Statistical learning for complex data to enable precision medicine strategies

Zwep, L.B.

Chapter 1

General Introduction and Outline

1.1 Introduction

This thesis addresses the use of advanced statistical learning techniques to characterize complex and large scale data in biomedical and pharmacological research to enable development of precision medicine strategies.

1.1.1 Treatment variability

Addressing individual variability in patients' responses to drug treatment is of crucial importance to provide adequate treatment for each patient. To optimize treatment outcomes, individualization of treatment strategies is needed. This is often referred to as precision medicine, an approach based on individuals, rather than on average population effects. Lack of individualization can lead to insufficient treatment efficacy if underdosed, and adverse drug effects if overdosed.

Variation in treatment response can be due to between-patient variability and within-patient variability. Between-patient variability occurs when patients react differently to the same treatment. By quantifying this variability, patient- or disease-specific predictors can be identified to inform design of precision treatment strategies. Within-patient variability concerns the change of treatment response within a patient, due to changes in progression or adaptation of the disease during treatment. Understanding and explaining these types of variability can improve early medical decision making, through monitoring of patient response biomarkers.

The explanation of treatment response variation in drug research is part of the discipline of (clinical) pharmacology. At its core, pharmacological research concerns characterization of the dynamics of drug exposure in the body, or pharmacokinetics (PK), and the dynamics of corresponding drug effects as measured through biomarkers, or pharmacodynamics (PD). The PK describes the way the drug is moving through and changing in the body, where aspects such as absorption, concentration at the target site and clearance of a drug play a central role. The PD is the effect a drug has on a patient, mostly clinical outcomes such as blood pressure for hypertension drugs or tumor size for anti-cancer drugs. Variability in treatment response between patients can often be attributed to patient-specific factors which affect PK or PD relationships. Quantitatively capturing PK-PD relationships and identifying factors associated with inter-individual variability through mathematical and statistical models, commonly referred to as pharmacometrics, has developed as an important tool to aid in the design of individualized treatment strategies.

An important topic in pharmacology is the occurrence and development of treatment resistance. Treatment resistance can occur when a population of targeted cells or pathogens adapts to the administered drug treatment. Many different mechanisms of treatment resistance can arise due to evolutionary processes, selection pressure and rapid cell division (zur Wiesch et al., 2011). Similar treatment resistance mechanisms are observed in both oncology and infectious diseases (Groenendijk & Bernards, 2014; zur Wiesch et al., 2011).

In oncology, treatment response variability is an important factor in treatment

failure. Within patients, tumors are shown to develop resistance, which contributes to high treatment failure (Sun & Hu, 2018). Understanding of resistance development has improved cancer treatment, but there is still a lot of unexplained variability between patients (Sun & Hu, 2018; Yin et al., 2019). Understanding underlying factors for resistance development and predictive biomarkers for treatment response could improve cancer treatment outcomes.

In infectious diseases, resistance to antimicrobial drugs represent a global health challenge (Talebi Bezmin Abadi et al., 2019; World Health Organization, 2014). Pathogens can develop resistance against multiple antimicrobial treatments, turning simple infections into serious health threats. Alternative treatment strategies could help prevent the development of resistance and even reduce resistance (Maltas & Wood, 2019). One such strategy is the use of collateral sensitivity, a phenomenon where resistance to one antibiotic reduces the resistance to a second antibiotic. Collateral sensitivity is one strategy which is of interest to design treatment strategies which suppress the risk of resistance (Aulin et al., 2021; Pál et al., 2015; Roemhild & Andersson, 2021).

1.1.2 Patient- and disease associated factors

Knowledge of underlying factors of treatment response variability, such as patient and disease characteristics, is pivotal to develop strategies which can improve treatment outcomes. Insight into the factors contributing to variation in treatment response can help to predict the treatment response in different patients, enabling precision treatment strategies. Data to support deriving such insights are increasingly available from clinical studies and from routine patient care (Morrato et al., 2007).

The variability in treatment response and PK and PD of drugs between patients is large. Different patient covariates can explain parts of this variance; these covariates are for example age and body weight, but also include measurable biological factors, known as biomarkers, which are concentrations of molecules, or other physiological measures that can indicate underlying biological processes at a molecular or cellular level (Depledge et al., 1993; Strimbu & Tavel, 2010). Most pharmacological studies characterize time-dependent trends in the patients with regards to drug concentrations, treatment response and biomarker levels. These trends can be determined by measuring biomarkers reflecting different aspects of the patient's physiological characteristics. Next to the dependence structures introduced by longitudinal measurements, most covariates are also interdependent. These covariates vary with each other, often due to physical properties or biological processes, such as height being related to weight physically.

Molecular profiling 'omics' technologies for characterization of DNA, RNA, proteins, and metabolites are increasingly used to characterize biological samples from patients and during drug research (Nice, 2018). Omics data are often high-dimensional, having more variables than patients ($p \gg n$), due to the possibility to measure hundreds (metabolomics), thousands (transcriptomics) or even millions (genomics) of variables. These large sets of omics data allow for thorough characterization of pa-

tients, but also pose a challenge in terms of data analysis, due to this high-dimensional nature, and that they can be measured over time.

The increasing use of electronic health record databases has provided new opportunities for using routine health care data collected from patients in scientific research (Currie & MacDonald, 2000; Swift et al., 2018). These real world data are used to monitor patients and their treatment response in the clinic, and to make decisions about treatments and dosing schedules. Improved data availability, due to developments in data management and sharing, creates opportunities for studying patient characteristics that can predict treatment response, enabling more individualized dosing regimens in the clinic.

Overall, the complexity of these pharmacological, molecular and health care data requires the use of appropriate statistical techniques that are able to address important biomedical questions require appropriate handling of the associated heterogeneous, high dimensional, and longitudinal data.

1.1.3 Statistical methods and pharmacometrics

Complex data, such as longitudinal and high-dimensional data, require different data analysis methods. Several methods have been developed in the fields of pharmacometrics and statistics with the aim to detect covariates and biomarkers that can explain the treatment response variability and estimating their effect size.

Longitudinal data allow for studying treatment responses over time, but pose a challenge for data analysis. Measurements within a patient often are typically more similar than measurements between patients, violating the assumption of independent residuals, which is assumed in standard regression models. Mixed effect models have been developed to include the dependency structure between different measurements, enabling the characterization of the inter patient variability (McCulloch & Searle, 2000). With patient characteristics and biomarkers, part of this inter-patient variability can be explained in order to better predict outcomes for specific patients.

Pharmacometrics concerns the modeling and prediction of PK and PD measures using longitudinal data analysis methods. Through, mostly nonlinear, mixed effect modeling, random effects are estimated which represent the individual variability, thereby quantifying how diverse the response to certain drugs is over different patients, and predicting the drug effects in the population. Pharmacometric models can then be used for simulations to predict treatment responses and variability in different patient populations. These simulations take into account the unexplained between-patient variability, as well as covariates used to explain part of the difference between patients (Mould & Upton, 2012, 2013; Upton & Mould, 2014).

Next to variability between measurements, modeling interdependence between covariates also poses a challenge. Pharmacometric models often include covariates that are interdependent. To simulate different (special) patient populations, simulation of realistic sets of patient covariates is crucial, but this requires an accurate estimation of the dependence between covariates (Smania & Jonsson, 2021).

A third data analysis challenge is posed by high-dimensional data, such as most

omics data, where standard linear regression and more complex nonlinear mixed effect models are not applicable anymore, because the parameters of a model cannot be uniquely estimated (Johnstone & Titterington, 2009). One way of circumventing this problem is to use shrinkage, where a penalty is placed on the size of the parameters (e.g., regression weights), which is a technique developed within the field of statistics. The two most common shrinkage methods for linear regression are Ridge regression (Hoerl & Kennard, 1970), which penalizes the sum of the squares of the parameter values, effectively shrinking large parameter values more, and the lasso (Tibshirani, 1996), which penalizes the sum of the absolute parameter values, which shrinks some parameters to zero. So the lasso selects the most relevant parameters, which are estimated to be non-zero. In both cases, a shrinkage parameter is used to determine how strong the penalty is. Another way to analyze high-dimensional data is by using dimension reduction techniques, such as principal component analysis and proximity mapping, where variability in the high-dimensional data is summarized into much less dimensions (Heiser et al., 2020).

Although methods for high-dimensional data, longitudinal data and other complex data have been extensively developed and used, combining different data analysis methods to study treatment response variability still remains a challenge. The combination of pharmacometric approaches and statistical methods, and the application of different statistical methods in pharmacological research, can potentially improve our understanding of treatment variability and allow for the optimization of treatment and dosing regimens for individual patients.

1.2 Scope

In this thesis, we studied the use of advanced statistical techniques for the analysis of biomedical datasets to enable development precision medicine strategies, with a particular focus on pharmacological applications. With an increase in data complexity, techniques from different disciplines need to be integrated to answer research questions regarding precision treatment and antibiotic resistance. This thesis first describes this increasing data complexity and different data science techniques in more detail (Section I). Next, the thesis aims to integrate statistical techniques for analyzing high-dimensional data and pharmacometric methods to facilitate omics biomarker research (Section II) and, finally, different statistical methods are used to build tools for the pharmacological studies in clinical pathogens and populations (Section III). Thus, the thesis contains the following sections.

Section I: Data science in pharmaceutical research

In Chapter 2, we discuss the use of different data types to enhance clinical pharmacological research. These complex data require the use of different data analysis techniques and could provide insights that are hard to obtain from randomized clinical controlled trials.

Section II: High-dimensional biomarker discovery

Section II focusses on detection of biomarkers in high-dimensional omics data, using methods from statistics and data science. First, in Chapter 3, we use the lasso in combination with a pharmacometric model for tumor growth dynamics to identify potential biomarkers for treatment response and resistance development. In Chapter 4, we focus on biomarker detection to monitor the clinical course of bacterial infections in patients with community acquired pneumonia (CAP), for early decision making concerning monitoring disease progression. To detect possible biomarkers for disease progression and treatment response, longitudinal, high-dimensional metabolomics data are analyzed with dimension reduction through PCA, to explore different biochemical metabolic classes and their roles in the changes over time.

Section III: Real world data

In Section III, we develop tools to study antibiotic resistance and patient characteristics in clinical routine health care data, to support translation of concepts studied in vitro and in silico to be researched in clinical pathogen and patient data. Chapter 5 describes a method for detection of collateral sensitivity in large clinical data on antibiotic susceptibility. Using this method, Chapter 6 explores collateral sensitivity in different bacterial species and over different antibiotic classes. In Chapter 7, the statistical concept of copulas is used as a method for simulation of virtual patients for pharmacometric research. Copulas are multivariate density functions that can be used to estimate the joint density of multiple variables. We evaluate its use for estimation joint densities and subsequent simulation of patient's covariates used in pharmacometric models.

Section IV: General discussion and summary

In Chapter 8 we discuss the findings in this thesis and the future prospects for the use of statistics and pharmacometrics to evaluate treatment response variability. We discuss overall themes that are of relevance for successful research in precision medicine.

References

Aulin, L. B. S., Liakopoulos, A., van der Graaf, P. H., Rozen, D. E., & van Hasselt, J. G. C. (2021, Sep). Design principles of collateral sensitivity-based dosing strategies. *Nature Communication*s, *12*(1), 5691. Retrieved from http://dx.doi.org/10.1038/s41467-021-25927-3https://www.nature.com/articles/s41467-021-25927-3 doi: 10.1038/s41467-021-25927-3

Currie, C. J., & MacDonald, T. M. (2000). Use of routine healthcare data in safe and cost-effective drug use. *Drug Safety*, *22*(2), 97–102. Retrieved from https://doi.org/10.2165%2F00002018-200022020-00002 doi: 10.2165/00002018-200022020-00002

Depledge, M. H., Amaral-Mendes, J. J., Daniel, B., Halbrook, R. S., Kloepper-Sams, P., Moore, M. N., & Peakall, D. B. (1993). The conceptual basis of the biomarker approach. In *Biomarker*s (pp. 15–29). Springer Berlin Heidelberg. Retrieved from https://doi.org/10.1007%2F978-3-642-84631-1_2 doi: 10.1007/978-3-642-84631-1_2

Groenendijk, F. H., & Bernards, R. (2014, May). Drug resistance to targeted therapies: Déjà vu all over again. *Molecular Oncology*, *8*(6), 1067–1083. Retrieved from https://doi.org/10.1016%2Fj.molonc.2014.05.004 doi: 10.1016/j.molonc.2014.05.004

Heiser, W. J., Busing, F. M. T. A., & Meulman, J. J. (2020). Mapping Networks and Trees with Multidimensional Scaling of Proximities. In (pp. 385–407). Retrieved from http://link.springer.com/10.1007/978-981-15-2700-5_24 doi: 10.1007/978-981-15-2700-5_24

Hoerl, A. E., & Kennard, R. W. (1970, Feb). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. Retrieved from https://doi.org/10.1080%2F00401706.1970.10488634 doi: 10.1080/00401706.1970.10488634

Johnstone, I. M., & Titterington, D. M. (2009, Nov). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *367*(1906), 4237–4253. Retrieved from https://doi.org/10.1098%2Frsta.2009.0159 doi: 10.1098/rsta.2009.0159

Maltas, J., & Wood, K. B. (2019, Oct). Pervasive and diverse collateral sensitivity profiles inform optimal strategies to limit antibiotic resistance. *PLOS Biology*, *17*(10), e3000515. Retrieved from https://doi.org/10.1371%2Fjournal.pbio.3000515 doi: 10.1371/journal.pbio.3000515

McCulloch, C. E., & Searle, S. R. (2000). *Generalized, linear, and mixed models*. Wiley. Retrieved from https://doi.org/10.1002%2F0471722073 doi: 10.1002/0471722073

Morrato, E. H., Elias, M., & Gericke, C. A. (2007, Dec). Using population-based routine data for evidence-based health policy decisions: lessons from three examples of setting and evaluating national health policy in australia, the UK and the USA. *Journal of Public Health*, *29*(4), 463–471. Retrieved from https://doi.org/10.1093%2Fpubmed%2Ffdm065 doi: 10.1093/pubmed/fdm065

Mould, D., & Upton, R. (2012, Sep). Basic concepts in population modeling, simulation, and model-based drug development. *CPT: Pharmacometrics & Systems Pharmacology*, *1*(9), 6. Retrieved from https://doi.org/10.1038%2Fpsp.2012.4 doi: 10.1038/psp.2012.4

Mould, D., & Upton, R. (2013, Apr). Basic concepts in population modeling, simulation, and model-based drug development-part 2: Introduction to pharmacokinetic modeling methods. *CPT: Pharmacometrics & Systems Pharmacology*, *2*(4), 38. Retrieved from https://doi.org/10.1038%2Fpsp.2013.14 doi: 10.1038/psp.2013.14

Nice, E. C. (2018, Jul). Challenges for omics technologies in the implementation of personalized medicine. *Expert Review of Precision Medicine and Drug Development*, *3*(4), 229–231. Retrieved from https://doi.org/10.1080%2F23808993.2018.1505429 doi: 10.1080/23808993.2018.1505429

Pál, C., Papp, B., & Lázár, V. (2015). Collateral sensitivity of antibiotic-resistant microbes. *Trends in Microbiology*, *23*(7), 401–407. doi: 10.1016/j.tim.2015.02.009

Roemhild, R., & Andersson, D. I. (2021, Jan). Mechanisms and therapeutic potential of collateral sensitivity to antibiotics. *PLOS Pathogens*, *17*(1), e1009172. Retrieved from https://doi.org/10.1371/journal.ppat.1009172.g001https://dx.plos.org/10.1371/journal.ppat.1009172 doi: 10.1371/journal.ppat.1009172

Smania, G., & Jonsson, E. N. (2021, Apr). Conditional distribution modeling as an alternative method for covariates simulation: Comparison with joint multivariate normal and bootstrap techniques. *CPT: Pharmacometrics & Systems Pharmacology*, *10*(4), 330–339. Retrieved from https://doi.org/10.1002%2Fpsp4.12613 doi: 10.1002/psp4.12613

Strimbu, K., & Tavel, J. A. (2010, Nov). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463–466. Retrieved from http://journals.lww.com/01222929-201011000-00003 doi: 10.1097/COH.0b013e32833ed177

Sun, X., & Hu, B. (2018, Nov). Mathematical modeling and computational prediction of cancer drug resistance. *Briefings in Bioinformatics*, *19*(6), 1382–1399. Retrieved from https://academic.oup.com/bib/article/19/6/1382/3886023 doi: 10.1093/bib/bbx065

Swift, B., Jain, L., White, C., Chandrasekaran, V., Bhandari, A., Hughes, D. A., & Jadhav, P. R. (2018, Sep). Innovation at the Intersection of Clinical Trials and Real-World Data Science to Advance Patient Care. *Clinical and Translational Science*, *11*(5), 450–460. Retrieved from http://doi.wiley.com/10.1111/cts.12559https://onlinelibrary.wiley.com/doi/10.1111/cts.12559 doi: 10.1111/cts.12559

Talebi Bezmin Abadi, A., Rizvanov, A. A., Haertlé, T., & Blatt, N. L. (2019, Dec). World Health Organization Report: Current Crisis of Antibiotic Resistance. *BioNanoScience*, *9*(4), 778–788. Retrieved from http://link.springer.com/10.1007/s12668-019-00658-4 doi: 10.1007/s12668-019-00658-4

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Royal Statistical Society*, *58*(1), 267–288. Retrieved from www.jstor.org/stable/2346178

Upton, R. N., & Mould, D. R. (2014). Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development: Part 3—Introduction to Pharmacodynamic Modeling Methods. *CPT: Pharma-*

cometrics & Systems Pharmacology, *3*(1), e88. Retrieved from http://doi.wiley.com/10.1038/psp.2013.71 doi: 10.1038/psp.2013.71

World Health Organization. (2014). *Antimicrobial resistance: global report on surveillance* (Tech. Rep.). Retrieved from https://apps.who.int/iris/bitstream/handle/10665/193736/9789241509763_eng.pdf

Yin, A., Moes, D. J. A., Hasselt, J. G., Swen, J. J., & Guchelaar, H. (2019, Oct). A Review of Mathematical Models for Tumor Dynamics and Treatment Resistance Evolution of Solid Tumors. *CPT: Pharmacometrics & Systems Pharmacology*, *8*(10), 720–737. Retrieved from https://onlinelibrary.wiley.com/doi/10.1002/psp4.12450 doi: 10.1002/psp4.12450

zur Wiesch, P. A., Kouyos, R., Engelstädter, J., Regoes, R. R., & Bonhoeffer, S. (2011, Mar). Population biological principles of drug-resistance evolution in infectious diseases. *The Lancet Infectious Diseases*, *11*(3), 236–247. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S1473309910702644 doi: 10.1016/S1473-3099(10)70264-4