# Medical Mining Seminar – Winter term 2021/2022

## 1. OVERVIEW

**Goal** of this seminar is to make students familiar with underpinnings and methods for learning on medical data. At the end of the seminar, the students will be able to (a) distinguish basic types of medical data collections, (b) formulate workflows to solve learning problems for each type of collection, (c) build and run workflows for example case studies.

**Structure:** The seminar consists of 4 blocks after the introduction. Each block encompasses a theory part, a case study, a discussion of papers and a Small Student Project (ssp) on the case study. Students work in teams of up to three members. The 6 ECTS of the seminar can be achieved in two ways:

**a) Paper presentations only:** The team must present one paper from each block, picking papers from the block's collection and/or choosing from a larger collection.

**b) Paper + ssp:** The team must present one paper and complete on ssp. The paper can come from any block. The ssp can be chosen from the same or another block; it is also possible to propose an ssp.

**Examination:** The examination takes the form of "Homework Report" (Hausarbeit), to be delivered by each team one day after their ssp presentation. The Homework Report is submitted by each team consists of following **deliverables**:

a) The 4 presentations of the papers assigned to the team **and** a detailed description of the work of each team member for each paper

b) The presentation of the one paper assigned to the team **and** the presentation and the code of the ssp assigned to the team **and** a detailed description of the work of each team member for the paper and for the ssp

A presentation is done in class before being submitted as part of the Homework Report. The deliverable of the presentation consists of the presented slides and, optionally, of notes to each slide.

**Prerequisites:**

- background in data mining or machine learning or computational intelligence
- programming skills for data preparation tasks
- familiarity with DM/ML libraries

***Note 1:*** Successful completion of "Data Science with R" is of advantage.
***Note 2:*** Some ssp may demand background on streams or time series, and thus familiarity with DM II; some papers demand background in specific DM/ML/AI areas. The teams are advised to choose papers and ssp carefully.

---

## 2. CONTENT

In this seminar, we go through solutions for healthcare tasks that involve mining medical datasets. Medical datasets are of different types. Each type has particular challenges, and it permits only specific tasks, disallowing others. For example, consider the task of finding factors that lead to a particular choice of treatment for a disease: this task is valid when run on all clinical records of a clinic, but may result in misleading insights if only some of the clinic's records is available, and is certainly nonsense for a Randomized Clinical Trial (RCT) run on treatments offered by this clinic. We will see that these differences come from the data collection process.

The specification of the medical mining tasks affects our choice of learning algorithms, our learning work-flows and how we evaluate them. In this seminar, you will become familiar with solutions, i.e. learning workflows and their evaluation.

Scientific publications on medical datasets come in two colors:

- Scientific publications on methods that attempt to solve a general medical problem
  An example task is *"How to distinguish among patients who benefit / do not benefit from a given treatment?"* For most blocks, the papers come from KDD 2021. Teams may pick papers from each block's collection or choose themselves relevant papers from: ECML PKDD 2020 / 2021, KDD 2020 / 2021, IEEE ICDM 2020.
- Scientific publications on solutions for medical problems
  An example task is *"What characterizes the patients who benefit from a given treatment?"* For such solutions, the learning method is of lesser importance than the model it delivers. For this seminar, papers of this type come from journals like Nature Scientific Reports, Frontiers, PlosOne. Teams may decide to choose papers from these or other journals appearing in pubmed, focussing on diabetes, glaucoma, retinopathy or tinnitus.

**Block 1 "Introduction":**

1. Introduction to the seminar, discussion of the seminar modalities, overview of the blocks
2. Presentation of basic terminology, distinction between observational data and data of clinical studies
3. Medical datasets of type *"Cohort"*: example datasets and example tasks on each of them
4. Medical datasets of type *"Observational healthcare data"*: example datasets and tasks on each of them
5. Overview of the paper collections in each block

**Block TRP "Treatment Response Prediction":**

1. Problem specification
2. Core tasks of the learning workflow
3. TRP Extension A: Dealing with more than one outcome
4. In-class discussion of a paper on one of the example datasets
5. Presentation of papers by the student teams

| **Choice of papers** | | **Prediction of response to treatment, where "treatment" is any kind of intervention** | |
|---|---|---|---|
| KDD 2021 - ADS | x | Individual Treatment Prescription Effect Estimation in a Low Compliance Setting | Thibaud Rahier et al |
| PlosOne 2021 | | Two birds with one stone. –Addressing depressive symptoms, emotional tension and worry improves tinnitus-related distress and affective pain perceptions in patients with chronic tinnitus | Benjamin Boecking et al |
| KDD 2021 - ADS | | Interpretable Drug Response Prediction using a Knowledge-based Neural Network | Oliver Snow et al |
| Nature SciRep 2021 | | MLIP genotype as a predictor of pharmacological response n primary open-angle glaucoma and ocular hypertension | Maria I. Canut et al |
| KDD 2021 - RT | | Graph Infomax Adversarial Learning for Treatment Effect Estimation with Networked Observational Data | Zhixuan Chu et al |
| KDD 2021 - RT | | Multi-Objective Model-based Reinforcement Learning for Infectious Disease Control | Runzhe Wan et al |
| Frontiers in Public Health | | Toward Personalized Tinnitus Treatment: An Exploratory Study Based on Internet Crowdsensing | Jorge Simoes et al. |

**Block LITTLEDATA "Missingness and Scarcity":**

1. Data ***M**issing **C**ompletely **A**t **R**andom* (MCAR), ***M**issing **A**t **R**andom* (MAR), ***M**issing **N**ot **A**t **R**andom* (MNAR)
2. *Treatment Response Prediction* (TRP) Extension: TRP under MNAR
3. In-class discussion of a paper on one of the example datasets
4. Presentation of papers by the student teams

| **Choice of papers** | | **Dealing with Missingness and Scarcity – not necessarily for TRP** | | **Task is** |
|---|---|---|---|---|
| KDD 2021 - ADS | | Tolerating Data Missing in Breast Cancer Diagnosis from Clinical Ultrasound Reports via Knowledge Graph Inference | Jianing Xi et al | Diagnosis |
| JDSA 2021 | | Analyzing the impact of missing values and selection bias on fairness | Y. Wang and L. Singh | |
| KDD 2021 - ADS | | FLOP: Federated Learning on Medical Datasets using Partial Networks | Qian Yang et al | Diagnosis |
| Arxiv 2020 | | "A kernel to exploit informative missingness in multivariate time series from EHR",<br>(arxiv version of publication in Explainable AI in Healthcare and Medicine, 2021, pp. 23-36. Springer, Cham) | K. Mikalsen et al | |
| KDD 2021 - ADS | | Task-wise Split Gradient Boosting Trees for Multi-center Diabetes Prediction | Mingcheng Chen et al | Diagnosis |
| Pattern Recog 2021 | | Time series cluster kernels to exploit informative missingness and incomplete label information | K. Mikalsen et al | |
| KDD 2021 - ADS | x | Hierarchical Reinforcement Learning for Scarce Medical Resource Allocation with Imperfect Information | Qianyue Hao et al | Value/Vector prediction |
| Stats in Med 2021 | | Analyzing categorical time series in the presence of missing observations | C. Weiss | |

**Block COMPLIANCE:**

1. Problem specification, forms of compliance
2. *Treatment Response Prediction* (TRP) Extension: Missingness as indicator of non-compliance
3. In-class discussion of a paper on one of the example datasets
4. Presentation of papers by the student teams

| **Choice of papers** | | **Compliance as influence factor** | | **Task is** |
|---|---|---|---|---|
| KDD 2021 - ADS | | Diet Planning with Machine Learning: Teacher-forced REINFORCE for Composition Compliance with Nutrition Enhancement | Changhun Lee et al | Planing |
| KDD 2021 - ADS | x | Individual Treatment Prescription Effect Estimation in a Low Compliance Setting | Thibaud Rahier et al | TRP |
| Nature SciRep 2020 | | Understanding adherence to the recording of ecological momentary assessments in the example of tinnitus monitoring | Miro Schleicher et al | |
| JMIR 2018 | | Predicting Dropouts From an Electronic Health Platform for Lifestyle Interventions: Analysis of Methods and Predictors | Daniel Hansen Pedersen et al | |
| Translational Behavioral Medicine 2018 | | Predicting user adherence to behavioral eHealth interventions in the real world: examining which aspects of intervention design matter most | Amit Baumel et al | |

**Block FORECASTING:**

1. Problem specifications: forecasting as TRP extension, as event prediction, as value/vector estimation
2. Temporal representations for time series forecasting
3. Core tasks of the learning workflows
4. In-class discussion of a paper on one of the example datasets
5. Presentation of papers by the student teams

| **Choice of papers** | | **Forecasting – different problem specifications** | | **Task is** |
|---|---|---|---|---|
| KDD 2021 - ADS | | Causal and Interpretable Rules for Time Series Analysis | Amin Dhaou et al | Event prediction |
| KDD 2021 - ADS | | All Models Are Useful: Bayesian Ensembling for Robust High Resolution COVID-19 Forecasting | Aniruddha Adiga et al | Value/vector prediction |
| KDD 2021 - ADS | | PAMI: A Computational Module for Joint Estimation and Progression Prediction of Glaucoma | Linchuan Xu et al | Progression prediction |
| KDD 2021 - ADS | | Predicting COVID-19 Spread from Large-Scale Mobility Data | Amray Schwabe et al | Progression prediction |

---

## 3. GRADING SCHEME

a) Each of the 4 paper presentations makes 25% of the grade.
b) The paper presentation makes 25% of the grade, the ssp makes 75%.

**Grading of a paper presentation:**

- Explanation of the problem solved (10%)
- Explanation of the method used by the paper's authors (25%)
- Critical appraisal of the literature discussion (20%)
- Explanation of the evaluation criteria and the evaluation workflow (20%)
- Critical appraisal of the evaluation of the method (25%)

**Note 1:** The team earns bonus points for finding literature that the authors should have discussed but did not, and for finding flaws in the evaluation.
**Note 2:** Quotations from the paper are permitted, since they may be needed for the explanations. Quotations and rephrasings are not explanations; they contribute to the grading with **ZERO** points.

**Grading of the ssp:**

- Challenges of the problem (weight with values: 0.5, 1.0, and exceptionally 1.5 for difficult problems)
- Originality of the proposed workflow (25%)
- Effort invested in the proposed workflow, including data cleaning and preparation (25%)
- Technical quality of the solution (25%)
- Quality of the evaluation, including algorithm performance and test(s) for significance (25%)

**Note 3:** Each team member is graded separately on the basis of their contributions to the paper(s), resp. the ssp.
**Note 4:** Team size is taken into account in the weight scheme; large teams should work on more challenging problems.
**Note 5:** A student may also work alone, as team-of-one. The additional effort is considered in the weight scheme for the ssp. There is no weight scheme for the papers.

## 4. TIMEPLAN (tentative)

| Date | Block | Description | Deadlines |
|---|---|---|---|
| Oct 14, 2021 | 1 | Introduction to the seminar | |
| **Oct 21** | TRP | Parts 1, 2 | **Teams specified** |
| Oct 28 | TRP | Parts 3; papers and ssp's assignments | |
| Nov 4 | LITTLEDATA | Parts 1,2 | |
| **Nov 11** | LITTLEDATA | Parts 3; papers and ssp's assignments | **Registration done** |
| Nov 18 | Advisory meeting | TRP papers & ssp, LITTLEDATA papers & ssp | |
| Nov 25 | COMPLIANCE | Parts 1,2 | |
| Dec 2 | COMPLIANCE | Parts 3; papers and ssp's assignments | |
| Dec 9 | Advisory meeting | COMPLIANCE papers & ssp | |
| Dec 16 | FORECASTING | Block content; papers and ssp's assignments | |
| Dec 23 | Christmas break | -- | |
| Dec 30 | Christmas break | -- | |
| Jan 6, 2022 | Epiphany | -- | |
| Jan 13 | Advisory meeting | FORECASTING papers and ssp | |
| Jan 20 | Presentations I | TRP papers, LITTLEDATA papers | |
| Jan 27 | Presentations II | COMPLIANCE papers, FORECASTING papers | |
| **Early February** | Presentations III | All ssp presentations, meetings to be scheduled | **Homework to be submitted in the next day at 18:00** |