Fernando Maestu - CTB
  - Ernestina Menasalvas - Univ. Polytecnica de Madrid
  - Jose (Chema) Pena – Univ. Polytecnica de Madrid

Thank you very much!

Questions



LITERATURE

Analysis of epidemiological data
with traditional methods and with mining methods

Cited Literature I: General Issues

- R.H. Fletcher and S.W. Fletcher (2011). Clinical Epidemiology. Lippincott Williams & Wilkins
- G. Groth (2012)" Analyzing Medical Data", CACM 55(6), pp. 13-15, 06/2012, DOI:10.1145/2184319.2184324
- M. Marathe and A.K.S. Vullikanti (2013) "Computational Epidemiology", CACM 56(7), pp. 88-96, 07/2013, DOI:10.1145/2483852.2483871
- B. Preim, P. Klemm, H. Hauser, K. Hegenscheid, S. Oeltze, K. Toennies and H. Voelzke (2014). "Visual analytics of image-centric cohort studies in epidemiology", Visualization in medicine and life sciences III, Springer.
- J.C. Weiss, S. Natarajan, P.L. Peissig, C.A. McCarty, and D.Page (2012) "Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records", AI Magazine, 33-45, Winter 2012, ISSN 0738-4602
- C. Zhanga, R.L. Kodell. Subpopulation-specific confidence designation for more informative biomedical classification. Artificial Intelligence in Medicine 58 (3), 155-163, (2013)

Literature II: cohorts

- N.P. Castellanos, N. Paul, V.E. Ordonez, O. Demuynck, R. Bajo, P. Campo, A. Bilbao, T. Ortiz, F. del-Pozo and F. Maestu (2010) "Reorganization of functional connectivity as a correlate of cognitive recovery in acquired brain injury", *BRAIN* (133), 2365–2381, DOI: 10.1093/brain/awq174
- S. Glaßer, U. Niemann, P. Preim and M. Spiliopoulou (2013) "Can we distinguish between benign and malignant breast tumors in DCE-MRI by studying a tumor's most suspect region only?" In Proc. of 26th IEEE Int. Symp. on Computer-Based Medical Systems (CBMS'13)
- U. Niemann, H. Voelzke, J.-P. Kuehn and M. Spiliopoulou (2014) "Learning and inspecting classification rules from longitudinal epidemiological data to identify predictive features on hepatic steatosis," *J. of Expert Systems with Applications,* 41(11), 5405–5415
- U. Preim, S. Glaßer, B. Preim, F. Fischbach and J. Ricke (2012) "Computer-aided diagnosis in breast DCE-MRI – Quantification of the heterogeneity of breast lesions", *Europ. Journal of Radiology*, 81(7): 1532–1538.
- H. Voezke, D. Alte, ..., U. John and W. Hoffmann (2011) "Cohort profile: the Study of Health In Pomerania," *Int. J. of Epidemiology* 40(2), 294–307

Literature III: Progression of discussed impairments

- Y. Li, S. Swift and A. Tucker (2013) "Modelling and analysing the dynamics of disease progression from cross-sectional studies", *J. of Biomedical Informatics*, 46(2), 266-274.
- A.I. Rughani, T.M. Dumont, Z .Lu, J. Bongard, M.A. Horgan, P.L. Penar and B. Tranmer. Use of an artificial neural network to predict head injury outcome: clinical article. *J. of Neurosurgery*, 113:585–590, 2010.
- A. Tucker, and D. Garway-Heath (2010) "The pseudotemporal bootstrap for predicting glaucoma from cross-sectional visual field data", *IEEE Trans. on Inf. Tech. in Biomedicine*, 14(1), 79–85.
- Z.F. Siddiqui, G. Krempl, M. Spiliopoulou, J. M. Pena, N. Paul, and F. Maestu. "Are some brain injury patients improving more than others?" In Proc. of Int. Conf. on Brain Informatics & Health (BIH 2014), Special Session on Analysis of Complex Medical Data, Warsaw, Aug. 2014, Springer, LNAI 8609

Literature III- additional: Progression of discussed impairments

- P.J.Andrews, D.H.Sleeman, P.F.Statham, A.McQuatt, V.Corruble, P.A.Jones, et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. *J. of Neurosurgery*, 97:326–336, 2002.
- A. Brown, J. Malec, R. McClelland, N. Diehl, J. Englander, and D. Cifu. Clinical elements that predict outcome after traumatic brain injury: a prospective multicenter recursive partitioning (decision-tree) analysis. *J. of Neurotrauma*, 22:1040–1051, 2005.
- A. Marcano-Cedeno, P. Chausa, A. Garcıa, C. Caceres, J.M. Tormos, and E.J. Gomez. "Data mining applied to the cognitive rehabilitation of patients with acquired brain injury", *J. of Expert Systems with Application*s, 40:1054–1060, 2013.
- A. Rovlias and S. Kotsou. Classification and regression tree for prediction of outcome after severe head injury using simple clinical and laboratory variables. *J. of Neurotrauma*, 21:886–893, 2004.
- H. Y. Shi, S. L. Hwang, K. T. Lee, and C. L. Lin. In-hospital mortality after traumatic brain injury surgery: a nationwide population-based comparison of mortality predictors used in artificial neural network and logistic regression models. Journal of Neurosurgery, 118, 746-752, (2013)

Additional Literature on the progression of other impairments

- S. Ebadollahi, J. Sun, D. Gotz, J. Hu, D. Sow, and C. Neti. Predicting patient trajectory of physiological data using temporal trends in similar patients: A system for near-term prognostics,. AMIA Annu. Symp. Proc., vol. 2010, pp. 192-196, (2010)
- H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen. High-order multi-task feature learning to identify longitudinal phenotypic markers for Alzheimer's disease progression prediction. In Adv. in Neural Inf. Processing Systems 25, eds., P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, 1286-1294, (2012)
- J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In Proc. of KDD 2012, pages 1095-1103. ACM, (2012)