

Medical Data Science
Diploma of Advanced Studies

Heidelberg University Hospital
Institute of Medical Biometry

Module Description

Date: 26.01.2026

1 year, 36 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	Institute of Medical Biometry
Title of the program	Medical Data Science
Study program	Postgraduate study program, block courses of 2-3 days
Version / date	Version 6.0 / 26.01.2026
Amounts of ECTS	36 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..... 4
- 2 Overview..... 4
- 3 Description of the modules 5
 - 3.1 Module Data Scientist’s Toolbox (M1) 5
 - 3.2 Module Statistical Modelling (M2)..... 6
 - 3.3 Module Machine Learning (M3)..... 8
 - 3.4 Practical Applications (M4)..... 9
- 4 Appendix..... 10
 - 4.1 Schedule of the course “Basics of Data Science with R” 10

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. The ECTS for each course include attendance time, preparation and post-processing time. In total, the program consists of seven mandatory courses (28 ECTS), an optional course and a project work (8 ECTS) of three-month preparation time.

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

Modules	ECTS	Examination
1 Data Scientist’s Toolbox (M1)		
1.1 Introduction into R (optional)	(no ECTS)	
1.2 Basics of Data Science with R	4	
Total number of ECTS	4	Homework
2 Statistical Modelling (M2)		
2.1 Regression Methods	4	
2.2 Generalized Additive Models	4	
2.3 Bayesian Statistics	4	
Total number of ECTS	12	Written examination
3 Machine Learning (M3)		
3.1 Supervised Learning	4	
3.2 Beyond Supervised Learning	4	
Total number of ECTS	8	Written examination
4 Practical Applications (M4)		
4.1 Data Science in Practice	4	
4.2 Project Work	8	
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 Scientist’s Toolbox	
Prior knowledge	Knowledge of mathematical principles including basic knowledge of probability theory and programming is needed.
Format	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	<p>The module is divided into two courses: An optional “Introduction into R” and “Basics of Data Science with R”.</p> <p>1. Course: “Introduction into R” General introduction to the R programming language</p> <ul style="list-style-type: none"> • Object and data types • Base R functions <p>2. Course: “Basics of Data Science with R” Working with data:</p> <ul style="list-style-type: none"> • Importing data from various sources (SAS, SPSS) (‘haven’ package) • Visualizing data using a ‘Grammar of Graphics’ (‘ggplot2’ package) • Transforming data (‘dplyr’ package) • Tidy data • Workflow advice and functional programming (‘purrr’ package) <p>Reproducibility:</p> <ul style="list-style-type: none"> • Why reproducible research is essential to good scientific practice • RMarkdown/Quarto and knitr for automatic report generation • Package dependencies / CRAN • Version control with git
Intended Learning Outcomes	<p>The participant will be able to:</p> <ul style="list-style-type: none"> • differentiate between medical data science and “classical” biostatistics and recognize potential fields of application. • program in the statistical programming language R. • import data from a wide variety of sources in the R environment. • explain the basic structure of relational databases and perform common SQL queries such as merges or select operations in R.

	<ul style="list-style-type: none"> • visualize data using a systematic, grammar-based approach. • pre-process data to obtain a dataset that can be used for machine learning tasks. • explain why reproducibility is important in medical data science. • generate reproducible research reports using RMarkdown and Quarto. • use version control software such as Git.
Workload	Total effort: 120h
Module examination	Graded homework: exploratory presentation of a data set of interest provided by a reproducible report

3.2 Module Statistical Modelling (M2)

Prior knowledge	Content of module M1
Format	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	<p>This module introduces regression modeling strategies and Bayesian statistics and consists of three courses. The first course covers “Regression Methods” and comprises the following topics:</p> <ul style="list-style-type: none"> • 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 <p>The second course covers “Generalized Additive Models” (which are an extension of regression methods) and comprises the following topics:</p> <ul style="list-style-type: none"> • Polynomial functions of covariates • Modeling using splines • Non-parametric modeling of covariates • Implementation in R <p>The third course covers “Bayesian Statistics” and comprises the following topics:</p>

	<ul style="list-style-type: none"> • Bayes' Theorem • Bayesian linear and non-linear regression models • Markov Chain Monte Carlo Methods and Gibbs sampling • Implementation in JAGS and R
Intended Learning Outcomes	<p>The participant will be able to:</p> <ul style="list-style-type: none"> • differentiate between various (Bayesian) regression modeling strategies. • explain the mathematical assumptions of different models. • explain which models are appropriate in which contexts. • choose a regression modeling strategy based on a description of the data and the research question. • fit the coefficients of these models using R. • interpret the output of the regression program and present the results appropriately.
Workload	Total effort: 360h
Module examination	Written exam. Duration: 3x45 min

3.3 Module Machine Learning (M3)

Prior knowledge	Content of modules M1 and M2
Format	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	<p>This module introduces machine learning and statistical pattern recognition. The module is divided into two courses: “Supervised Learning” and “Beyond Supervised Learning”. In the course “Supervised Learning”, the following topics are included:</p> <ul style="list-style-type: none"> • Regularization methods for linear regression • Model assessment and selection • Neural networks • Decision trees • Random forests • Bagging and boosting <p>Topics covered in the “Beyond Supervised Learning” course:</p> <ul style="list-style-type: none"> • Concept of unsupervised learning, based on applications such as clustering and principal component analysis (PCA) • methods to capture feature importances including local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) • uncertainty quantification using conformal prediction
Intended Learning Outcomes	<p>The participant will be able to:</p> <ul style="list-style-type: none"> • explain the basic principle of each of the covered methods. • understand unsupervised learning approaches and distinguish them from supervised learning. • decide which of the methods is appropriate in which situation. • interpret statistical models using methods from explainable machine learning. • apply the methods to data using R. • interpret the output of the programs that implement these methods.
Workload	Total effort: 240h
Module examination	Written examination, duration: 2x45 min.

3.4 Practical Applications (M4)

Title of the module: Practical Applications	
Prior knowledge	Content of modules M1, M2, and M3
Format	Teaching forms are mainly group work, presentations, and discussions in the plenary. Additionally, independent working phases supervised by responsible persons are included.
Topics	<p>This module consists of two parts: 1. "Data Science in Practice" and 2. "Project Work".</p> <p>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.</p> <p>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.</p>
Intended Learning Outcomes	<p>The participant will be able to:</p> <ul style="list-style-type: none"> • collaborate in a small group and coordinate an analysis strategy for a large dataset. • apply statistical and graphical methods to analyze a dataset. • present and discuss their results. • write a statistical analysis report following best practices for academic writing.
Workload	Total effort: 360h
Module examination	Presentation of results (results of group work as part of practical experience in "Data Science in Practice", not graded) and project work (grade consists of 70% for written project work and 30% for presentation of project work)

4 Appendix

In the following, we show some examples of timetables.

4.1 Schedule of the course “Basics of Data Science with R”

TIME	THURSDAY	TIME	FRIDAY	TIME	SATURDAY
9.00 – 10.30	Introduction into data science and the study program	9.00 – 10.30	Data Visualization	9.00 – 10.30	Version Control with Git
10.30 – 11.00	COFFEE BREAK	10.30 – 11:00	COFFEE BREAK	10.30 – 11:00	COFFEE BREAK
11.00 – 12.30	Statistical Learning: the two cultures	11.00 – 12.30	Data Visualization	11.00 – 12.30	Version Control with Git
12.30 – 13.30	LUNCH BREAK	12.30 – 13.30	LUNCH BREAK		
13.30 – 15.00	Data Manipulation	13.30 – 15.00	Reproducible Reports		
15.00 – 15.30	COFFEE BREAK	15.00 – 15.30	COFFEE BREAK		
15.30 – 17.00	Data Manipulation	15.30 – 17.00	Reproducible Reports		