Medical Data Science
Diploma of Advanced Studies

Heidelberg University Hospital

Institute of Medical Biometry and Informatics

Department of Medical Biometry

Module Description

Date: 05.05.2020

1 year, 38 ECTS

Summary of the program

Degree of the study program	Diploma of Advanced Studies
Name of the university	Ruprecht-Karls-University Heidelberg
Name of the responsible unit	Department of Medical Biometry at the Institute of Medical Biometry and Informatics
Title of the program	Medical Data Science
Study program	Postgraduate study program, block courses of 2-3 days
Version / date	Version 2.0, 05.05.2020
Amounts of ECTS	38 ECTS
Duration	1 year (2 semester)
Short overview of the modules	Data Scientist's Toolbox, Statistical Modelling, Machine Learning, Practical Application
Target competencies	The present program equips students with statistical and computational methods for managing and analyzing large and complex data sets. Additionally, students learn how to extract and present information from these data sets in a meaningful way.

Content

1 Introduction

The program *"Medical Data Science"* is a postgraduate study program which is designed to provide a deeper, more specialized knowledge of statistical tools to analyze (big) data sets in medical research projects. Students learn to manage, analyze, and visualize the data, as well as to provide appropriate reports and interpretation of results. Besides theoretical considerations, applications are especially focused.

In the following sections, the modules with the respective courses are described.

2 Overview

Overall, the program consists of four modules. Table I gives an overview of the modules with the respective examinations and the credit points (ECTS) assigned to each course. Additionally, the time of attendance in days is given. The ECTS for each course include preparation and post-processing time. In total, the program consists of eight courses (30 ECTS) and a project work (8 ECTS) of three month preparation time.

Table I: Overview of modules of the program "*Medical Data Science*".

Modules	ECTS, Days	Examination
1 Data Scientist's Toolbox (M1)		
1.1 Introduction into Data Science	2, 1.5	
1.2 Working with Data, Plotting, Reproducibility, and Presentation	4, 2.5	
Total number of ECTS	6	Homework
2 Statistical Modelling (M2)		
2.1 Regression Methods	4, 2.75	
2.2 Generalized Additive Models	4, 2.75	
2.3 Bayesian Statistics	4, 2.75	
Total number of ECTS	12	Written examination
3 Machine Learning (M3)		
3.1 Supervised Learning	4, 2.75	
3.2 Unsupervised Learning	4, 2.75	
Total number of ECTS	8	Written examination
4 Practical Applications (M4)		
4.1 Data Science in Practice	4, 2.75	
4.2 Project Work	8, 3 month	
Total number of ECTS	12	Project thesis

3 Description of the modules

3.1 Module Data Scientist’s Toolbox (M1)

Title of the module: Data Scientists Toolbox	
Prior knowledge	None
Responsible persons	Dr. Marietta Kirchner
Didactics	Subject matter will be taught by alternating teacher-oriented presentations with prolonged practical tasks. Students will be encouraged to find own solutions by discussing the practical tasks in-class under the supervision of the teacher.
Topics	The module is divided into two courses: “Introduction into Data Science” and “Working with Data, Plotting, Reproducibility, and Presentation”. 1. Course: “Introduction into Data Science” Medical Data Science: definition and applications • Stake out term “medical data science” • Broad overview of topics and applications covered in subsequent courses General introduction in R programming language 2. Course: “Working with Data, Plotting, Reproducibility, and Presentation” Working with data: • Importing data from various sources (SAS, SPSS) (‘haven’ package) • Visualizing data using a ‘Grammar of Graphics’ (‘ggplot2’ package) • Transforming data (‘dplyr’ package) • Basics of relational data bases • Tidy data • Workflow advice and functional programming (‘purrr’ package) Reproducibility: • Why reproducible research is essential to good scientific practice • RMarkdown and knitr for automatic report generation • Package dependencies / CRAN / MRAN (‘checkpoint’ package) • ‘packrat’ package • Outlook: container-based workflow using Docker

	Presentation: - Creating interactive analyses using Shiny - Deploying interactive analyses as Web-apps
Acquisition of competencies	The participant is able to differentiate between medical data science and "classical" biostatistics and knows about potential fields of application. He/she is able to program in R. The participant is able to import data from a wide variety of sources in the R-environment. He/she understands the basic structure of relational databases and is able to perform SQL joins and transforms in R. He/she is able to visualize data using a systematic, grammar-based approach. The participant is aware of the need for reproducibility of his/her research and is able to implement basic technical reproducible research methods to generate reproducible reports. The participant knows about the possibilities of making reports interactive using the Shiny technology and is able to present his/her work using web applications.
Workload	Attendance time: 15h + 20h = 35h Preparation and post processing time: 40h + 90h Total effort: 165h
Module examination	Graded homework: exploratory presentation of a data set of interest; reproducible report of web app possible
Single booking costs	645 € for the second course "Working with Data, Plotting, Reproducibility, and Presentation" (reduced rate: 430 €)

3.2 Module Statistical Modelling (M2)

Prior knowledge	The participants must have - Basic knowledge of statistics and probability theory - Basic knowledge in R
Responsible persons	Dr. Lorenz Uhlmann
Didactics	The contents are taught in the form of conventional lectures. The lectures consist of various forms of teaching, e.g. discussions, group work, and classical teacher-centered parts. A special emphasis is put on practical training phases where the students learn to apply the taught methods.
Topics	This module provides an introduction to regression modeling strategies and Bayesian statistics and consists of three courses. The first course covers "Regression Methods" and comprises the following topics: - Linear and nonlinear regression (exponential family, link function)

	- Variable or model selection methods (Subset selection, forward, backward, and stepwise selection) - Model evaluation (Akaike/Bayesian Information Criterion (AIC, BIC), Deviance, Mellow' Cp, Mean squared error, Brior Score) - Resampling methods (Bootstrapping, jackknife, cross-validation) - Implementation in R The second course covers "Generalized Additive Models" (which are an extension of regression methods) and comprises the following topics: - Polynomial functions of covariates - Modeling using splines - Non-parametric modeling of covariates - Implementation in R The third course covers "Bayesian Statistics" and comprises the following topics: - Bayes' Theorem - Bayesian linear and non-linear regression models - Markov Chain Monte Carlo Methods and Gibbs sampling - Implementation in JAGS and R
Acquisition of competencies	The students are familiar with (Bayesian) regression modeling strategies. They know how to apply the taught methods and which assumptions are to be met to obtain sensible results. Furthermore, they know how to interpret the results. They are able to advise practitioners on statistical (regression) models as well as plan and implement these models (in R) and present the results in an appropriate way.
Workload	Attendance time: 3*22h=66h. Preparation and post processing time: 3*90h= 270h. Total effort: 336h.
Module examination	Written exam. Duration: 90 min.
Single booking costs	645 € per course (reduced rate: 430 €)

3.3 Module Machine Learning (M3)

Prior knowledge	The participants must have - Basic knowledge of statistic and probability theory - Knowledge of regression analysis.
Responsable persons	Dr. Katharina Hees and Dr. Lorenz Uhlmann
Didactics	The contents are taught in the form of conventional lectures. The lectures consist of various forms of teaching, e.g. discussions, group work, and classical-teacher

	centered parts. A special emphasis lies also on practical training phases where the students learn to apply the taught methods independently.
Topics	This module provides an introduction to machine learning and statistical pattern recognition. The module is divided into two courses: “Unsupervised Learning” and “Supervised Learning”. In the course “Unsupervised Learning”, methods to describe associations and patterns in data are discussed. The topics included are: • Clustering • Dimension reduction • Introduction to deep learning • Generative models Topics included in the “Supervised Learning” course: • Regularization methods for linear regression • Model assessment and selection • Neural networks • Decision trees • Random forests • Bagging and boosting
Acquisition of competencies	The participant knows the basic methods of supervised and unsupervised learning. He/She can decide, which of the methods is appropriate in a special situation. Furthermore, he/she is able to apply the methods to data (using R) and to interpret the results correctly.
Workload	Attendance time: 2*22h = 44h. Preparation and post processing time: 2*90h = 180h. Total effort: 224h.
Module examination	Written examination, duration: 90 min.
Single booking costs	645 € per course (reduced rate: 430 €)

3.4 Practical Applications (M4)

Title of the module: Practical Applications	
Prior knowledge	Content of modules M1, M2, and M3
Responsable persons	Prof. Dr. Meinhard Kieser, Dr. Lorenz Uhlmann, Dr. Marietta Kirchner
Didactics	Teaching forms are mainly group work, presentations, and discussions in the plenary. Additionally, independent working phases supervised by responsible persons are included.
Topics	This module consists of two parts: 1. “Data Science in Practice” and 2. “Project Work”. The course “Data Science in Practice” includes working with data-analytic methods, which are taught in the first three modules.

	Students will work in small groups on practical problems in the field of data science. The course focuses on tackling methodological problems in the analysis of the data and on presenting and discussing the results. The second part of this module is the project thesis which concludes the study program. The project thesis should be stimulated by a practical problem which can be an extension of the material discussed in the course "Data Science in Practice". Students work independently on their project.
Acquisition of competencies	• Gather practical experience • Learn to work in small groups with (large) datasets • Presentation and discussion of results • Consolidation of acquired knowledge • Independent scientific study
Workload	Attendance time: 20h Preparation and post processing time: 90h + 220h Total effort: 330h
Module examination	Presentation of results (not graded) and project thesis
Single booking costs	645 € for the first course "Data Science in Practice" (reduced rate: 430 €)

4 Appendix

In the following, we show some examples of timetables.

4.1 Schedule of the course “A Data Scientist’s Toolbox”

Time	Thursday	Friday	Time	Saturday
9.00 – 10.30	Getting data in R and exploratory data analysis using systematic visualizations	Why reproducibility? Reproducible reports using RMarkdown + knitr	**9.00 – 10.30**	Interactive reports using Shiny
10.30 – 11.00	Coffee Break	Coffee Break	**10.30 – 10.45**	Coffee Break
11.00 – 12.30	Practical exercise in small groups: Importing, cleaning, transformation and presentation	Practical exercise (small groups): Transfer yesterday’s results in a reproducible report [‘Prettiest Report Contest’]	**10.45 – 12.30**	Practical exercise: Publishing an interactive analysis as Web-App
12.30 – 13.30	Lunch Break	Lunch Break		
13.30 – 15.00	Advanced data manipulation using nested data frames and functional programming	Advanced methods for reproducible research: packrat, MRAN, Docker		
15.00 – 15.30	Coffee Break	Coffee Break		
15.30 – 17.00	Practical exercise in small groups: Nested data frames and functional programming	Practical exercise (individual): Using packrat with RMarkdown; Checkpoint package; Running R in a Docker container on Windows		

4.2 Schedule of the course “Regression Methods”

TIME	THURSDAY	TIME	FRIDAY	TIME	SATURDAY
9.00 – 10.30	Linear Regression I	9.00 – 10.30	Generalized Linear Models	9.00 – 10.30	Survival Analysis
10.30 – 11.00	COFFEE BREAK	10.30 – 11.00	COFFEE BREAK	10.30 – 11.00	COFFEE BREAK
11.00 – 12.30	Linear Regression II	11.00 – 12.30	Generalized Linear Models II	11.00 – 12.30	Survival Analysis
12.30 – 13.30	LUNCH BREAK	12.30 – 13.30	LUNCH BREAK		
13.30 – 15.00	Model Evaluation	13.30 – 15.00	Variable Selection		
15.00 – 15.30	COFFEE BREAK	15.00 – 15.30	COFFEE BREAK		
15.30 – 17.00	Resampling Methods	15.30 – 17.00	Mixed Models		

4.3 Schedule of the course “Supervised Learning”

Time	**Thursday**	**Friday**	**Time**	**Saturday**
9.00 – 10.30	Introduction (Machine Learning vs. Data Mining, Supervised vs. Unsupervised, etc.)	Model Assessment and Selection I (e.g. Bias and Variance, AIC, BIC, Subset Selection, Cross Validation, Bootstrap)	**9.00 – 10.30**	Prototype methods
10.30 – 11.00	Coffee Break	Coffee Break	**10.30 – 10.45**	Coffee Break
11.00 – 12.30	Regularized regression methods I	Model Assessment and Selection II (e.g. Bias and Variance, AIC, BIC, Cross Validation, Bootstrap Methods)		Tree based methods (e.g. Decision Trees, Random Forests)
12.30 – 13.30	Lunch Break	Lunch Break	**12.15 – 12.45**	Lunch Break
13.30 – 15.00	Regularized regression methods II	Neural Networks and Deep Learning	**12.45 – 14.15**	Ensemble Methods (e.g. Bagging, Boosting)
15.00 – 15.30	Coffee Break	Coffee Break		
15.30 – 17.00	Regularized regression methods II	Neural Networks and Deep Learning		