DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
11

TÕNIS TASA

Bioinformatics Approaches in Personalised Pharmacotherapy

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

11

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
11

TÕNIS TASA

Bioinformatics Approaches in Personalised Pharmacotherapy

UNIVERSITY OF TARTU
Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in informatics on June 21th, 2019 by the Council of the Institute of Computer Science, University of Tartu.

Supervisors

Prof. Jaak Vilo
Institute of Computer Science, University of Tartu, Tartu, Estonia

Prof. Lili Milani
Insitute of Genomics, University of Tartu, Tartu, Estonia

Tuuli Metsvaht, MD, PhD
Department of Pediatrics, University of Tartu, Tartu, Estonia

Opponents

Prof. Inge Jonassen
Department of Informatics, University of Bergen, Bergen, Norway

Prof. William Hope
Department of Molecular and Clinical Pharmacology, University of Liverpool, Liverpool, UK

The public defense will take place on August 26th, 2019 at 14:15 in J. Liivi 2, room 405.

This publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906
ISBN 978-9949-03-094-1 (print)
ISBN 978-9949-03-095-8 (PDF)

University of Tartu Press
`http://www.tyk.ee/`

Pühendatud mu kõigile eelkäijatele, perele ja järeltulijatele.

ABSTRACT

The amount of collected health data is growing fast. Insights from these data allow using biological patient specifics to improve therapy management with further individualisation. This has spurred personalised medicine that has come to represent all therapeutic developments based on genomic and other types of individualised biological data. Eventual goals require development of applicable methods and tools, interpretable analysis results and usable interfaces of communication. This thesis addresses problems in multiple sub-fields of personalised medicine.

Therapeutic drug monitoring relies on drug administration adjustments during treatment. However, drug metabolism is difficult to predict because individual biological differences cause variability in patient responses. Fortunately, drug concentrations collected in-treatment can often be associated with outcomes and can therefore guide personalised dosing decisions. To address the growing need for such tools, we have developed and externally evaluated a precision dosing tool that allows individualised dosing of vancomycin in neonates.

Genetics is also a rich source of information for treatment individualisations. Other than drugs used in therapeutic drug monitoring, most pharmacotherapies can not rely on continuous input data because self-medication complicates measurements of outcomes. In many of these cases, pre-emptive use of genetic information can help avert unwanted outcomes. Effects of many genetic variants are often large enough to warrant changes in drug prescriptions or dosing schedules. On-going initiatives in the field aim to identify, validate and implement tests for genetic variants that manifest drug effects. We have applied an hypothesis-free population-based approach in testing drug related adverse effects to genomic loci, and found and validated a novel variant in CTNNA3 gene that increases the prevalence of adverse drug effects in patients with oxicam prescriptions. This was made possible by the nation-wide genomic data collection initiative coordinated by the Estonia Genome Center.

Computational genetics relies on quantitative methods. The most common method to study relations between genomic markers and individual traits is the genome-wide association analysis (GWAS). Downstream analysis of association results often applies time-consuming custom approaches. A common step involves visual assessment of the distribution of genetic variants and GWAS p-values. Our previous study on pharmacogenetic variants led to acknowledging the need for a more automated detection of "interesting" visual peaks worth further assessment and to the development of Manhattan Harvester. These tools automate the detection and quality scoring of genomic regions based on GWAS summary statistics that considerably decreases the time an analyst spends on individual plots. The quality scores were designed to emulate the subjective assessment by human evaluators.

CONTENTS

LIST OF PUBLICATIONS

List of publications included in this thesis are referred to by Roman numerals (Ref I - IV).

INCLUDED PUBLICATIONS

I Tõnis Tasa, Tuuli Metsvaht, Riste Kalamees, Jaak Vilo, and Irja Lutsar. Dosopt: a tool for personalized Bayesian dose adjustment of vancomycin in neonates. *Therapeutic Drug Monitoring*, 39(6):604–613, 2017

II Tõnis Tasa, Riste Kalamees, Jaak Vilo, Irja Lutsar, and Tuuli Metsvaht. External evaluation of population pharmacokinetic models for vancomycin in neonates. *bioRxiv*, page 458125, 2018. *Manuscript is available in bioRxiv*

III Tõnis Tasa, Kristi Krebs, Mart Kals, Reedik Mägi, Volker M Lauschke, Toomas Haller, Tarmo Puurand, Maido Remm, Tõnu Esko, Andres Metspalu, et al. Genetic variation in the Estonian population: Pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, 27(3):442, 2019

IV Toomas Haller, Tõnis Tasa, and Andres Metspalu. Manhattan Harvester and Cropper: a system for GWAS peak detection. *BMC Bioinformatics*, 20(1):22, 2019

PUBLICATIONS NOT INCLUDED IN THIS THESIS

Other publications with my contributions that have not been included in this thesis.

V Mithu Guha, Mario Saare, Julia Maslovskaja, Kai Kisand, Ingrid Liiv, Uku Haljasorg, Tõnis Tasa, Andres Metspalu, Lili Milani, and Pärt Peterson. DNA breaks and chromatin structural changes enhance the transcription of autoimmune regulator target genes. *The Journal of Biological Chemistry*, 292(16):6–542, 2017

VI Maarja Hallik, Mari-Liis Ilmoja, Tõnis Tasa, Joseph F Standing, Kalev Takkis, Ruta Veigure, Karin Kipper, Tiiu Jalas, Maila Raidmäe, Karin Uibo, et al. Population pharmacokinetics and dosing of milrinone after patent ductus arteriosus ligation in preterm infants. *Pediatric Critical Care Medicine*, 2019

VII Maarja Hallik, Tõnis Tasa, Joel Starkopf, and Tuuli Metsvaht. Dosing of milrinone in preterm neonates to prevent postligation cardiac syndrome: Simulation study suggests need for bolus infusion. *Neonatology*, 111(1):8–11, 2017

VIII Helgi Padari, Kersti Oselin, Tõnis Tasa, Tuuli Metsvaht, Krista Lõivukene, and Irja Lutsar. Coagulase negative staphylococcal sepsis in neonates: do we need to adapt vancomycin dose or target? *BMC Pediatrics*, 16(1):206, 2016

IX Helgi Padari, Tuuli Metsvaht, Eva Germovsek, Charlotte I Barker, Karin Kipper, Koit Herodes, Joseph F Standing, Kersti Oselin, Tõnis Tasa, Hiie Soeorg, et al. Pharmacokinetics of penicillin G in preterm and term neonates. *Antimicrobial Agents and Chemotherapy*, pages AAC–02238, 2018

X Kadri Rekker, Tõnis Tasa, Merli Saare, Külli Samuel, Ülle Kadastik, Helle Karro, Martin Götte, Andres Salumets, and Maire Peters. Differentially-expressed miRNAs in ectopic stromal cells contribute to endometriosis development: The plausible role of miR-139-5p and miR-375. *International Journal of Molecular Sciences*, 19(12), 2018

XI Kadri Rekker, Merli Saare, Elo Eriste, Tõnis Tasa, Anne Mari Roost, Viktorija Kukuškina, Kristi Anderson, Külli Samuel, Helle Karro, Andres Salumets, et al. High-throughput mRNA sequencing of stromal cells from endometriomas and endometrium. *Reproduction*, 154(1):93–100, 2017

XII Kadri Tamme, Kersti Oselin, Karin Kipper, Tõnis Tasa, Tuuli Metsvaht, Juri Karjagin, Koit Herodes, Helmut Kern, and Joel Starkopf. Pharmacokinetics and pharmacodynamics of piperacillin/ tazobactam during high volume haemodiafiltration in patients with septic shock. *Acta Anaesthesiologica Scandinavica*, 2(60):230–240, 2016

INTRODUCTION

Conventional therapeutic medicine is built on systematic classification of pathologies that are followed up with chemical, surgical or radiation-based interventions. Clinicians currently use basic patient characteristics such as age and sex in intervention decisions but not much regard is still given to other individual patient attributes due to lack of knowledge about patient biology and availability of methods to include detailed patient information. However, all intervention methods are subject to large between-subject differences and thus a great promise lies in fine-tuning the use of existing therapeutic methods by patient specifics.

There is an on-going data revolution in health related fields [135]. Increases in computational capacities and storage make wide-spread data collection feasible for a growing proportion of stakeholders and rapid advances in genomic sequencing technologies have massively increased the amount of available biological information [5]. Genomic biobanks contain biological samples that provide the foundation for subsequent information retrieval. At the same time, there is a slow but steady movement towards common data sharing practices [158]. Requirements for data availability statements and deposition of intermediate data in publicly available repositories are becoming more common for scientific journals [90]. These provide a foundation for research and development in academia and industry. The genomic data is also supplemented by increasingly effective measurement and collection of observable patient traits and other data about the treatment management into electronic health data repositories [153]. Heterogeneous sources and types of data are expected to help gain new insights and applications in the real-world.

This thesis explores several aspects that these new previously unavailable data sources provide towards improving the outcomes of currently existing treatments. Broadly, such personalised, individualised or genomic medicine approaches require quantitative methods to analyse and apply genomic and other types of biological patient data to make the patient more likely to be responsive to treatment.

Personalised medicine is an umbrella-term often used to relate to therapeutic provision developments based on genetic data. However, there are increasing attempts to attain similar targets of increased personalised care using sources of other individualised health data. Therapeutic drug monitoring (TDM) is a branch of clinical pharmacology that has traditionally been used to adapt therapeutic drug doses [88]. Its practitioners are increasingly applying data-rich quantitative methods to guide on-going therapy in response to observed events. One objective is to use these methods to guide patients' drug concentration profiles towards therapeutic windows that increase the receptivity of a drug. This aim can be achieved by adjusting doses based on individually measured drug concentrations and other measurable variables. This thesis presents two articles that demonstrate the workings of a web-based tool that simplifies individualised dosing of vancomycin in neonates.

- **Ref I** - "Dosopt: A Tool for Personalized Bayesian Dose Adjustment of Vancomycin in Neonates". This novel web-based TDM tool, DosOpt, allows optimisation of in-treatment vancomycin doses in neonates, and is the first freely available tool focused on this cohort. The adjustments in DosOpt are based on combing population information with patient concentration measurements.
- **Ref II** - "External Evaluation of Population Pharmacokinetic Models for Vancomycin in Neonates". To further elucidate the pharmacokinetics of vancomycin in neonates, we also aimed to benchmark academically published population models that are used as the basis for DosOpt TDM adjustments.

In real life, biological concentration measurements are mostly not available in self-managed drug treatments but largely immutable biological data such as DNA can still be used for treatment individualisation. Pharmacogenetics aims to explain drug response variability in relation to genetic differences. Identification of patient biomarkers that associate with changes in drug response help lowering rates of unexpected side effects and hospitalisations [48]. The effect sizes of genomic loci are often large enough to warrant changes in treatment guidances [118]. In this thesis we have used population-based data to find genomic loci that are associated with specific drug usage related adverse effects.

- **Ref III** - "Genetic Variation in the Estonian Population: Pharmacogenomics Study of Adverse Drug Effects Using Electronic Health Records". Our contribution is to identify novel and confirm previously known genetic markers linked to drug related adverse effects. A distinctive feature of this work is the use of population-based data in the study of pharmacogenetic relationships.

Next, we cover the development of a tool that enables automated detection of interesting genetic signals from results of associated genomic markers and traits of interest. The need for such a tool was acknowledged during a previous study that analysed a large number of genetic regions. The "interestingness" of regions that are selected for further assessment is in large part subjective and study-dependent. In our work, we aimed to help reduce the time needed for selecting genetic validation targets by emulating human quality assessments of genomic association results.

- **Ref IV** - "Manhattan Harvester and Cropper: a System for GWAS Peak Detection". Assessment of Manhattan plots is a common step in the genome wide association analysis pipeline. We developed a software that enables automatic assessment and evaluation of regional genome plots of genome wide association study (GWAS) summary data.

Overall, this thesis aims to contribute to the tool-set and knowledge applied within the framework of personalised medicine as opposed to population-based pharmacotherapy. The main aims of this thesis are:

1. To develop and qualify a novel TDM personalised dose optimisation tool using vancomycin in neonates as an example;
2. To evaluate the effect of using different published pharmacogenetic models as Bayesian population priors within the developed tool;
3. To identify novel and confirm previously known genetic associations to drug related adverse effects using data from electronic health records;
4. To simplify the detection of potential signals in genome-wide association analyses.

The Preliminaries chapter of this thesis provides the background by introducing concepts of precision medicine, electronic health data, genetics and therapeutic drug monitoring. Subsections of the preliminaries focus on specifics that lead to thesis results. The two covered personalised medicine subsections are pharmacogenetics and therapeutic drug monitoring. Concepts in genetics relate to pharmacogenetics and its clinical applications. One of the main aims of pharmacogenetic research is to reduce adverse drug effects. Gene-drug associations that increase target attainment are morphed into therapeutic guidelines and subsequently applied in clinical practice. We outline the state of the field and discuss methods commonly applied in association discovery pipelines. Relevant concepts in TDM provide context for methods and prerequisites of developing quantitative TDM strategies. The other chapters summarize the author contributions from publications included in this thesis (Ref I-IV).

1. PRELIMINARIES

1.1. Precision medicine

Increasingly unsustainable costs of health-care systems in the developed world have forced administrators to re-evaluate their approaches to providing healthcare [144]. Pharmaceutical industry struggles with lagging research productivity [183]. The precision medicine approach has held promise for major improvements through more effective resource management and effective use of patient data in re-imagining therapeutic medicine. Although precision/ personalised/ individualised medicine has been most pre-eminently associated with advances in genomics, a wider scope includes all health-related data inputs [41].

Currently, pharmacotheraphy is broadly population-based. Instead, individualised approaches use methods that customise treatment administration and management according to the past and present specifics of the patient. This approach is largely data driven - anything that can be collected, counted or measured might be included in treatment decisions. The promise of precision medicine is to make health care more personalised, preventative, patient inclusive and more cost-effective. Ideally, this brings about both improvements in health-outcomes and introduces cost-savings [10].

Personalised medicine is still a developing field with many unknowns and potential pitfalls [84]. Success in reformulating existing systems depends on perceived economic and health gains. Development of digital support systems that guide medical decisions makes the system more complex. More of complexity is expensive to maintain and develop and likely exacerbates the medical outcome differences between high-and low-income countries [5]. Patients with higher genetic risk may visit physicians more because of increased worry [16]. Physicians need to develop new skills as they need to learn to use and apply the decision support systems. The workload of the medical workforce increases as they need to include analytical computerised interface in their work and in communication of inferences [12]. This means they need to understand and be convinced of the benefits [9]. The improvements are expected to come from higher efficacy of healthcare systems and increased economic output from added qualitative life-years [42]. Evidence about the meaningful public health-effect of personalised medicine tests in actual use remains inconclusive. Some sources indicate that most tests providing better health at a somewhat higher cost and only a minority present eventual cost savings [154].

Personalised medicine programs are nevertheless being adopted all over the world. In 2015, the US president proposed an investement of 215 million for a corresponding national initiative for developing and implementing customised treatments. Estonian Ministry of Social Affairs administrates the implementation of Estonian personalised medicine programme 2016-2020 following a feasibility study. Other national efforts in Europe include SPHT in Switzerland and

Aviesan national alliance in France amongst several others [109, 124, 141]. The International Consortium for Personalised medicine brings together both national entities and EU representatives to forge a common development framework in Europe [26]. Work for foundations of similarly co-ordinated initiatives are being laid in parts of Asia [146].

Personalised medicine is powered by considerable advances in emerging biological fields such as genomics and related sub-fields, and data analysis, storage and collection capacities. The outcomes can be applied in different domains. Genomic studies have uncovered thousands of trait related markers [199]. Between-subject drug effect differences can be quantitatively explained by variations in the genome, transcriptome, microbiome and environment [86, 205]. Advanced analytics methods can be applied at bed-side at monitoring and patient specific treatment adjustments [137]. Machine learning methods can help medics detect malignancies based on medical image processing and diagnostics [51].

1.2. Electronic health data

Predictive models in medicine aim to quantify medical phenomena with the use of past data to describe the future. These rely on availability and access to data. Data input is required to investigate and derive inferences for individual decision support. Therefore considerable efforts in developing precision medicine initiatives are going into systemic organisation of electronic data collection and management. Medical systems all over the world are replacing paper based health records with electronic data [149]. Benefits of computerised health records include the availability, legibility, continuity and completeness. These systems aim to collect and aggregate medical care information of all type starting from patient life choices, familial health histories, related health-affecting habits and patients own disease history including case durations, used and prescribed medications, adverse effects, co-morbidities, symptom descriptions and outcomes [69]. Linked in digital systems they also provide functions for exchange of data between clinicians and interaction with patients [126]. Overview of comprehensive medical histories is invaluable for providing the best possible care. Besides administrative effectiveness, the existing system is made more accessible and simultaneously enables the use of novel clinical data for research.

Electronic health records can substantially improve wellness and disease management by providing the basis for individual decision support, population management and analytics routines [153]. Patient portals, in-treatment analytics tools and predictive clinical models aim to transform patient data to an integrated component of on-line care. Currently, electronic health records have not yet filled their full potential even though national initiatives for adoption and system integration are on-going [14]. However, patient data management tools are moving from static dashboards and reports to real-time health assessment reports and calibrated predictive probabilistic modelling [37]. Evidently, electronic health records have

a two-fold utility for the patient. First, these can be used to develop predictive models and patient care routines, and also that individual data can be used in such models. Use of health records in clinical research for regulatory, observational and safety studies is being widely used in an expanding set of applications [30].

Collection and measurement of genomic, metabolic or microbiome data is expensive and administratively difficult [104]. Handling requires careful methods of storage and transport. Human biological samples are vulnerable to environmental degradation and require special laboratory conditions. Donor information and data usage needs to confirm to pre-set requirements and regulatory conditions. For this, population-based biobanks have emerged to store and manage the samples safe from physical harm in conditions that minimise their deterioration over time.

Use of genomic and medical data is riddled with ethical, moral and legislative questions which complicate adoption. Regulations regarding biobank activities and their data processing are still evolving and are often handled case-by-case. Regulatory constraints regarding private medical data severely complicate efforts in data sharing and aggregation. Patient consent and public approval of reasonable use policies are crucial, otherwise key stakeholders could be alienated [19, 56]. The general public regards scientific research that translates to public health improvements as one of the most desirable use cases [105]. Breach of trust may decisively turn public opinion against development of novel health approaches and thus public must be included in the decision making process and kept knowledgeable of the actions regarding their private data [156]. Precaution is warranted for potential cases of discrimination and data misuse. Large inconsistencies remain within data safety and privacy standards of healthcare providers with key concerns around data access and accountability questions [50]. Data leakage to external actors is a concern. Public upheaval can result when holders of genetic data do not understandably inform the donors of their data sharing practices with third parties [63]. Patients are not comfortable sharing all their medical health data with any other healthcare provider associated practitioner than the treating physician [158]. Laws regarding lawful use of data contents are largely not agreed upon between-systems. For example, calculation of insurance premiums based on medical data has ethical implications. Other barriers to adoption include lack of funds in the public system, and for the physicians the problems may be in missing skills and training, lack of desire for additional control measures and little belief in the utility of innovations [11]. Lack of interest in adopting new technological measures is grounded in several issues. The physicians are not always educated on the benefits and taught how to properly use the system [178]. This means that the implemented systems do not necessarily contain features required for successful use [181]. In case of compulsory, bureaucratically mandated use, the physicians start using short-cuts in the digital system by missing fields, inconsistent updates and copy-pasting previous inputs [14].

In Europe, countries in the Nordic region and UK are especially active in using

biobanks in planned precision medicine programs [142]. Specifically, we highlight the Estonian success story for its approach to data-driven healthcare. The Estonian Genome Center affiliated with University of Tartu (EGCUT) collects and manages the genomic data of volunteered adult Estonian population participants [103]. It was set-up as a longitudinal, population-based Biobank with ability to re-contact patients but currently also performs the duties of a research institute. Its aims included promotion of genetic research development, collection of health and genetics data and public health improvements [104]. Now, EGCUT also manages the population-wide analysis and communication of genetic feedback on selected diseases and drugs to the first 52,000 participants. The genomic data is supplemented by an extensive questionnaire describing the life-history of the participating individuals [103]. Current iteration has expanded the genotyped population to around 150,000 individuals. Publicly oriented Biobank that focuses on communication of insights has an important role in popularising the idea of genomic based medicine. About 70% of the population in Estonia is in favour of the work done in EGCUT [104]. This has been made possible by constant engagement with the public.

The data in EGCUT is periodically linked to other national health registries and databases including data registries of main hospitals, Estonian Cancer Registry, and Death Registry [104] that helps the Center fulfil its research goals. All research is performed following the broad consent form of University of Tartu Research Ethics Committee. Additionally, initiatives such as the Estonian e-health system support the development of comprehensive personalised medicine in a top-to-bottom fashion and is the centralised focal point for nationwide infrastructure. Estonian National Health Information System collects the health information provided by patients themselves, general practitioners or specialised physicians. Databases are connected via a governmental IT framework X-Road [197]. Standardised data presentation, clear guidelines for access, usage and review provide the basis for future work and feature additions. In all, a comprehensive personalised approach relies on a heterogeneous set of data sources and registries.

1.3. Basics of genetics

Human genetic material containing biological instructions is packed into deoxyribonucleic acid (DNA). DNA is made up of 4 nucleotides: adenine (A), thymine (T), guanine (G) and cytosine (C). All of organism's genetic material makes up its genome. DNA sequences in the genome that produce biologically functional proteins are called genes. One of the most important biological tenets and the founding informational pathway is the transcription of double-helix DNA into single stranded messenger RNA. Three bases of RNA are translated into a single amino acid. All amino acids derived from a single ribosome mediated translation event form a protein.

No two humans are genetically identical. A human offspring gets around 50%

of its genetic material from each of the two parents. To a lesser extent, novel individual mutations are introduced to heritable and non-heritable cells at a very slow rate and genetic material can be recombined between and within chromosomes during meiosis and mitosis. The resulting genetic between-subject variation ensures that the overall ability of a species to respond to changing environmental and societal changes remains flexible. This genetic shuffle is nature's method to heuristically search for combinations that improve the fitness and survival of the species. The genetic composition of an individual, made up of around 3.6 billion base-pairs, is unique. Genetics is mostly interested in characterisation, functionality and mechanisms of the structural units that vary between individuals. Genetic variation exhibits itself in very heterogeneous forms: from one-base single nucleotide variant (SNV) changes to large scale structural variations encompassing millions of base-pairs in the genome.

Biological processes are difficult to explain mechanistically. Each measurable trait or phenotype is a result of numerous complex biological interactions on molecular, cell and tissue level with added random effects from environmental variables. Most phenotypes' exact formation mechanisms and pathways are unknown. Therefore one of the most popular first approaches for finding genetic variations related to some expression of a trait is to evaluate the strength of associations using statistical methods. Statistically significant associations indicate that the variation is a plausible candidate to have a measurable effect on the outcome variable and is potentially involved in the system of trait development. A genetic perturbation can affect the trait by switching genes on/off, decrease/increase the expression of the gene, alter the protein, mediate regulatory pathways or alter interactions with epigenetic elements and environmental attributes.

Many technologies exist for interrogating the exact DNA composition of an individual. Early genotyping technologies involved a painstaking work performed with techniques such as Sanger sequencing [172]. Preparation required custom-prepared lab primers and manual labour. DNA microarrays allowed rapid targeted genotyping of prioritised gene regions and SNVs [61]. Despite that DNA microarray use requires strong foresight of regions of interest and would not enable to evaluate out-of-region associations, even now DNA microarrays are the most cost-effective option for determining the composition of specific genomic loci.

The nucleotides of the first human genome were sequenced by determining and ordering them for millions of euros by the year 2000 [28]. This mega project highlighted a need for more scalable techniques which emerged with the next generation sequencing (NGS) techniques that streamlined genotyping capacity. The range of available technologies has now greatly expanded the range of testable genomic regions. NGS techniques include a plethora of methods and technologies that allow deep characterisation of parts of the genome or its products such as DNA (DNA-Seq), RNA (RNA-Seq), protein-DNA interactions (Chip-Seq) *etc.* [125]. Whole genome sequencing captures (almost) the entirety of the organism genome, whereas whole exome sequencing encompasses a functionally very im-

portant subset of all protein coding genes. Rapid advancement of next-generation technologies has lowered the cost of whole genome sequencing to around a thousand dollars per individual. The decreasing sequencing cost has caused a proliferation of studies that study genomic effects in relation to observable traits in large groups of individuals [114]. Whole genome sequencing (WGS) of populations has also led to construction of reference panels for probabilistically imputing the untyped genotypes based on a genotyped subset of markers [25, 122, 127]. Combining reference panels with genotyping arrays provides an relatively cheap way for obtaining comprehensive individual genetic profiles. The gnomAD browser (v 2.1), which has aggregated data on 125,748 exomes and 15,708 whole-genome sequences, includes around 230 million unique variant forms, alleles [89]. This number is likely to continue growing as more individuals are sequenced.

1.3.1. Genome wide association discovery

One of the main aims in genetics is to explain observable traits through genetic variants. Direct observational studies do not scale to the human genome and to the large number of statistically significant variants with small individual effects [151]. Moving from targeted genotyping to whole-genome approaches allows usage of data-driven approaches. Quite commonly investigators apply the framework of statistical hypothesis testing on a whole genome scale, a genome wide association study (GWAS), to evaluate genomic variants for statistical correlations with the phenotype [198].

GWAS approach tests the genotype effect in relation to the dependent phenotype. Single marker association tests the alleles, realised variant forms, one at a time. Since humans have two copies of each chromosome then in cases where a SNV varies between two nucleotides then a genotype configuration at a specific loci can contain either two copies of either allele or one of each. For example, if at a certain loci C is the most commonly observed allele and also in the reference genome (reference allele) but T is also sometimes manifested (alternative allele) then the genotypes at this position can be CC, CT or TT. For quantitative analyses the alleles need to be coded so that it reflects the genotype effect change on the phenotype. Dominant model posits that a single risk allele is needed to exhibit the effect, a recessive model requires two copies of the risk allele. However, the analyses most commonly apply additive models which assume that the risk increases linearly with each risk allele so that the genotypes can be coded as 0,1 and 2 [18].

Let's consider a case for associating genotypes with a phenotype that can take continuous numeric values. One approach to test the relationship between the phenotype and genotype would be to use statistical hypothesis testing which compares an actual scenario with an expected one given a set of constraints and compares them based on a test-statistic value. The hypothesis under any null scenario is that a difference between the two scenarios does not exist. The simplest statistical test, one sample t-test, compares the mean of a sample to a constant. This test is often

the basis for testing significance of the predictors in a linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is the residual error vector and its elements are independently and identically distributed by $N(0, \sigma_\varepsilon^2)$, where σ_ε^2 is the variability of the residual error, $\mathbf{y}$ is the dependent variable, $\mathbf{X}$ the design matrix that consists of genotypes coded in accordance with the genetic model and other predictive variables and $\boldsymbol{\beta}$ a vector of coefficients that are to be estimated for predictive variables. One way of solving this would be to minimise the sum of squared residuals for which there exists a closed-form solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}.$$

Assuming a number of independent variables, the null hypothesis of the coefficient for the genotype β_1 can be tested for linear relationship with $\mathbf{y}$ using hypotheses

$$H_0 : \beta_1 = 0,$$

$$H_1 : \beta_1 \neq 0,$$

using a test-statistic that approximates the normal distribution

$$t = \frac{\hat{\beta}_1 - \mu}{se(\hat{\beta}_1)},$$

where μ evaluates to zero as specified in the hypothesis and $se(\hat{\beta}_1)$ is the standard error of the β_1. Next, the value of this test-statistic (t-statistic) translates to the probability of observing at least as extreme of a test-statistic value from an underlying distribution (t-distribution) known as the p-value. The p-values are used as the main decision criteria for either rejecting or staying with the null hypothesis of the test so to decide of the predictor is significant in terms of explaining the outcome [101].

A second very common set-up would associate genotypes to a binary predictor. For example, one analysis aim would be to relate genotypes with diseased (case) and healthy (control) statuses of patients. Such a design allows grouping the patients by their outcome (case/control) and presence of an allele in a gene. Binary outcomes can be tabulated into frequency tables for which a plethora of statistical tests such as Chi-squared or Fisher's exact test can be used for testing group differences. Fisher's exact test allows evaluation of direct probabilities from a hyper-geometric distribution. If the probability of observed data that is used as the p-value is lower than the threshold then null is rejected and differences in category frequencies are declared. Chi-squared test relies on the expectation of a test-statistic from chi-distribution. The extension of 2x2 frequency table is a logistic model

$$\text{logit}(p_i) = \log\frac{p_i}{1-p_i} = \mathbf{x}_i\boldsymbol{\beta},$$

where $\mathbf{x}_i$ is a row of design matrix $\mathbf{X}$ that contains predictors for the element i where $i = 1 \ldots n$, to predict the probability p_i that the binomially distributed outcome $y_i = 1$. Logistic regression does not have a closed form solution. Instead a maximum likelihood estimate, differentiated logarithm of the loss-function, is obtained using some iterative methods such as Newton-Rhapson or gradient descent [165]. The maximum likelihood estimates are parameter values that maximise the likelihood function. Hypothesis testing of logistic regression coefficients is commonly performed using Wald test [200]

$$W = \frac{\hat{\beta}_1 - \mu}{se(\hat{\beta}_1)}.$$

Since both β_1 and μ are assumed to come from normal distribution so the square of their differences is assumed to arise from Chi-squared distribution. Odds ratios (OR) from GWAS are used to describe the extent that a genotype affects the outcome.This measure is derived from logistic regression analysis by exponentiating the coefficients. An odds ratio derived from a coefficient of a genotype represents the odds of having a particular risk genotype in cases compared to controls [76]

$$OR_{\beta_1} = \exp(\beta_1).$$

In a single statistical test, a p-value that is often use as a threshold in hypothesis testing is 0.05. If the p-value probability is lower than the threshold then a statistically significant difference between observed conditions is declared and the null is overturned. Increasing the number of statistical tests increases the probability of observing false associations. A common approach is to perform threshold corrections corresponding to the number of performed tests. The gold standard for adjusted threshold in GWAS studies in usually set at p-value $5 * 10^{-8}$ which roughly estimates the number of independent tests to around 1,000,000 [27]. Often, additional safeguards are applied to guard against spurious results. This includes using bioinformatics methods for assessing the functionality of the region, external validations in independent datasets and evaluation of background knowledge from previous studies.

The number of significant associations identified with GWAS that relate to a multitude of traits is already in thousands and growing [114]. Ideally, data-driven GWAS insights are placed in biological context and explained as part of mechanistic pathways for the associated traits. In recent years, several calls have been made to guide research towards translational approaches and clinical applications [23, 115]. This has proven to be a challenge [45]. Most of the associated variants are in complex intergenic or intronic regions so the effect of SNVs is not directly inferrable from the effect on resulting protein [72]. Instead, these variants seem to have a role in a dynamic system between various regulatory, epigenomic, transcriptomic elements that can be tissue and development cycle specific [110, 173]. Mechanistic translational science is more expensive, time-consuming and less accessible, alas the population-based GWAS approach is still

wildly popular and even increasing in usage [55, 199]. It is unlikely that adding many more samples to new GWAS would much increase the currently explained variability of most commonly studied complex disease traits [168]. With smaller effect sizes, the statistical significance does not translate to clinical significance.

Still, the success of a data-driven and statistical GWAS approach is testified in a number of developments. It has significantly increased the amount of explained genomic variation of many phenotypes, GWAS studies have identified many SNVs that replicate well and have also yielded some SNVs with clear medical utility [198]. Genomic data used for GWAS can be combined with other sources to provide the basis for systemically disentangling the mechanisms in complex biological systems [68]. In drug effect related genomics studies, many associations between drugs and a single allele have been shown to have effect sizes that warrant changes in treatment [33]. Individual GWAS SNVs are now being used in predicting complex traits and common diseases by combining multiple allele contributions in polygenic risk scores that explain larger proportions of total variance [29]. These scores have now started to show predictive performances which may propel translation of polygenic risk scores to clinical practice [92].

1.3.2. Pharmacogenetics

Pharmacogenetic studies mostly investigate the relationship between genomic regions and responses to pharmacological medications. Genetic variation has been shown to alter drug metabolism, absorption and elimination [71]. These can result in markable differences in dosages needed to produce similar concentration-time profiles. Use of a drug with a conflicting genotype could also lead to a completely missing positive response, unexpected severe health effects, hospitalisation or even death [44]. Adapting the genetic information to personal drug selection and adjusted administration holds promise in improving therapeutic care [169].

Under-estimation of drug variability may lead to failure of therapy due to dosing failures, low efficacy or toxicity events. Prevention of adverse drug effects (ADE) is the motivating factor of studying pharmacogenetics as ADEs account for around 6.5% of hospital admissions in the western countries [155]. Preventative assessment of ADEs can decrease the expected patient treatment costs but also improve the rates of successful therapies. However, drug switches for the primary condition and medications to the resulting ADEs drives up costs [85]. In the United States, ADE management is estimated to cost up to 30 billion United States dollars annually [185].

Most drugs have listed some potential side effects but on a population level most of these have relatively small epidemiological prevalences [15]. Evidence on ADEs in treatment centres is collected through physician documentations. The physician needs to recognise the ADE and also link it to drug usage by some universal identification. Treating physicians are likely to miss related side effects that have not directly been implicated with the drug before-hand [44]. Pre-

market testing is unlikely to have identified all possible ADEs that happen in low frequencies due to limited sample sizes, limited duration of studies and unfocused cohorts [49]. Co-administration of several drugs complicates the outlook. Thus, estimates of the extent of ADEs are huge but their cost and extent are most certainly underestimated. In Kuwait, only 34% of the physicians have ever reported ADEs [4]. Hazell and Shakir reported that the median under-reporting rate from 37 studies conducted in wide range of different countries was a staggering 94% [70]. The high cost of adverse drug effects provides abundant cost-saving opportunities.

Drug metabolism pathways are relatively well-studied. A set of around 60 genes also known as Very Important Pharmacogenes (VIP) are known to be mechanistically involved in metabolism of many drugs [46]. These genes have become the first focus points when investigating genetic associations to drug response. Quite uncommonly for many traits, alterations in metabolising genes can result in drug response events with large effects which makes prevention of ADEs a clinically implementable task [99]. Firstly, limited genomic area of interest means that the regions are well studied which improves functional profiling and annotation of variants. Unacceptable ADEs can be avoided by timely drug switches to efficient and non-toxic alternatives. In other cases, knowing the effect and magnitude of the alterations enables changes in dosing which lead to more preferable drug concentration profiles leading to desirable therapeutic effects and fewer adverse effects/toxicity. Therefore, personalised genome-based approach to prevention of adverse drug effects can lower the rate of re-medications, treatment interventions and can increase the dosing accuracy as there is a better model for the individual's response [24].

Genetics is a significant contributor to the between-subject variability of performance for many drugs but genetics-guided provision is not applicable for all drugs. Actionability due to genetics is estimated to be the case for around 7% of drugs which account for around 20% of all prescriptions [40]. The list of gene-drug pairs with applicable interventions is still growing but it is likely that contributors with largest effects have been identified. It has been estimated that around 99% of people have an actionable variant affecting future drug prescriptions [83]. A classic pharmacogenetics success case is NAT2 testing for tuberculosis treatment with isoniazid. It was one of the first pharmacogenetic interventions shown to be clinically relevant and has thus far become standard practice [47]. Implications of lower doses were based on statistical associations between the allele variants and observed efficacy. Another classic case concerns testing CYP2C9*2/CYP2C9*3 alleles that affect warfarin pharmacokinetics (PK) [2]. Evidently, S-warfarin, one of the two warfarin components, is largely metabolised by CYP2C9. The allele carriers require lower doses of warfarin and have a higher risk of bleeding complications due to impaired hydroxylase that lowers the binding affinity of the drug [152]. For most drugs the pharmacogenomic pathways have not been fully elaborated. Uncharacterised variants and regulatory and epi-

genetic mechanisms are all possible factors in mediating the genomic effect on drug dosage, toxicity and efficacy. Possible environmental interactions such as drug-drug interactions and other sources of variability mean that drug monitoring via genotype assessment is not enough and in-treatment follow-ups are still required.

1.3.3. Clinical applications of pharmacogenetics

Coordinated efforts by medical and scientific societies aim to translate pharmacogenomic information into clinical practice. Several professional societies and consortia are working on developing practical pharmacogenetics guidelines and curated databases. In Europe consortia such as The European Pharmacogenetics Implementations Consortium and Ubiquitous Pharmacogenomics Consortium work towards enabling availability of pharmacogenetic testing in clinics and dissemination of guidelines for European populations [116]. Members of United States based Pharmacogenomics Research Network develop uniform guidelines for Clinical Pharmacogenetics consortium [43, 163]. Their European analogue, Royal Dutch Association for the Advancement of Pharmacy presented the first set of guidelines in 2011 [187]. Both focus on actionable translation of genetic information to clinical environments. The guidelines link genetic variation with a drug and give a recommended action with a certainty based on the amount of accumulated evidence. However, they do not inform clinical decision making and there are differences between the two guidelines. Curated guidelines need to be validated but randomised clinical studies are expensive and time-consuming to conduct [182]. New evidence may change and complicate existing interpretations if the two conflict. Ancestry based population stratification and differences in allele frequencies complicate transferability of guidelines between populations [121]. Evidently, practical difficulties in monitoring and curating the genome associated drug guidelines arise as information accumulates. Contradictory effect directions from multiple contributing alleles may suggest different courses of action. For this purpose, databases such as PharmGKB and PharmVar aim to aggregate current pharmacogenetic knowledge in concise summarised format and results from incoming independent studies with sometimes conflicting information are aggregated [54, 71]. Centralised curators disseminate the data whereas implementation guidelines are left to independent stakeholders and guideline-developers.

Development of pharmacogenetic guidelines aims to maximally account for evidence-based clinical pharmacogenetic utility. Genetics based decision making introduces a new source of information which requires novel competencies. The stakeholders are unlikely to accept new complexity and extra expenses if the utility of incorporation of genetic information in standard routine care is not clear [162]. Clinicians, the information gatekeepers need to be educated in transmitting the findings back to the individual patient. The treating physician would be greatly aided by supportive infrastructure, decision support methods and guidelines that

are handily accessible. This requires proactive planning and targeted development. The patient genotype data needs to be collected, stored and analysed in unison with other relevant medical data. Information technology systems are required to be able to process, maintain and communicate the data [196]. With the availability of patient genomic data most analysis steps can be pre-processed which cuts the required time and costs [97].

The incentives to nudge pharmacogenetic testing into mainstream use already exist. Patients pro-actively research their ailments online and increasingly demand state-of-the-art care. Single marker tests are most often performed when ordered but multi-gene testing panels may be shifting genetic testing towards preemptive approaches [162]. Not testing clinically identified pharmacogenetic markers has precedents of initiated class action law-suits [113].

1.4. Basics of therapeutic drug monitoring

Drug response variability may be attributed to physiological and environmental factors. Therapeutic drug monitoring (TDM) aims to guide pharmacotherapy using information gleaned from patient concentration measurements to attain therapeutically effective ranges of specified targets, therapeutic windows. This approach can help achieving individualisation of therapy, drug interaction monitoring, minimisation of under- and overdosing and withdrawal management [88]. Drug therapy that does not attain therapeutic windows can result in adverse effects [13]. For this, quantitative techniques help in quantifying the relationships between pharmacological concentration profiles and treatment response [189] and help in optimising expected outcomes in regard to dose amounts, administration times and dosing intervals [176].

For many drugs therapeutic drug monitoring is not needed or applicable but those that do benefit share several common characteristics [81]. Utility of TDM requires that the concentration measurements can be approximated to the therapeutic effect and that these can be used to derive interpretable clinical outcomes, and limited extent of intermediate metabolism processes of the active compound [88]. Environmental or physiological differences in drug metabolism can result in large between-subject concentration variability at effect site that increases uncertainty of outcomes. Essentially, the utility depends on the width of therapeutic window that can be narrower if between-subject variability is small but needs to be wider when the variability is larger [81]. Often clinicians need to avoid the development of antimicrobial-resistance through higher concentrations and considerations for special populations where the inter-individual variability may be larger [31, 167]. This decreases the size of effective target range and increases the likelihood of toxicity thus warranting more precise TDM. Drugs routinely monitored in clinical practice include digoxin, lithium, tacrolimus and vancomycin [58].

Several practical issues restrict the extent and optimal use of TDM applica-

tions. As in genetics, there is always a need for additional evidence of relationships between observable concentrations and the outcome to justify additional complexity. Each use case requires individual evaluation of efficacy and prospects [176, 193]. The drug must have an interpretable correlation between the concentration and effect [184]. Physiological differences between some subcohorts require care in development of TDM guidelines [8]. The outcome of TDM also depends on several external factors. Variation can be included from errors in timing of concentrations sampling, storage conditions, and measurement assay specifics [87]. Specific measurement assays may not always be available [58]. Drugs with better explained pharmacological properties and pharmacodynamics are therefore more adaptable to monitored treatment. There are also requirements for skill development in interpreting and communicating the TDM information which introduces new tasks for the physicians.

Personalised medicine is conventionally associated with genomics-based health care. Nevertheless, TDM belongs to same broad-based family of approaches [130]. It enables optimisation of therapeutic goals based on accrued evidence from related individuals with similar conditions and is based on patient data based guidances used in personalised clinical care provision. The aim is to derive definitive improvement of patient outcomes through incorporation of evidence other than clinical judgment [48].

1.4.1. Pharmacokinetics in therapeutic drug monitoring

Drugs vary in terms of disposition processes, therapeutic sites and physical effects [60]. Post-administration the drug is absorbed, distributed, metabolised and eliminated from the body. Pharmacokinetics aims to explain, model and further the understanding of drug disposition based on the drug concentration changes at the site of measurement in relation to time [81]. Modelled processes encompass all the effects of individual pathways related to drug function. The underlying assumption is to base basic pharmacokinetic techniques on physiology in the context of absorption-distribution-metabolism-elimination (ADME) processes. Pharmacokinetic methods and techniques rely on simplifications to approximate the true physiological processes and drug-body interactions. Models and techniques for this rely on a set of established principles such as explaining concentration courses through parameters that approximate some physiological property. Direct observance of drug disposition in *in vivo* systems is complicated. Mathematical approximations of physiological phenomena incorporate and are dependent on external treatment specifics such as method of drug administration, protein/tissue binding and pathways of elimination. In the following, we focus on intravenous infusion administration. This entails administering the drug directly to the bloodstream over a short period of time. As opposed to oral administration, all of the drug is instantly absorbed into bloodstream.

Drug concentration amount (Cp) in serum, plasma or blood is the pre-eminent

attribute that characterises all other dependent pharmacokinetic terms. It is directly measured. The volume of distribution (V) is a parameter that is estimated to approximate for the volume of drug distribution in the soluble space of an organism. These two parameters, at time t, are related through relationship $Cp(t) = \frac{Y(t)}{V}$, where $Y(t)$ is the amount of drug at time t. The drug disposition throughout the body is not usually uniform. For simplification, the organism is often divided in parts of equal solubility and access for the drug called compartments. As such V often signifies the volume of "central compartment" which represents well perfused regions (liver, blood) or plasma. Individual volume of distribution is a result of an interplay between the drug's fat and water solubility and body composition, and the activity of drug transporters and other enzymes involved in metabolism pathways [177]. Clearance (CL) is the volume from which the drug is eliminated per unit of time. Total clearance is the sum of all individual organ contributions towards elimination. Two main pathways for drug elimination in human are kidneys and liver. A smaller fraction may be eliminated via other metabolic organs. Elimination rate constant, k, links the two primary terms by division (CL/V) and expresses the fraction of the remaining drug that is eliminated per unit of time [36].

An analytical parameter that captures all concentration changes throughout therapy administration is the area under the curve (AUC). In a one-time administration it links the drug dose and the total clearance by evaluating the area of concentration curve over time $AUC = \int_0^\infty Cp(t)dt = \frac{D}{CL}$, where D is the amount of administered dose. Often, concentration measurements are sparsely available and simplifications are applied in practical AUC estimation with methods such as linear trapezoidal rule AUC_{0-t_n} for periods where concentrations increase

$$AUC_{0-t_n} = \sum_{i=1}^{n} (\frac{Cp(t_i) + Cp(t_{i+1})}{2}) * (t_{i+1} - t_i)$$

and the log-linear trapezoidal when concentrations decrease [53]

$$AUC_{0-t_n} = \sum_{i=1}^{n} \frac{Cp(t_i) - Cp(t_{i+1})}{\ln \frac{Cp(t_i)}{Cp(t_{i+1})}} (t_{i+1} - t_i).$$

Here, t_i represents the time of a specific concentration measurement and $Cp(t_i)$ the corresponding value, where $i = 1 \ldots n$ are indices for available measurements. The general idea is to extrapolate between available measurements. Log-linear method reduces overestimation compared to linear estimates if the speed of the process is related to the remaining drug amount in plasma. In treatment scenarios with multiple intermittent doses the aim is to arrive to steady state where subsequent dose administrations with consistent intervals yield identical time-concentration curves and therapeutic range is most accurately targeted. Importantly, AUC of a single dose is equal to its interval AUC in steady-state [81].

ADME drug disposition processes characterise drug transmission through the body. A central problem is modelling the amount of drug Cp at the effect site at a given time t. A full time-concentration profile consists of concentration increase by administration and depletion by elimination. This process of concentration changes can be captured with description of rate changes in the general form:

$$\pm\frac{d(Cp(t))}{dt} = A * (Cp(t))^n.$$

Exponent n determines the order of processes and A is the rate of concentration change. In pharmacokinetics, a zero-order kinetic process expresses the rate of administration/elimination that is independent of the existing drug concentration eg. elimination of ethanol:

$$Cp(t) = c \pm A * t.$$

Here, c is a constant signifying the initial amount. In a first-order kinetic process the rate of change is dependent on the existing concentrations and increases:

$$Cp(t) = c * e^{\pm A * t}.$$

Combination of directional rate processes provide the basis for drug concentration change modelling for the vast majority of drugs [81].

The rate processes provide a coherent way to model physiological drug PK by relating dose information and time to concentration changes. An important simplification for concentration modelling using rate processes relies on grouping the body in compartmental sections. This assumes similar disposition properties within a compartment and enables capturing drug movements with a single rate process between connected compartments. Combination of rate processes of various order can simulate different modes of administration, amounts of compartments, elimination, absorption pathways and PK dynamics [81]. These considerations guide the selection of a structural model. Fortunately, analytical solutions for differential equations exist for several commonly found set-ups of drug kinetics changes. A standard one compartmental model assumes that the drug is administered into a compartment of certain volume by some rate and eliminated at a different rate. A standard two compartmental model includes a second compartment which exchanges contents with the former at different speeds. This aids in simulating a two-tiered drug elimination curve whereby the faster elimination phase is followed by a slower one [132]. This thesis highlights the one-compartmental model with multiple administration episodes of intravenous infusion with zero-

order absorption and linear elimination of first-order rate:

$$Cp(t) = \begin{cases} \frac{D_n}{Tinf_n}\frac{1}{kV}(1-e^{-k(t-t_{D_n})}), & \text{if } t-t_{D_n} \leq Tinf_n \;\&\; n=1 \\ \sum_{i=1}^{n-1}\frac{D_i}{Tinf_i}\frac{1}{kV}(1-e^{-kTinf_i})e^{-k(t-t_{D_i}-Tinf_i)} & \\ \quad +\frac{D_n}{Tinf_n}\frac{1}{kV}(1-e^{-k(t-t_{D_n})}), & \text{if } t-t_{D_n} \leq Tinf_n \;\&\; n>1 \\ \sum_{i=1}^{n-1}\frac{D_i}{Tinf_i}\frac{1}{kV}(1-e^{-kTinf_i})e^{-k(t-t_{D_i}-Tinf_i)} & \\ +\frac{D_n}{Tinf_n}\frac{1}{kV}(1-e^{-kTinf_n})e^{-k(t-t_{D_n}-Tinf_n)}, & \text{otherwise.} \end{cases}$$

Variable D_i is dose amount, t_{D_i} is the time of dose, $Tinf_i$ is duration of infusion for the dose administration i where $i = 1,\ldots,n-1$. Also k is the elimination rate, V is volume of distribution and $Cp(t)$ is concentration at time t. Variables D_n, t_{D_n}, $Tinf_n$ represent analogous quantities but for the last, nth, observed dose administration [39]. The idea is to characterise concentration movements over time given an arbitrary dosing schedule. The concentrations increase when less time than the duration of infusion has passed from the dosing event and decrease otherwise. The first case is used to evaluate concentrations for single-administration events at such times t when the infusion is on-going. The second case is used to evaluate concentrations in cases with more than 1 administration at time-points t till the end of last infusion event and the third case is used to describe the decreases in concentration for any $t > Tinf_n + t_{D_n}$.

The compartmental models aim to link time and dosages with concentrations. The concentration predictions are also influenced by values of physiological PK parameters such as clearance and volume of distribution that are not readily available. Therefore, the non-linear compartmental models are often used for modelling the drug concentrations so to estimate the related pharmacokinetic parameters. Additionally, properties of the drug often warrant inclusion of other covariates that affect and stratify PK such as age, weight, co-administration of a different drug *etc*. These other measurable and predictive variables are often modelled as covariates of the PK parameters through allometric, proportional, linear or exponential functions and standardised in relation to population or expected reference values. A realisation of pharmacokinetic parameter θ_{pop} may then in reality be a function of any interlinked explanatory variables $\theta_{pop} = f(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ is the vector of additional covariates and $f()$ is the linking function. Relevance of covariates is determined by relying on physiological assumptions and their effect on improving model performance. The latter is measured with residual analyses and statistical tests [132, 170] .

Pharmacokinetic data are longitudinal and often sparse; datasets are made up of a fairly small number of samples with unbalanced number of measurements. To extract most value from these data, in time, several approaches have been used

in parameter estimation. Previously, the two-stage approach relied on distinct estimation of individual kinetic parameters [20]. First stage estimates individual parameters by non-compartmental approaches and *AUC* based concentration curves. In the second phase, individual estimates are aggregated by standard statistical estimates such as standard deviation and the mean. Another popular way is to apply weighted non-linear least squares regression [134]. The weakness of that approach lies in sensitivity to the number of samples available per patient, the issue of weight selection and threat of over-fitting. Two-stage and weighted regression are not population-based as estimations are aggregated from individual patient samples.

Currently, the non-linear mixed-effects modelling is the most popular method for developing population-based PK models [132]. Hierarchical modelling captures between-subject variability of physiological parameters as a surrogate for population-wide concentration variability. A pharmacokinetic non-linear mixed effect model could be represented as $Cp(t) = f(t, \boldsymbol{\theta})$, where vector $\boldsymbol{\theta}$ represents physiological parameters that are assumed to be normally distributed and t is time. Assuming independence between different physiological parameters, the vector of random effects $\boldsymbol{\eta}_{\theta_1}$ describes population-wide concentration variability for a specific pharmacokinetic parameter θ_1 and its individual elements are distributed by $N(0, \omega^2_{\theta_1})$ where $\omega^2_{\theta_1}$ is the population variance of the PK parameter. Another expansion of structure accounts for model residual variability with error terms. In the context of longitudinal concentration values, a fair presupposition is to expect larger unexplained variability at higher concentrations. As opposed to additive residual structure ($Cp(t) = f(t, \boldsymbol{\theta}) + \varepsilon_{0t}$), proportional ($Cp(t) = f(t, \boldsymbol{\theta}) * (1 + \varepsilon_{0t})$), exponential ($Cp(t) = f(t, \boldsymbol{\theta}) * \exp(\varepsilon_{0t})$) and combined models ($Cp(t) = f(t, \boldsymbol{\theta}) * (1 + \varepsilon_{0t}) + \varepsilon_{1t}$) allow for heteroskedastic residual modelling. Here, a common simplification assumes independence between different error terms and between error terms and parameter values. Here, instances for ε_{zt} at time t for error component z are assumed to be drawn from the distribution $N(0, \sigma^2_z)$), where σ^2_z is the variability estimate of the zth error term [132, 170].

Pharmacokinetic model fitting produces models for the PK parameters such as clearance and volume of distribution that consist of estimates accompanied by any additional variables explanatory variables. In the case of a mixed-effect model the between-subject variability of PK estimates can be described as the population-wide variability of the concentrations. As such this model of pharmacokinetic parameters allows simulation of individual PK parameter values for a population using draws from the random effect distribution as $\theta_i = g(\eta_i, \theta_{pop})$ where $g()$ is a link function that captures the relationship between individual values to fixed effect population level estimates for the individual i and θ_{pop} is the population value of the PK parameter. As above η_i is an individual estimate for random variation around the population estimate of the corresponding PK parameter θ_i. The link function is often expressed through an exponential component $\theta_i = \theta_{pop} * \exp(\eta_i)$. Proportional ($\theta_i = \theta_{pop} * (1 + \eta_i)$) and additive structural forms ($\theta_i =$

$\theta_{pop} + \eta_i$) are also in common use.

Pharmacokinetic modelling is relevant for most drugs which can benefit from TDM for dose optimisation and concentration monitoring [57]. However published range of results, structural models and covariates often differ even within one drug and cohort [120, 203]. Several attributes affect the success of pharmacokinetic study outcomes. Sampling design is crucial because of limited availability of samples and per patient concentration measurements [123]. Sample unavailability may be caused by the rarity of the condition or because the drug concentration measurements pose additional health threats. In turn, limited sample sizes reduce the number of covariates that can be modelled and increase estimation biases [139]. These in turn affect model validation which translates to generalisability and over-fitting issues [17].

1.4.2. Therapeutic targets

Structurally modelled pharmacokinetics provides a way to link individual drug kinetics estimates to context specific therapy adjustments. Drug administration in therapeutically monitored treatment is followed by a concentration measurement that is converted to a desired target index and evaluated against index values associated with improved treatment outcomes. Modifications for the next dose administration are done so that discrepancies between observed and expected targets are reduced and to improve the probability of target attainment (PTA). This process is repeated for the duration of the treatment [88]. TDM requires availability of target metrics and their optimal values or ranges. Such pharmacokinetic/pharmacodynamic (PK/PD) indices are commonly established by associating the treatment endpoints to drug kinetics.

The most common treatment endpoint that is monitored in TDM via PK/PD indices is efficacy [32]. For some drugs such as gentamicin and vancomycin, toxicity monitoring is also important [67, 95]. In the broad class of antimicrobial drugs, the mechanism of anticipated drug effect is a good indicator of the most suitable PK/PD index [195]. Still, final index selection is mostly based on best observed correlations with the endpoint [133]. Penicillins, macrolides and carbapenems exhibit time-dependent effect and minimal persistent effects, and are often evaluated by the time of concentration above the minimum inhibitory concentration ($t > MIC$). The PK/PD index of concentration dependent drugs with prolonged persistent effects such as ketolides and aminoglycosides is the maximal concentration over minimum inhibitory drug concentration (Cp_{max}/MIC). On the other hand, concentration-dependent drugs with moderate to prolonged persistent effects are often evaluated by dividing area-under-curve of the pathogen by minimum inhibitory concentration (AUC/MIC) [6]. Vancomycin and clindamycin are examples of the last group [108]. Consequently, drug related PK/PD index discriminates efficiency outcomes on a certain threshold. For most drugs the debate about the optimal values for desired PK/PD indices is on-going.

Computation of pharmacodynamic indices by derivation from concentration measurements may result in loss of accuracy. AUC estimates are not always consistent due to differences in the choices of time-concentration curve fitting model, sampling times and estimation method [94]. In some cases, the approximate PK/PD indices can be substituted with direct indicators such as peak or trough concentration values (Cp_{trough}) or respective concentration ranges thanks to correlations between different PK/PD indices. For example, ample evidence for associations between trough concentration values to AUC/MIC for vancomycin in paediatric populations show that trough concentrations of 10 mg/L are sufficient to achieve $AUC/MIC > 400$ [22, 52, 98, 194] but contradicting evidence has also been presented [138]. Directly correlating concentration profiles with efficiency indices can provide a more straightforward application of TDM. Because they are easier to measure and to interpret, the Cp_{trough} values are somewhat easier to use in clinical settings.

Quantitative TDM requires both known efficiency related indices and a systemic approach to performing individual adjustments [167]. Traditional approaches rely on nomograms or charts designed for the current drug and patient cohort set. Simpler rule of thumb adjustments apply proportional dose adjustments depending on the direction of error. Model based monitoring is most useful for drugs with narrow therapeutic windows and high inter-individual variability such as vancomycin [136]. Larger differences in population-wide variability and narrower therapeutic range make patients more susceptible to toxic over-dosing or inefficient under-dosing. Therefore, related improvements in target attainment can have a considerable real-world therapeutic effect. For this, precision dosing methods aim to apply computational methods to improve treatment provision by higher target attainment of therapeutic indices [137].

1.4.3. Precision dosing

Bayesian framework derives probabilistic inferences by updating prior beliefs with observed data. Solution of a Bayesian parameter estimation, conditional on the observed data, is a full distribution of parameter values instead of a point estimate. These are derived on the basis of a Bayes rule that is used to combine evidence with prior beliefs

$$P(\theta|x) = \frac{(P(x|\theta) * P(\theta))}{P(x)}.$$

Here, $P(\theta|x)$ is the posterior distribution of parameters conditioned on the data, $P(x|\theta)$ is the likelihood function for probabilities of observing the data conditional on unknown parameter values. $P(\theta)$ is the initial belief or prior about the distributions of parameter values. The evidence $P(x)$ is the scaling factor for the density function to scale the sum of probabilities to 1. The evidence can be evaluated as an integral over all possible parameter values in continuous parameter space

$$P(x) = \int P(x|\theta)P(\theta)d\theta.$$

In the absence of conjugate priors for likelihood distributions, the analytical evaluation involves often intractable high-dimensional integration [100].

The foundations for Bayesian applications in pharmacokinetics were made in the 1980s [175]. Bayesian methods are a natural extension to compartmental population PK models which allow including new concentration measurements to update population estimates. In Bayesian context, kinetic compartmental models serve as likelihood function which ties the measured individual drug concentrations with the dosing information (dosing times, duration of doses, amounts) and the PK parameters as $Cp(t) = f(t, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the physiological pharmacokinetics parameters. Commonly, population model between-subject estimates and residual model error estimates would be respectively used for parameter variability and unexplained variability estimates with the corresponding structural model that was used in population-wide PK estimation.

A solution for the posterior PK parameters with a Bayesian PK model is the product of the compartmental likelihood function and the prior distributions of PK parameters normalised by the prior-likelihood product over the space of all possible pharmacokinetics parameter values. First, the normalising constant is intractable and involves a multi-dimensional integration. Secondly, the likelihood function is non-linear and conjugate solutions are not often available. Such cases have spurred development of approaches that provide approximate solutions. Intractable Bayesian equations are often solved using numerical Monte Carlo Markov Chains. The aim is to sample the parameter space so that the retrieved samples produce a posterior distribution of target parameters. For example, given the likelihood function and prior distributions, the Metropolis-Hastings algorithm is used to perform a random walk in the parameter space. Assuming the ergodicity and stationarity of the Markov chain, the walk converges towards a stable final distribution [96, 166].

Non-linear mixed PK models can be used to predict concentration profiles stratified by selected model attributes. The main benefit of a Bayesian approach is in providing iterative updates to the final PK model with individual concentration data. Each timed concentration measurement weighs the population PK distribution towards a distribution that would attain observed concentrations under the likelihood model. Posterior estimates provide the basis for accurate individual treatment course simulations and dose optimisations. Predictive improvements that result from inclusion of individual data over general population modelling has been demonstrated for many drugs, including tacrolimus, ciclosporin, vancomycin, voriconazole, busulfan and many others [73, 75, 117, 137, 143, 164, 202]. Bayesian TDM tools have been developed by both both academic and private actors. Tools such as NONMEM, TDMx (http://www.tdmx.eu/), BestDose (http://www.lapk.org/ bestdose.php) and DoseMe ((https://doseme.com.au/)) pro-

vide the computational frameworks to combine population information with individual data.

Even though precision based dosing of drugs with narrow therapeutic windows and large between-subject variability have the potential for large cost-savings [12], the clinical adoption of precision based dosing methods has not been straightforward [12, 38]. Computational TDM requires novel skills and clinically validated tools which are to be implemented to accompany other bed-side apparatus [140]. Such systems need to be prospectively and clinically validated for feasibility before clinical use [59]. Organisational problems include lack of sound guidelines and uniform standards that creates mistrust as physicians are not convinced of the benefits.

1.4.4. Vancomycin pharmacokinetics

Vancomycin is a commonly used antibiotic used in neonates for decades. Its popularity is driven by high methicillin-resistance rates of coagulase negative staphylococci (CoNS) and the spread of Methicillin Resistant Staphylococcus Aureus (MRSA) [80]. Vancomycin has a relatively narrow therapeutic target and time-dependent efficacy. Still, there is disagreement about exposure levels of under- or overdosing and concentration profiles that improve treatment outcome. In adults, the PK/PD index which is best associated with efficacy is AUC/MIC [171]. There is some evidence that in adults PK/PD target $AUC/MIC > 400$ improves treatment outcome with MRSA pneumonia [129]. This result is inconclusive for neonates. Optimal target index values may vary by pathogens, clinical conditions and patient sub-cohort specifics [74, 79]. Some evidence indicates that CoNS infections require lower exposure in neonates compared to *staphylococcus aureus* conditions [180]. Bedside derivation of AUC from concentration measurements is inconvenient and inconsistent [82]. Instead, Cp_{trough} monitoring is widely used as surrogate for the AUC/MIC. In neonates Cp_{trough} values around 10 mg/L are claimed to be sufficient for achieving $AUC/MIC > 400$ [22, 52, 98, 194]. Evidence on toxicity in neonates is limited [107], Lodise suggests that increases start from Cp_{trough} values above 20 mg/L [112]. Correspondingly, the higher end of AUC/MIC values that does not increase adverse effects has been claimed to be around 700 $\frac{mg*h}{L}$ [138, 186]. Others have associated vancomycin nephrotoxicity with cumulative exposure [95]. Values of AUC/MIC are highly dependent on the minimum inhibitory concentration of the pathogen. Therefore, common dosing schemes do not apply due to potentially toxicity inducing concentrations in cases of $MIC > 2$ mg/L. Padari *et al.* reported underachieving $AUC/MIC > 400$ and $AUC/MIC > 300$ compared to targeted attainment levels with common dosing schedules [148]. This may corroborate that fear of toxicity may guide neonatologists to choose more cautious dosing schemes and err on the side of underdosing [78].

High population variability and need for efficiency related targets have made

vancomycin widely investigated in PK/PD studies. Many studies focused on vancomycin PK characterisation exclusively on neonates [3, 7, 52, 93, 111, 119, 145, 174, 204]. Elimination clearance in neonates with normal renal function is around 0.04-0-09 $\frac{L}{kg*h}$ and volume of distribution is around 0.57-0.7 L/kg. The clearance of pre-term neonates is lower due to renal and hepatic immaturity. In turn, proportionally more extracellular liquid increases the volume of distribution [171]. Marsot *et al.* estimated that the mean between-subject variability of vancomycin clearance parameter was 30% and 23% in case of volume of distribution [120]. PK in neonates is often described using a one-compartmental model whereas adult PK is commonly modelled using two compartments. Most models include age (postnatal, post-menstrual or gestational) and birth- or current weight. Attributes such as creatinine, co-administration effect with inotropes and artificial ventilation have also been used in published models [120]. Generalisability of the models remains an issue but more are being externally validated, standard processes for model development are forming and sample sizes are increasing [203]. This state of improvement is a starting point for translating model information towards practical scenarios.

High variability of population PK, uncertainties regarding therapeutic PK/PD targets in special cohorts, changes in pathogen resistance rates and administrative complexities all attribute to low precision dosing method adoption in TDM [12, 38]. Currently, dose modifications are mostly guided by nomograms/charts due to their applicability, simplicity of use and robustness as computerised clinical systems have not really gained widespread use [140]. Nevertheless, vancomycin is used for treating therapeutically serious conditions, it has a narrow therapeutic target and preliminary evidence exists for efficacy related PK/PD indices so one-size-fits-all criteria rarely applies in vancomycin treatment and TDM is also actively encouraged for patients with altered PK [74]. These features make vancomycin a good candidate for precision based TDM to benefit from individualised patient treatment [150].

2. PRECISION DOSING OF VANCOMYCIN IN NEONATES

Even though vancomycin has been marketed for more than 50 years, its use is steadily growing due to increasing rates of resistance of gram positive bacteria to standard lines of treatment [128]. The pharmacokinetic/pharmacodynamic (PK/PD) indices that associate with optimal treatment outcome in neonates are not well ascertained but effective standardised treatment regimens rely on established PK/PD targets that maximise efficacy and minimise toxicity. Our previous work in University of Tartu showed that doses commonly used for neonates do not attain the recommended AUC/MIC targets [148]. The primary goal of the Neovanc consortium (2014-2019; Horizon2020; Framework Programme 7) was to resolve the uncertainty about PK/PD indices for vancomycin in neonates using a randomised clinical trial. It consisted of 12 European partners including a group in University of Tartu led by prof. Irja Lutsar.

To supplement the on-going consortium work, we additionally saw the need for a tool that would help the clinicians better achieve fixed PK/PD targets through dosing changes as high variability of PK also complicates target attainment. In Ref I, this inspired us to develop a web-based dosing tool DosOpt that we have made available at www.biit.cs.ut.ee/DosOpt. Other than developing a tool for simulating personalised therapy scenarios with variable PK/PD targets, dosings and individual responses, it provided us with an opportunity to observe the dynamics in attainment of therapeutic windows resulting from simulated dosing changes based on individually modelled concentration data.

Concentration measurements in TDM are scarce. DosOpt overcomes the lack of individual data by using estimates for the pharmacokinetic parameters from academically publicised population PK models as priors to be combined with concentration measurements. Since published models include a variety of different covariates with highly varying estimates then DosOpt naturally emerged as a platform for performing comparative evaluations of different PK models based on assessments of differences in their simulated attainment rates and predictive accuracy. Therefore, in Ref II we set out to elucidate the landscape of vancomycin PK models in neonates with the goal of informing the DosOpt population prior.

DosOpt provides a practical use case for TDM based individual dose optimisation through the example of vancomycin in neonates. Based on retrospective validation, it serves to illustrate the feasibility of computational TDM for special cohorts, the benefits of individualised dose optimisation but also provides a basis for continued developments into bringing such methods more widely into clinical care.

2.1. Web-based dosing tool - DosOpt (Ref I)

Practical use-cases highlight the utility of model based TDM for vancomycin [73, 106, 143, 150]. Computational frameworks such as NONMEM provide the user with functionality to develop the models and code their own analyses but practical applications also need to have user-geared interfaces and third party usability [1, 38]. These applications are still largely missing for special populations such as neonates [34].

DosOpt is an on-line tool that applies Bayesian methods in combining in-treatment individual concentration measurements with population level estimates for population PK parameters to obtain individual estimates. Generally, the use of DosOpt follows a set of operations. The user selects the prior and uploads in-treatment concentration measurements that are used to model individual pharmacokinetics estimates. Based on these the users can simulate individualised time-concentration profiles and optimise doses against therapeutic indices. Modelled scenarios can include any user provided custom design for dose administration time, infusion length, dosing interval or desired PK/PD index. The primary use of DosOpt is the simulation of optimal dosing schemes under a desired therapeutic PK/PD target value based on previous dosing and concentration measurement history. The optimal dose is probabilistically most likely to attain the desired target. This dose is provided back to the user with the corresponding probability.

DosOpt was implemented with R Shiny [21] and the Bayesian modelling operations used JAGS (Just Another Gibbs Sampler) [157]. The first model for priors that was implemented in DosOpt was from Anderson *et al.* [7] that uses a one-compartmental population model with zero-order input and first order elimination. This model estimated the volume of distribution as $V_{pop} = V_{std} * V_{allom} * 1.18^{Inot}$, where allometric component $V_{allom} = (Wt/70)$, Wt is body weight and $Inot$ indicates the use of inotropes. Population level estimate $V_{std} = 39.4$ and the between-subject variability $\omega_V^2 = 0.197$. Correspondingly, clearance CL was modelled as $CL_{pop} = CL_{std} * CL_{allom} * CL_{renalmat} * 1.03^{Vent}$, where $CL_{allom} = (Wt/70)^{0.75}$, renal maturation component $CL_{renalmat} = PMA^{3.68}/(PMA^{3.68} + 33.3^{3.68})$, PMA is post-menstrual age in weeks and $Vent$ indicates the use of positive pressure artificial ventilation. Population level estimate $CL_{std} = 3.79$ and the between-subject variability $\omega_{CL}^2 = 0.209$. The distribution of PK parameters was assumed to be log-normal then pharmacokinetic parameter values for the ith individual were modelled as $CL_i = CL_{pop} * \exp(\eta_{CL,i})$ and $V_i = V_{pop} * \exp(\eta_{V,i})$. The one-compartmental model that ties together the concentration data with PK parameters was used as a likelihood function. Posterior estimates for the physiological pharmacokinetic parameters are simulated using the Metropolis algorithm in JAGS (Ref I, Fig. 1) which are then used with dosing information to construct individual time-concentration curves and optimise for doses that attain therapeutic targets with maximal attainment.

We used several different datasets in evaluating the performance of the Dos-

Opt tool with the Anderson *et al.* model. First, we assessed the modelling bias and precision by using a simulated population with fixed pharmacokinetics. This was done by simulating attributes required in the Anderson model *et al.* for 1,000 patients and then used the model of PK parameters to simulate individual PK values. We then used the simulated PK values for time-concentration curves under a pre-designated dosing schema assuming a 15 mg/kg loading dose followed by three 10 mg/kg infusions with 1-hour duration in 12-hour intervals, and extracted concentrations at five reference time points. Then, we modelled these individuals with DosOpt with variable number of included individual concentration points to predict concentrations not included in modelling at remaining reference time points. We evaluated the prediction errors using the normalised prediction distribution error (NPDE) approach and several other error measures (mean absolute error [MAE], mean absolute percentage error [MAPE], mean percentage error [MPE]).

Secondly, we used the same simulated population to assess the accuracy of the probabilities assigned to target attainment in dose optimisation. For each simulated individual with fixed PK, we established the range of doses that obtained Cp_{trough} within 10-15 mg/L. The modelled comparison data was obtained by using the extracted reference concentration measurements to retrieve the most probable dose suggested by DosOpt, the corresponding predicted probability and if this dose was within the range of feasible doses. Any observed predictive biases in the modelling would thus implicate problems in DosOpt modelling processes, algorithm or implementation.

Thirdly, we used a retrospective clinical dataset to evaluate the predictive performance of DosOpt on real patients from the paediatric intensive care unit of Tartu University Hospital that received vancomycin treatment within January 1, 2010, and December 31, 2015. Each included patient had at least one measured trough concentration. Assessments of predictive accuracy were evaluated in terms of MAE, MAPE and MPE as in simulated data. Here, we used a different number of available concentrations starting from base PK models to predict concentrations in all time points with known concentrations that were not included in modelling. For assessments of real world performance, we also used patients in our retrospective dataset to optimise for doses that gave highest attainment probabilities in Cp_{trough} range 10-15 mg/L and compared these to attainment proportions observed in the retrospective dataset. For this we used individually modelled PK estimate distributions that were adjusted by prediction error distributions specific to the population with the matching number of concentrations included in the model. Then each individual's documented dosing schedule was extended with another dose that was optimised so that it maximised the probability of resulting in a Cp_{trough} within a target window.

The predictive performance analysis results on simulated data showed that mean prediction error estimates were not different from zero regardless of the number of modelled individual concentrations. Increase in the number of individ-

ual measurements decreased mean absolute errors and mean absolute percentage errors as new data shrank estimates towards true values (Ref I, Table 2). The result was the same when simulated concentrations were added inter-occasional variability with zero-mean and standard deviation of 15% of the concentration value (Ref I, Supplementary Table 3). However, variability of normalised prediction distribution errors decreased significantly below 1 with more than two individual concentrations. DosOpt model predictions converging towards true values faster than expected means that the probabilities of attaining the best doses were underestimated. This phenomenon was illustrated in the target attainment probability evaluations in the simulated dataset. On average, DosOpt provided a 48% probability of target attainment in the 10-15 mg/L trough target when no additional concentrations were included to modelling. Since values of the pharmacokinetic parameters were known in the simulated dataset then we could confirm that in reality the dose suggested by DosOpt was sufficient for attaining desired trough levels in 45.8% of cases. When the number of concentrations that was included in modelling was increased to 3 then the mean probability from DosOpt increased to 81.5% but the actual proportion of simulations within the target window rose to 96.2% (Ref I, Fig 4).

The qualifications on clinical data were performed on a retrospective test dataset consisting of 149 individual treatment episodes from 121 patients with 1-10 measurements (median 2) from each (Ref I, Table 3). DosOpt allowed us to use each patients' own dosing schedule. Mean percentage errors with Anderson *et al.* model were biased when less than two individual concentrations were included (Ref I, Table 4) but only one concentration was required to improve predicted optimised dose target attainment in Cp_{trough} range of 10-15 mg/L above observed rates in the retrospective data set (Ref I, Fig. 5). Each additional individual concentration also decreased MAPE and MAE.

DosOpt is publicly accessible at www.biit.cs.ut.ee/DosOpt. It is designed to have a user-friendly interface and requires no technical skills besides formatting TDM data into pre-designated format. Tests with simulated data indicated that DosOpt modelling is unbiased and increasing the number of individual concentrations decreases forecasting error. Application of the Bayesian approach based on Anderson *et al.* model on retrospective data improved therapeutic attainment rates above those observed in clinic with just one individual concentration. As the lead author, my contributions were the design of the study, development and implementation of the tool, all analyses and writing up the manuscripts.

2.2. External evaluation of pharmacokinetics models (Ref II)

Initial evaluation of DosOpt was based on PK population estimates from Anderson *et al.* We observed that this model prior combined with a single individual concentration was sufficient to improve target attainment in the 10-15 mg/L Cp_{trough} range above hospital attainment rates even whilst with one concentration

the predictions remained systemically biased. We then hypothesised that these results could likely be improved with priors that match more closely with the target population as the number of neonatal vancomycin PK models published in academic literature is quite extensive [3, 7, 52, 93, 111, 119, 120, 145, 174, 204]. Relatively more precise prior estimates would require fewer individual measurements to adapt to individual PK which in turn would lead to more accurate dosing and concentration predictions.

In a clinical TDM scenario, the aim of collecting concentration predictions is to guide patient dosing. Individual concentration measurements provide individual level information that is not accounted in population model based predictions. Data based adjustments are expected to decrease unexplained variability as demonstrated in a similar analysis for tacrolimus [202]. External evaluations test generalisability of model performance outside training data but in reality many developed pharmacokinetics models have not been externally validated owing to high capital and time costs of collecting such datasets [17,203]. Therefore, we set out to test performance differences from the use of different literature based PK priors.

We aimed to gather all population PK models for neonates in vancomycin. The models needed to have been described in sufficient detail for implementation in DosOpt with all the variability measures and structural models fully described. All evaluations were performed using a retrospective dataset of patients collected in Tartu University Hospital between 2010 and 2015 with postnatal age < 90 days. This was the same dataset as used for retrospective evaluations in Ref I. Population model predictions were assessed using normalised prediction distribution errors (NPDE), MAE, MAPE and MPE over all individually available concentrations. Evaluations of Bayesian predictive accuracy correspondingly evaluated absolute error (AE), percentage errors (PE) and absolute percentage errors (APE) to predict a value for a known concentration using a specified number of previously known measurements. We also assessed the proportion of percentage errors within 20% and 30% of true concentrations. Next, we employed a simulated dosing scenario with patients in our retrospective dataset to estimate model-wise probability of target attainments in Cp_{trough} between 10-15 mg/L and 10-20 mg/L.

The retrospective dataset contained 309 concentration measurements from 149 treatment episodes from 121 patients (Ref II, Table 1). There was a total of 149, 84 (56.4%), and 38 (25.5%) patient treatment episodes that had at least 1, 2, and > 2 vancomycin time- concentration points available, respectively. We performed a literature review for population pharmacokinetic models from academic literature (Ref II, Fig. 1). Final model selection yielded 12 candidate models including the Anderson *et al.* model (Ref II Supplementary Table 1) [3, 7, 52, 93, 111, 119, 120, 145, 174, 204].

All published models were biased in population-model based validation as evidenced by NPDE analysis results (Ref II, Supplementary Fig. 1). In comparison of predictions made without any individual data and all individual data included

(Ref II, Table 2; Supplementary Fig. 2), we observed improvements in adjusted R^2 values, reduced median MPE and MAPE values for all models. This indicates that compartmental models are a suitable fit to the data.

Inclusion of a single individual concentration in forecasting improved both precision and accuracy metrics compared to results from only model based prediction. The inclusion of a second concentration did not result in additional major improvements (Ref II, Fig. 3; Supplementary Table 3). All models got less than half of predictions within 30% of the measured concentrations when forecasting did not include individual concentrations. Inclusion of a single concentration attained more than half of the predictions for 8 of the twelve models within the 30% range. Six models had more than 40% of predictions within 20% percentage error range. Importantly, predictions for the subset of individuals with more collected concentrations generally had larger percentage errors.

Probability of target attainment evaluations indicated improvements for targeted Cp_{trough} dosing within 10-15 mg/L and 10-20 mg/L when base model predictions were compared with those based on one individual concentration. On average, the probability of target attainment estimates in Cp_{trough} 10-15 mg/L improved to around 40% with two included concentrations and to 60% within Cp_{trough} range 10-20 mg/L. We observed that PTA estimates improved with individual concentrations by about 25% (Ref II, Fig. 4; Supplementary Table 5).

This study is the first to consider the effects of using individual concentrations in pharmacokinetics re-estimation for vancomycin administration in neonates. Improved precision and accuracy from precision dosing translate to higher probabilities of attaining therapeutic targets. Relatively low predictive precision of predictive models limits attainment in narrow therapeutic targets. In our simulations of PTA estimates, the best results were obtained with the model by Zhao *et al.* which is a prior candidate for prospective DosOpt evaluation but different PK models may have use cases depending on available covariates and specific patient cohorts [204]. All evaluated models were implemented in DosOpt. My contributions in this article were in the same extent as in Ref I.

3. GENETICS OF ADVERSE DRUG EFFECTS

Tools like DosOpt are useful in raising efficacy for treatments that can be actively monitored and are directly actionable for instance through dosing changes. This is not the case for many drugs self-administered at home additionally complicated by the difficulty of keeping to the prescribed medication schedule in an every day life. Patient DNA is largely immutable and provides a source of information for raising the probability of a successful treatment for many such cases. Since the metabolism processes of many drugs are affected by products of only a limited number of genes then variation in the gene products or their regulation can have a large effect on the drug response.

Most pharmacogenomic studies focus on limited sets of drugs and genomic regions. In contrast, we were able to leverage the whole genome sequenced and imputed samples in the Estonian Biobank linked with population registries on drug prescriptions and documented illnesses. In Ref III we performed a population-based assessment by looking for adverse drug effects among individuals that had been prescribed various drugs. This allowed us to test previously known associations but also use a hypothesis-free approach in assessing novel regions.

3.1. Population-based discovery of pharmacogenetic adverse drug effects (Ref III)

Adverse drug effects are a frequent and severe problem in drug therapy [99]. In fact, ADEs cause up to a tenth of all hospitalisations with considerable costs on healthcare services [155]. Many of these effects could be predicted as many drugs are metabolised by the same set of pharmacologically important VIP genes (Ref III, Supplementary Table 2) and in many instances variants in VIP genes have been shown to have a large effect on drug ADME processes [46]. However, drug and health related phenotypes are slow and expensive to collect. Associations studies on drug response have been lagging compared to other phenotypes [201]. Compared to studies with targeted recruitment, population-based studies use a hypothesis-free approach to amass larger sample sizes. This can result in unexpected discoveries. To elucidate the field, we designed a study that investigates the relationships of ADEs related to drug prescriptions and genetics. Ref III is a proof-of-concept study to confirm and discover biomarkers related to adverse drug effects using multiple sources of population-based data: whole genome sequences, electronic health data and drug prescription records.

Our analysis was based on the 2,240 whole-genome sequences and more than 16,000 imputed genotypes of Estonian Biobank participants. The imputation applied a custom Estonian reference panel of $16.5 * 10^6$ SNVs [127]. ADE instances were defined based on a pre-defined list of 79 medical condition related, ICD-10, codes. These were additionally grouped into 12 mechanistic categories (Ref III, Supplementary Table 1). The ICD-10 codes and drug prescription data were ex-

tracted from Health Insurance Fund Treatment Bills (from 2004), Tartu University Hospital (from 2008) and North Estonia Medical Center (from 2005) electronic data registries [102].

We overlapped these heterogeneous sources of data to characterise relationships between genetic polymorphisms and ADEs among individuals with purchases of specific drug prescriptions. First, we described the genomic variation of pharmacogenetic VIP genes in Estonian population and in context with other genomic regions. Our second set of analyses aimed to replicate variant-drug ADE associated with levels of evidence 1A-2B in the PharmGKB database. The designations of alleles in CYP2D6 and HLA-B genes were handled using special purpose allele calling tools. Thirdly, we tested non-synonomous pharmacogenetic variation detected in VIP genes in gene-drug pairs that had been previously implicated in PharmGKB on any evidence level (evidence 1-4). For these, we also tested conditional independence of previously associated variants in test genes. Fourth, we conducted a genome-wide association study for adverse drug effects with drugs of more than 1000 prescriptions. For this, all variants with allele frequency greater than 1% were included. All the association tests above were evaluated using logistic regression of cases and controls that adjusted for body-mass index, 4 principal components, sex, age and genotyping platform. Findings were selected for follow-up based on p-value, variant frequency greater than 5% and subjective visualisations of the loci and biological plausability based on a variety of databases. Next, we re-evaluated these SNVs using newly genotyped individuals with associated drug prescriptions of the Estonian Genome Center data that were not available in the preliminary GWAS analysis. Up to 500 cases with an ICD-10 code for the most significant ADE subgroup were included to combine for a total of 1000 individuals. In case of fewer available cases, we used maximally three controls per one case. Replication findings were declared significant based on Bonferroni corrected p-values.

This study was the first to combine EHR and WGS data for ADE population scale investigation. We identified 1,314 putative high-impact variants in 64 pharmacogenes of which 80.3% were found with lower frequency than 1% and 20.3% were novel population-specific variants (Ref III, Table 1). Around 3% of the variants were predicted as loss-of-function (LoF) (Supplementary Table 6). We found that 32.5% of the participants carried at least one loss-of-function variant in a pharmacogene, with 3.5% of the individuals being homozygous for at least one inactivated gene. However, none of the LoF variants could be detected as directly associated with drug related ADEs in our sample. Next, we selected 337 previously described high-confidence associations in 64 pharmacogenes of which we were able 37 associations satisfied the condition of at least 500 samples (Ref III, Supplementary Table 7). The CYP2D6*6 allele was associated with higher incidence of ADEs in patients that had taken tramadol or amitriptyline (p-value $<$ 0.05). Additionally, we discovered nine independent non-synonomous variants in genes predominantly associated to drug related ADEs by other variants in the

same genes (Ref III, Table 2). In the GWAS analysis of 43 phenotypes (Supplementary Table 3), we filtered down candidate variants (Ref III, Supplementary Table 4, Supplementary Table 8) to five promising novel gene-drug associations. However, only the association between a CTNNA3 intronic variant and myopathy related ADEs among individuals taking oxicam was replicated in an extended cohort of 706 individuals from the Estonian Biobank (Ref III, Fig. 3, Supplementary Table 5).

In summary, we used data from Estonian Biobank to use the cohort in identifying novel and previously reported pharmacogenetic associations. We were able to identify and externally evaluate a novel association between a single nucleotide polymorphism in the CTNNA3 gene and higher prevalence of drug induced myositis among users of oxicam class drugs. This study provided further knowledge about the prevalence of population-based variants, including loss of function type of variants and showed that population-based cohorts can be used for pharmacogenetic association studies. As the co-lead author of this study, my contributions included performing most of the the analyses, coming up with the relevant research questions and significant participation in writing the manuscript.

4. AUTOMATED REGIONAL VISUALISATIONS OF GENOME WIDE ASSOCIATION STUDY RESULTS

Our previous population-based pharmacogenomics study relied on testing associations between genomic loci and adverse drug effect status. The number of such tests is large and due to pragmatic reasons only a limited number of associations can be selected for further evaluation. One very common step in assessing GWAS results is to visualise genomic regions using the association p-values and genomic positions (Manhattan plots). Regional visualisation of significant loci is commonly used as an instrument to find regions that may be of interest for further investigation [159]. Even though the visual plot does not provide any information about the biological function then several geometric properties have come to signal interestingness. Some desired properties are presence of SNVs with low p-values, a relatively larger number of SNVs in linkage disequlibrium and a symmetric peak-like shape. However, there are no gold standards for selection and the inspection process remains a largely subjective and time consuming endeavour. Manual inspection means that investigators can only focus on a subset of significant regions for further analyses downstream of GWAS. This becomes particularly challenging when hundreds of phenotypes are available for analysis.

To our knowledge, there are no such tools that automate the detection of regions of high potential interest. Therefore, to decrease the time evaluators spend on this laborious stage, we developed two tools in Ref IV - Manhattan Harvester and Cropper - that allow automated handling of GWAS summary statistics for viewing, zooming and cropping regions of interest. Both tools are available at https://www.geenivaramu.ee/en/tools/.

4.1. Manhattan Harvester and Cropper (Ref IV)

A considerable portion of knowledge discovery in genetics relies on GWAS. A recent push directs genetics towards developing translational applications for the clinic in order to challenge the status quo of GWAS based genetics research [23] but the number of GWAS is actually still increasing [199]. The core methodology of correlative associations in GWAS is still a highly effective way for communicating findings that are interpretable and easily convertible for usage in scientific publications [35]. Therefore, the increasing sample sizes, hundreds of phenotypes that become available through electronic health records, and technological complexity still requires development of scalable methods to analyse and extract GWAS output effectively. Common aims are to elaborate causal variants, to identify targets for biological validation and to predict functionality [77].

Our tools Manhattan Harvester and Cropper automate the detection and ranking of genomic peak regions based on GWAS summary data. This aims to emulate the quality assessment of human evaluators and aims to considerably reduce time

spent on manual visual inspection of genomic regions around tag-SNVs.

Manhattan Harvester applies several techniques to identify peak areas and assign border positions. Its first aim is to find peak borders for any SNVs in GWAS summary statistics with p-values less than 0.001. Each SNV in a genomic region is characterised by a p-value and its chromosomal position. These are used as the basis to perform a series of transformations that allow heuristically finding distinct genomic regions with an arbitrary number of SNVs that form a peak area. Peak border detection methodology was developed by Toomas Haller, the lead author of Ref IV. First, the SNV p-values are smoothed using linear regression on five SNVs, the two closest flanking SNVs in both directions and the smoothed SNV. The original p-value is replaced with the prediction on the middle SNV. Next, height-based compression transforms base-pair distances by bringing SNVs with smaller p-values closer together after which the positions of SNVs are reassigned to correspond to middle distance between their flanking SNVs. The final step identifies final peak borders using vector fragmentation by sequentially creating chunks containing SNVs by moving from largest inter-SNV distances to the smallest until a stopping criteria that defines a sufficiently dense cluster of SNVs. Two parameters are computed for each chunk $i = 1 \dots n$. First stopping parameter as $stop1_i = \frac{max(G_i)}{mean(G_i)}$ and $stop2_i = \frac{max(G_i)}{mean(G_{i+1})}$, where $maxG_i$ is the maximal inter-point gap size and $meanG_i$ is the mean inter-point gap size. Optimal chunk was chosen to be either the maximal value $stop1_i$ or in cases where $stop1_i - stop2_i > 2$, it was selected as $stop2_i$. After a successful run the data within the detected area is removed and the iteration continues with the next round of vector fragmentation to identify additional peaks (Ref IV, Fig. 1).

We used a publicly available metabolite GWAS dataset result to generate visualisations for tests and evaluations [91]. Automatic peak border assignment was evaluated by a single expert on 100 randomly chosen peaks. Additionally, all identified peak areas were quantified using 16 parameters that characterise the peak. To develop the GQS we extracted 277 Manhattan plots with Cropper and asked 20 experts from University of Tartu, knowledgeable in GWAS, to grade these peaks on a 5 point scale. These scores were used as data for a mixed-effects proportional odds model that assumes an ordinal response and accounted for expert specific effects. We applied step-wise model development to identify which of the 16 parameters explain variation in expert scores. The aim was to minimise the average mean square error of model predictions by using a five-fold out-of-sample cross validation. Final model parameters were re-estimated on a full dataset to assess correspondence between the means of 20 evaluator scores and expected values of model based score predictions: $\mathbf{E}(score) = \sum_{i=1}^{5} P(score = i) * i$. Lastly, Cropper and Manhattan Harvester were also subjected to qualitative evaluation in terms of their processing speed, ease of usage and peak identification quality.

Manhattan Harvester and Cropper were implemented in C++ and made avail-

able for downloads from www.geenivaramu.ee/en/tools. A summary statistics file with 560,000 rows was analysed in 3.07 seconds. With current sequencing technologies, genome wide analyses filtered for allele frequencies of more than 1% would not exceed 10,000,000 variants so computational speed is not likely to become a limiting factor (Ref IV, Table 1). In terms of border assignment quality, border-points on test peaks were agreed in 97% of cases as coinciding between the expert and Manhattan Harvester. A single peak was mis-assigned in terms of the width of the base. The other errors were caused by Manhattan Harvester extracting a sub-peak from the complete. Also the expert scores were never more than one unit different from the model prediction.

No between-experts scores correlated less than 0.5 with most experts' pairwise correlations between 0.7 and 0.8 (Ref IV, Fig. 5). We were able to include two parameters that yielded improvements to the null mixed-effect proportional odds model. These were the *log max p-value* in the peak region and also the *best slope*, the largest coefficient of the regression slope for any two points in the peak region. Final score predictions from Manhattan Harvester and mean expert scores correlated with r=0.88 (Ref IV, Fig. 6) and mean squared error of the model was 0.92.

To our knowledge, Manhattan Harvester and Cropper, is the first published tool that allows the automated assessment of Manhattan plots created based on GWAS summary statistics (Ref IV). Visual plots often accompany published GWAS studies to show that a genomic region of significance includes many loci with a distinctly separable "peak". A region with more significant and closely located variants is commonly more likely to be selected for further evaluation. Manhattan Harvester uses summary statistics for automatic peak detection and scores the region by emulating the assessment of human reviewers. This decreases analysis-time and opens avenues for more exploration studies on hundreds of phenotypes simultaneously. We showed that the predicted results correlate well with expert evaluations. My contribution to this paper was the development of general quality scores. I took part in expert evaluator grade collection, chose the model and performed the model fitting and secondary analyses. I also wrote the corresponding section of the manuscript.

CONCLUSIONS

Precision medicine applies genomic and other biological and health data to develop novel methods and approaches to treatment provision, administration and monitoring. Real-world applications rely on integration of methods, tools, clinically interpretable insights and communication channels. The work presented in this thesis covers several different stages in the application of precision methods for provision and monitoring of pharmacotherapy and the discovery of novel genomic associations.

To demonstrate specific examples of precision medicine that do not rely on genomic data, we developed a therapeutic drug monitoring tool, DosOpt. It uses individual concentrations to optimise for drug doses that maximise therapeutic target attainment for vancomycin in neonates (Ref I). Vancomycin has a high between-subject variability so pharmacokinetic models provide a good base for use in TDM. Our study showed that dosing guided by information from on-going treatment can help improve therapeutic target attainment rates in clinical settings with very little input data. This applies also when population priors are not in perfect alignment with the test population which shows that developing and introducing precision dosing methods in TDM to clinical practice has great potential in increasing the attainment rates in desired therapeutic windows. Still, transition from proof-of-concept application to clinical practice requires validation so our future aim is to prospectively validate the tool by neonates receiving vancomycin therapy in Tartu University Hospital and Tallinn Children's Hospital.

There is a serious generalisability issue to be considered when applying pharmacokinetic models outside the model training population. Even though non-conclusive evidence exists on the desirable target attainment estimates for vancomycin administration [143, 164], we aimed to retrospectively externally evaluate several population priors within DosOpt to estimate changes in predictive performance that are dependent on the underlying model (Ref II). Notably, concentration predictions improved in our studies with all PK priors given that individual concentrations data were accounted for. We claim that for Cp_{trough} target window 10-15 mg/L, inclusion of individual concentrations could result in an improvement to around 40% from around 30% attainment based on population model predictions. Our study indicated that the choice of the population prior has a considerable effect on the target attainment but different PK models may still have separate use cases depending on the proximity of the population where the underlying model was developed and the target population, and the overlap of covariates between the PK model and available patient data. Some of the benefits of the precision dosing approach include less subjectivity and added automation of the decision making process through quantitative methods. Potential complications are in the technical complexity and organisational overheads of setting up and managing such systems.

Our pharmacogenetics study (Ref III) showed that combination of population-

based genomic data linked with other independent sources can aid in pharmacogenomic association discovery. Unlike targeted studies, population-based studies identify markers outside pre-specified target regions and pathways. These types of approaches are likely to become more frequent as comprehensive approaches to management and collection of health data are improving the data quality and access. Improving cost-efficiency of sequencing-based technologies and imputation panels also democratises the use of genotype data. Hypothesis free combination of heterogeneous sources of complex data guides research in novel directions and uncovers unexpected findings [131]. Discoveries can be followed up with mechanistic evaluations and translated to clinically applicable guidelines [179]. This is important as genomic medicine has much potential in improving pharmacotherapy. Population-based studies can be successfully applied in elucidating relationships between drug responses and genetic variation based on electronic health records and genotype information.

Association studies in genomics commonly include visualisation of significances in a genomic region. The shape of a regional locus plot is dependent on the underlying biological aspects but does not give information about the peaks biological functionality. Instead their "interestingness" corresponds to subjective preferences. In Ref IV we automated this evaluation and selection process by developing software that mimics the preferences of human evaluators based on visually observable characteristics. These do not use biological background information. In the end, the human evaluators still make the decision on which peaks to follow up on. Thus may be dependent on other aspects that Manhattan Harvester does not include such as any biological mechanisms or attributes, previously known information about the peak, the capacity of an evaluator to follow up based on available resources. Even though Manhattan Harvester incorporates p-values in the general quality score evaluation as experts do, the initial peak discovery relies instead on a set of data manipulation techniques that are able to filter for genomic regions based on quality controlled GWAS summary statistics inputs. As such Manhattan Harvester and Cropper integrate into a standard GWAS pipeline and make standard analyses more time efficient.

Related work on the topic of drug pharmacokinetics modelling [65, 66, 147, 148, 188]. and analysis of next-generation sequencing data [62, 160, 161] by the author includes several other publications not included in this thesis.

This thesis aims to illustrate that precision medicine emerges from different sources. Firstly, bedside monitoring of pharmacotherapies can be made more precise when drug administration adjustments are directly informed by patient responses. Secondly, drug related genetic markers can improve pharmacotherapies with finer individualisation of drug prescriptions. Analysis and identification of such targets can be more effective with computational approaches that reduce analysis time and help assessment of more potential candidate associations. Fulfilment of these promises relies in the continued collaboration of medical, scientific and legislative partners.

BIBLIOGRAPHY

[1] Leon Aarons. Software for population pharmacokinetics and pharmacodynamics. *Clinical Pharmacokinetics*, 36(4):255–264, 1999.

[2] G. P. Aithal, C. P. Day, P. J. Kesteven, and A. K. Daly. Association of polymorphisms in the cytochrome P450 CYP2c9 with warfarin dose requirement and risk of bleeding complications. *Lancet*, 353(9154):717–719, February 1999.

[3] Karel Allegaert, Brian J Anderson, John N van den Anker, Sophie Vanhaesebrouck, and Francis de Zegher. Renal drug clearance in preterm neonates: relation to prenatal growth. *Therapeutic Drug Monitoring*, 29(3):284–291, 2007.

[4] FM Alsaleh, J Lemay, RR Al Dhafeeri, S AlAjmi, EA Abahussain, and T Bayoud. Adverse drug reaction reporting among physicians working in private and government hospitals in kuwait. *Saudi Pharmaceutical Journal*, 25(8):1184–1193, 2017.

[5] Akram Alyass, Michelle Turcotte, and David Meyre. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Medical Genomics*, 8(1):33, 2015.

[6] Paul G Ambrose, Sujata M Bhavnani, Christopher M Rubino, Arnold Louie, Tawanda Gumbo, Alan Forrest, and George L Drusano. Pharmacokinetics-pharmacodynamics of antimicrobial therapy: it's not just for mice anymore. *Clinical Infectious Diseases*, 44(1):79–86, 2007.

[7] Brian J Anderson, Karel Allegaert, John N Van den Anker, Veerle Cossey, and Nicholas HG Holford. Vancomycin pharmacokinetics in preterm neonates and the prediction of adult clearance. *British Journal of Clinical Pharmacology*, 63(1):75–84, 2007.

[8] Brian J Anderson and Nicholas HG Holford. Mechanism-based concepts of size and maturity in pharmacokinetics. *Annual Review of Pharmacology and Toxicology*, 48:303–332, 2008.

[9] Samuel J. Aronson and Heidi L. Rehm. Building the foundation for genomics in precision medicine. *Nature*, 526(7573):336–342, October 2015.

[10] Euan A Ashley. The precision medicine initiative: a new national effort. *JAMA*, 313(21):2119–2120, 2015.

[11] Haleh Ayatollahi, Nader Mirani, and Hamid Haghani. Electronic health records: what are the most important barriers? *Perspectives in Health Information Management*, 11(Fall), 2014.

[12] Arnaud Belard, Timothy Buchman, Jonathan Forsberg, Benjamin K Potter, Christopher J Dente, Allan Kirk, and Eric Elster. Precision diagnosis: a view of the clinical decision support systems (cdss) landscape through

the lens of critical care. *Journal of Clinical Monitoring and Computing*, 31(2):261–271, 2017.

[13] Leslie Z Benet and Jere E Goyan. Bioequivalence and narrow therapeutic index drugs. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, 15(4):433–440, 1995.

[14] James L. Bernat. Ethical and quality pitfalls in electronic health records. *Neurology*, 80(11):1057–1061, March 2013.

[15] Dianne C Berry, Peter Knapp, and DK Raynor. Provision of information about drug side-effects to patients. *The Lancet*, 359(9309):853–854, 2002.

[16] Cinnamon S Bloss, Nicholas J Schork, and Eric J Topol. Direct-to-consumer pharmacogenomic testing is associated with increased physician utilisation. *Journal of Medical Genetics*, 51(2):83–89, 2014.

[17] Karl Brendel, Céline Dartois, Emmanuelle Comets, Annabelle Lemenuel-Diot, Christian Laveille, Brigitte Tranchand, Pascal Girard, Céline M Laffont, and France Mentré. Are population pharmacokinetic and/or pharmacodynamic models adequately evaluated? *Clinical Pharmacokinetics*, 46(3):221–234, 2007.

[18] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS Computational Biology*, 8(12):e1002822, 2012.

[19] Kelly Caine and Rima Hanania. Patients want granular privacy control over health information in electronic medical records. *Journal of the American Medical Informatics Association*, 20(1):7–15, 2012.

[20] Alison A Carter, Sara E Rosenbaum, and Michael N Dudley. Review of methods in population pharmacokinetics. *Clinical Research and Regulatory Affairs*, 12(1):1–21, 1995.

[21] Winston Chang, Joe Chang, JJ Allaire, Yihui Xie, and Jonathan McPherson. *shiny: Web Application Framework for R*, 2017. R package version 1.0.0.

[22] Yewei Chen, Dan Wu, Min Dong, Yiqing Zhu, Jinmiao Lu, Xiaoxia Li, Chao Chen, and Zhiping Li. Population pharmacokinetics of vancomycin and auc-guided dosing in chinese neonates and young infants. *European Journal of Clinical Pharmacology*, 74(7):921–930, 2018.

[23] Francis S Collins. Reengineering translational science: the time is right. *Science Translational Medicine*, 3(90):90cm17–90cm17, 2011.

[24] Susannah L Collins, Daniel F Carr, and Munir Pirmohamed. Advances in the pharmacogenomics of adverse drug reactions. *Drug Safety*, 39(1):15–27, 2016.

[25] 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[26] ICPerMed International Consortium. Action plan. actionable research and support activities identified by the international consortium for personalised medicine. 2017, 2017.

[27] The International HapMap Consortium. The international hapmap project. *Nature*, 426(6968):789, 2003.

[28] The International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860, 2001.

[29] The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748, 2009.

[30] Martin R Cowie, Juuso I Blomster, Lesley H Curtis, Sylvie Duclaux, Ian Ford, Fleur Fritz, Samantha Goldman, Salim Janmohamed, Jörg Kreuzer, Mark Leenay, et al. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1):1–9, 2017.

[31] William A Craig. Antimicrobial resistance issues of the future. *Diagnostic Microbiology and Infectious Disease*, 25(4):213–217, 1996.

[32] William A Craig. Pharmacokinetic/pharmacodynamic parameters: rationale for antibacterial dosing of mice and men. *Clinical infectious diseases*, 26(1):1–10, 1998.

[33] Ann K Daly. Genome-wide association studies in pharmacogenomics. *Nature Reviews Genetics*, 11(4):241, 2010.

[34] AS Darwich, K Ogungbenro, AA Vinks, JR Powell, J-L Reny, Niloufar Marsousi, Youssef Daali, D Fairman, James Cook, LJ Lesko, et al. Why has model-informed precision dosing not yet become common clinical reality? lessons from the past and a roadmap for the future. *Clinical Pharmacology & Therapeutics*, 101(5):646–656, 2017.

[35] Philip L De Jager. The era of GWAS is over–no. *Multiple Sclerosis Journal*, 24(3):258–260, 2018.

[36] Matthew P Doogue and Thomas M Polasek. The ABCD of clinical pharmacokinetics, 2013.

[37] Dawn Dowding, Rebecca Randell, Peter Gardner, Geraldine Fitzpatrick, Patricia Dykes, Jesus Favela, Susan Hamer, Zac Whitewood-Moores, Nicholas Hardiker, Elizabeth Borycki, et al. Dashboards for improving patient care: review of the literature. *International Journal of Medical Informatics*, 84(2):87–100, 2015.

[38] Philip Drennan, Matthew Doogue, Sebastiaan J van Hal, and Paul Chin. Bayesian therapeutic drug monitoring software: past, present and future, 2018.

[39] Anne Dubois, Julie Bertrand, and France Mentré. Mathematical expressions of the pharmacokinetic and pharmacodynamic models implemented in the PFIM software. *Université Paris Diderot and INSERM*, 2011.

[40] Henry M Dunnenberger, Kristine R Crews, James M Hoffman, Kelly E Caudle, Ulrich Broeckel, Scott C Howard, Robert J Hunkler, Teri E Klein, William E Evans, and Mary V Relling. Preemptive clinical pharmacogenetics implementation: current programs in five us medical centers. *Annual Review of Pharmacology and Toxicology*, 55:89–106, 2015.

[41] Henry M Dunnenberger and Janardan D Khandekar. Value of personalized medicine. *JAMA*, 315(6):612–613, 2016.

[42] Victor J Dzau, Geoffrey S Ginsburg, Karen Van Nuys, David Agus, and Dana Goldman. Aligning incentives to fulfill the promise of personalized medicine. *Lancet (London, England)*, 385(9982):2118, 2015.

[43] Kelly E Caudle, Teri E Klein, James M Hoffman, Daniel J Muller, Michelle Whirl-Carrillo, Li Gong, Ellen M McDonagh, Katrin Sangkuhl, Caroline F Thorn, Matthias Schwab, et al. Incorporation of pharmacogenomics into routine clinical practice: the clinical pharmacogenetics implementation consortium (cpic) guideline development process. *Current Drug Metabolism*, 15(2):209–217, 2014.

[44] I Ralph Edwards and Jeffrey K Aronson. Adverse drug reactions: definitions, diagnosis, and management. *The Lancet*, 356(9237):1255–1259, 2000.

[45] Stacey L Edwards, Jonathan Beesley, Juliet D French, and Alison M Dunning. Beyond GWASs: illuminating the dark road from association to function. *The American Journal of Human Genetics*, 93(5):779–797, 2013.

[46] Michel Eichelbaum, Russ B Altman, Mark Ratain, and Teri E Klein. New feature: pathways and important genes from PharmGKB. *Pharmacogenetics and Genomics*, 19(6):403, 2009.

[47] GA Ellard. Variations between individuals and populations in the acetylation of isoniazid and its significance for the treatment of pulmonary tuberculosis. *Clinical Pharmacology & Therapeutics*, 19(5part2):610–625, 1976.

[48] Mary HH Ensom, Thomas KH Chang, and Payal Patel. Pharmacogenetics. *Clinical Pharmacokinetics*, 40(11):783–802, 2001.

[49] Alvan R Feinstein and Ralph I Horwitz. Problems in the 'evidence' of 'evidence-based medicine'. *The American Journal of Medicine*, 103(6):529–535, 1997.

[50] Jose Luis Fernandez-Aleman, Inmaculada Carrion Senor, Pedro angel Oliver Lozoya, and Ambrosio Toval. Security and privacy in electronic health records: a systematic literature review. *Journal of Biomedical Informatics*, 46(3):541–562, June 2013.

[51] Kenneth R. Foster, Robert Koprowski, and Joseph D. Skufca. Machine learning, medical diagnosis, and biomedical engineering research - commentary. *Biomed Eng Online*, 13:94, July 2014.

[52] Adam Frymoyer, Adam L Hersh, Mohammed H El-Komy, Shabnam Gaskari, Felice Su, David R Drover, and Krisa Van Meurs. Association between vancomycin trough concentration and area under the concentration-time curve in neonates. *Antimicrobial Agents and Chemotherapy*, 58(11):6454–6461, 2014.

[53] Johan Gabrielsson and Daniel Weiner. Non-compartmental analysis. In *Computational toxicology*, pages 377–389. Springer, 2012.

[54] Andrea Gaedigk, Magnus Ingelman-Sundberg, Neil A Miller, J Steven Leeder, Michelle Whirl-Carrillo, Teri E Klein, and PharmVar Steering Committee. The pharmacogene variation (pharmvar) consortium: incorporation of the human cytochrome p450 (cyp) allele nomenclature database. *Clinical Pharmacology & Therapeutics*, 103(3):399–401, 2018.

[55] Michael D Gallagher and Alice S Chen-Plotkin. The post-GWAS era: from association to function. *The American Journal of Human Genetics*, 102(5):717–730, 2018.

[56] Karin Garrety, Ian McLoughlin, Rob Wilson, Gregor Zelle, and Mike Martin. National electronic health records and the digital disruption of moral orders. *Social Science & Medicine*, 101:70–77, 2014.

[57] Leonard E Gerlowski and Rakesh K Jain. Physiologically based pharmacokinetic modeling: principles and applications. *Journal of Pharmaceutical Sciences*, 72(10):1103–1127, 1983.

[58] RA Ghiculescu. Therapeutic drug monitoring: which drugs, why, when and how to do it. *Australian Prescriber*, 31(2):42–4, 2008.

[59] Daniel Gonzalez, Gauri G Rao, Stacy C Bailey, Kim LR Brouwer, Yanguang Cao, Daniel J Crona, Angela DM Kashuba, Craig R Lee, Kathryn Morbitzer, J Herbert Patterson, et al. Precision dosing: public health need, proposed framework, and anticipated impact. *Clinical and Translational Science*, 10(6):443–454, 2017.

[60] David J Greenblatt and Jan Koch-Weser. Clinical pharmacokinetics. *New England Journal of Medicine*, 293(14):702–705, 1975.

[61] Michael Grunstein and David S Hogness. Colony hybridization: a method for the isolation of cloned dnas that contain a specific gene. *Proceedings of the National Academy of Sciences*, 72(10):3961–3965, 1975.

[62] Mithu Guha, Mario Saare, Julia Maslovskaja, Kai Kisand, Ingrid Liiv, Uku Haljasorg, Tõnis Tasa, Andres Metspalu, Lili Milani, and Pärt Peterson. DNA breaks and chromatin structural changes enhance the transcription of autoimmune regulator target genes. *The Journal of Biological Chemistry*, 292(16):6–542, 2017.

[63] Matthew Haag. FamilytreeDNA admits to sharing genetic data with FBI. *New York Times*, Feb 2019.

[64] Toomas Haller, Tõnis Tasa, and Andres Metspalu. Manhattan Harvester and Cropper: a system for GWAS peak detection. *BMC Bioinformatics*, 20(1):22, 2019.

[65] Maarja Hallik, Mari-Liis Ilmoja, Tõnis Tasa, Joseph F Standing, Kalev Takkis, Ruta Veigure, Karin Kipper, Tiiu Jalas, Maila Raidmäe, Karin Uibo, et al. Population pharmacokinetics and dosing of milrinone after patent ductus arteriosus ligation in preterm infants. *Pediatric Critical Care Medicine*, 2019.

[66] Maarja Hallik, Tõnis Tasa, Joel Starkopf, and Tuuli Metsvaht. Dosing of milrinone in preterm neonates to prevent postligation cardiac syndrome: Simulation study suggests need for bolus infusion. *Neonatology*, 111(1):8–11, 2017.

[67] M Hansen, LL Christrup, JO Jarløv, JP Kampmann, and J Bonde. Gentamicin dosing in critically ill patients. *Acta Anaesthesiologica Scandinavica*, 45(6):734–740, 2001.

[68] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, 2017.

[69] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *International Journal of Medical Informatics*, 77(5):291–304, 2008.

[70] Lorna Hazell and Saad AW Shakir. Under-reporting of adverse drug reactions. *Drug Safety*, 29(5):385–396, 2006.

[71] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Research*, 30(1):163–165, 2002.

[72] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[73] Y Hiraki, T Onga, A Mizoguchi, and Y Tsuji. Investigation of the prediction accuracy of vancomycin concentrations determined by patient-specific parameters as estimated by Bayesian analysis. *Journal of Clinical Pharmacy and Therapeutics*, 35(5):527–532, 2010.

[74] Joseph Hong, Lynne C. Krop, Tracy Johns, and Manjunath P. Pai. Individualized vancomycin dosing in obese patients: a two-sample measurement approach improves target attainment. *Pharmacotherapy*, 35(5):455–463, May 2015.

[75] William W Hope, Michael VanGuilder, J Peter Donnelly, Nicole MA Blijlevens, Roger JM Brüggemann, Roger W Jelliffe, and Michael N Neely. Software for dosage individualization of voriconazole for immunocompromised patients. *Antimicrobial Agents and Chemotherapy*, 57(4):1888–1894, 2013.

[76] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[77] Lin Hou and Hongyu Zhao. A review of post-GWAS prioritization approaches. *Frontiers in Genetics*, 4:280, 2013.

[78] Paul R Ingram, David C Lye, Paul A Tambyah, Wei P Goh, Vincent H Tam, and Dale A Fisher. Risk factors for nephrotoxicity associated with continuous vancomycin infusion in outpatient parenteral antibiotic therapy. *Journal of Antimicrobial Chemotherapy*, 62(1):168–171, 2008.

[79] Evelyne Jacqz-Aigrain, Stephanie Leroux, Wei Zhao, John N. van den Anker, and Mike Sharland. How to use vancomycin optimally in neonates: remaining questions. *Expert Reviews in Clinical Pharmacology*, 8(5):635–648, 2015.

[80] Evelyne Jacqz-Aigrain, Wei Zhao, Mike Sharland, and John N van den Anker. Use of antibacterial agents in the neonate: 50 years of experience with vancomycin administration. In *Seminars in Fetal and Neonatal Medicine*, volume 18, pages 28–34. Elsevier, 2013.

[81] Sunil S Jambhekar and Philip J Breen. *Basic pharmacokinetics*. Pharmaceutical Press London, UK:, 2012.

[82] Wojciech Jawień. Searching for an optimal AUC estimation method: a never-ending task? *Journal of Pharmacokinetics and Pharmacodynamics*, 41(6):655–673, 2014.

[83] Yuan Ji, Jennifer M Skierka, Joseph H Blommel, Brenda E Moore, Douglas L VanCuyk, Jamie K Bruflat, Lisa M Peterson, Tamra L Veldhuizen, Numrah Fadra, Sandra E Peterson, et al. Preemptive pharmacogenomic testing for precision medicine: a comprehensive analysis of five actionable pharmacogenomic genes using next-generation DNA sequencing and a customized CYP2D6 genotyping cascade. *The Journal of Molecular Diagnostics*, 18(3):438–445, 2016.

[84] Michael J Joyner and Nigel Paneth. Seven questions for personalized medicine. *JAMA*, 314(10):999–1000, 2015.

[85] Lisa M Kalisch, Gillian E Caughey, Elizabeth E Roughead, and Andrew L Gilbert. The prescribing cascade. *Australian Prescriber*, 34(6):162–6, 2011.

[86] Werner Kalow, Urs B Meyer, and Rachel F Tyndale. *Pharmacogenomics*. CRC Press, 2001.

[87] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348, 2010.

[88] Ju-Seop Kang and Min-Ho Lee. Overview of therapeutic drug monitoring. *The Korean Journal of Internal Medicine*, 24(1):1, 2009.

[89] K Karczewski and L Francioli. The genome aggregation database (gnomad). *MacArthur Lab*, 2017.

[90] Jane Kaye, Catherine Heeney, Naomi Hawkins, Jantina De Vries, and Paula Boddington. Data sharing in genomics re shaping scientific practice. *Nature Reviews Genetics*, 10(5):331, 2009.

[91] Johannes Kettunen, Ayşe Demirkan, Peter Würtz, Harmen HM Draisma, Toomas Haller, Rajesh Rawal, Anika Vaarhorst, Antti J Kangas, Leo-Pekka Lyytikäinen, Matti Pirinen, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nature Communications*, 7:11122, 2016.

[92] Amit V Khera, Mark Chaffin, Krishna G Aragam, Mary E Haas, Carolina Roselli, Seung Hoan Choi, Pradeep Natarajan, Eric S Lander, Steven A Lubitz, Patrick T Ellinor, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9):1219, 2018.

[93] Toshimi Kimura, Keisuke Sunakawa, Nobuo Matsuura, Hiroaki Kubo, Shigehiko Shimada, and Kazuo Yago. Population pharmacokinetics of arbekacin, vancomycin, and panipenem in neonates. *Antimicrobial Agents and Chemotherapy*, 48(4):1159–1167, April 2004.

[94] Omayma A Kishk, Allison B Lardieri, Emily L Heil, and Jill A Morgan. Vancomycin auc/mic and corresponding troughs in a pediatric population. *The Journal of Pediatric Pharmacology and Therapeutics*, 22(1):41–47, 2017.

[95] Frank Kloprogge, Louise F Hill, John Booth, Nigel Klein, Adam Irwin, Garth Dixon, and Joseph F Standing. Revising paediatric vancomycin dosing accounting for nephrotoxicity in a pharmacokinetic-pharmacodynamic model. *Antimicrobial Agents and Chemotherapy*, pages AAC–00067, 2019.

[96] John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.

[97] Adam Lavertu, Greg McInnes, Roxana Daneshjou, Michelle Whirl-Carrillo, Teri E Klein, and Russ B Altman. Pharmacogenomics and big genomic data: from lab to clinic and back again. *Human Molecular Genetics*, 27(R1):R72–R78, 2018.

[98] Jennifer Le, John S Bradley, William Murray, Gale L Romanowski, Tu T Tran, Natalie Nguyen, Susan Cho, Stephanie Natale, Ivilynn Bui, Tri M Tran, et al. Improved vancomycin dosing in children using area-under-the-curve exposure. *The Pediatric Infectious Disease Journal*, 32(4):e155, 2013.

[99] JW Lee, F Aminkeng, AP Bhavsar, K Shaw, BC Carleton, MR Hayden, and CJD Ross. The emerging era of pharmacogenomics: current successes, future potential, and challenges. *Clinical Genetics*, 86(1):21–28, 2014.

[100] Peter M Lee. *Bayesian statistics*. Arnold Publication, 1997.

[101] Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

[102] Liis Leitsalu, Helene Alavere, Mari-Liis Tammesoo, Erkki Leego, and Andres Metspalu. Linking a population biobank with national health registries-the Estonian experience. *Journal of Personalized Medicine*, 5(2):96–106, April 2015.

[103] Liis Leitsalu, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, Pauline C Ng, Reedik Mägi, Lili Milani, et al. Cohort profile: Estonian biobank of the Estonian Genome Center, university of tartu. *International Journal of Epidemiology*, 44(4):1137–1147, 2014.

[104] Liis Leitsalu and Andres Metspalu. From biobanking to precision medicine: The estonian experience. In *Genomic and Precision Medicine*, pages 119–129. Elsevier, 2017.

[105] Amy A Lemke, Wendy A Wolf, Jennifer Hebert-Beirne, and Maureen E Smith. Public and biobank participant attitudes toward genetic research participation and data sharing. *Public Health Genomics*, 13(6):368–377, 2010.

[106] Stéphanie Leroux, Evelyne Jacqz-Aigrain, Valérie Biran, Emmanuel Lopez, Doriane Madeleneau, Camille Wallon, Elodie Zana-Taïeb, Anne-Laure Virlouvet, Stéphane Rioualen, and Wei Zhao. Clinical utility and safety of a model-based patient-tailored dose of vancomycin in neonates. *Antimicrobial Agents and Chemotherapy*, 60(4):2039–2042, 2016.

[107] Jodi M. Lestner, Louise F. Hill, Paul T. Heath, and Mike Sharland. Vancomycin toxicity in neonates: a review of the evidence. *Current Opinion in Infectious Diseases*, 29(3):237–247, 2016.

[108] Matthew E Levison and Julie H Levison. Pharmacokinetics and pharmacodynamics of antibacterial agents. *Infectious Disease Clinics*, 23(4):791–815, 2009.

[109] Yves Levy. Genomic medicine 2025: France in the race for precision medicine. *Lancet*, 388(10062):2872, 2016.

[110] Mingfeng Li, Gabriel Santpere, Yuka Imamura Kawasawa, Oleg V Evgrafov, Forrest O Gulden, Sirisha Pochareddy, Susan M Sunkin, Zhen Li, Yurae Shin, Ying Zhu, et al. Integrative functional genomic analysis of human brain development and neuropsychiatric risks. *Science*, 362(6420):eaat7615, 2018.

[111] Yoke-Lin Lo, Johan G. C. van Hasselt, Siow-Chin Heng, Chin-Theam Lim, Toong-Chow Lee, and Bruce G. Charles. Population pharmacokinetics of vancomycin in premature Malaysian neonates: identification of predictors for dosing determination. *Antimicrobial Agents and Chemotherapy*, 54(6):2626–2632, June 2010.

[112] Thomas P. Lodise, Nimish Patel, Ben M. Lomaestro, Keith A. Rodvold, and George L. Drusano. Relationship between initial vancomycin concentration-time profile and nephrotoxicity among hospitalized patients. *Clinical Infectious Diseases*, 49(4):507–514, August 2009.

[113] D. M. Louie. HI Plavix Lawsuit news-release. 2014.

[114] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2016.

[115] Teri A Manolio. Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics*, 14(8):549, 2013.

[116] Lisanne En Manson, Cathelijne H. van der Wouden, Jesse J. Swen, and Henk-Jan Guchelaar. The Ubiquitous Pharmacogenomics consortium: making effective treatment optimization accessible to every European citizen. *Pharmacogenomics*, 18(11):1041–1045, 2017.

[117] Jun-Jun Mao, Zheng Jiao, Hwi-Yeol Yun, Chen-Yan Zhao, Han-Chao Chen, Xiao-Yan Qiu, and Ming-Kang Zhong. External evaluation of population pharmacokinetic models for ciclosporin in adult renal transplant recipients. *British Journal of Clinical Pharmacology*, 84(1):153–171, 2018.

[118] Joseph C Maranville and Nancy J Cox. Pharmacogenomic variants have larger effect sizes than genetic variants associated with other dichotomous complex traits. *The Pharmacogenomics Journal*, 16(4):388, 2016.

[119] María-Remedios Marques-Minana, Anas Saadeddin, and Jose-Esteban Peris. Population pharmacokinetic analysis of vancomycin in neonates. A new proposal of initial dosage guideline. *British Journal of Clinical Pharmacology*, 70(5):713–720, November 2010.

[120] Amélie Marsot, Audrey Boulamery, Bernard Bruguerolle, and Nicolas Simon. Vancomycin. *Clinical Pharmacokinetics*, 51(1):1–13, 2012.

[121] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly,

Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[122] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics*, 48(10):1279, 2016.

[123] Bernd Meibohm, Stephanie Läer, John C Panetta, and Jeffrey S Barrett. Population pharmacokinetic studies in pediatrics: issues in design and analysis. *The AAPS journal*, 7(2):E475–E487, 2005.

[124] Peter J Meier-Abt, Adrien K Lawrence, Liselotte Selter, Effy Vayena, and Torsten Schwede. The swiss approach to precision medicine. *Swiss Medical Weekly*, 2018.

[125] Michael L Metzker. Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1):31, 2010.

[126] Robert H Miller and Ida Sim. Physicians use of electronic medical records: barriers and solutions. *Health Affairs*, 23(2):116–126, 2004.

[127] Mario Mitt, Mart Kals, Kalle Pärn, Stacey B Gabriel, Eric S Lander, Aarno Palotie, Samuli Ripatti, Andrew P Morris, Andres Metspalu, Tõnu Esko, et al. Improved imputation accuracy of rare and low-frequency variants using population-specific high-coverage wgs-based imputation reference panel. *European Journal of Human Genetics*, 25(7):869, 2017.

[128] Jr. Moellering, Robert C. Vancomycin: A 50-year reassessment. *Clinical Infectious Diseases*, 42(Supplement 1):S3–S4, 01 2006.

[129] Pamela A. Moise-Broder, Alan Forrest, Mary C. Birmingham, and Jerome J. Schentag. Pharmacodynamics of vancomycin and other antimicrobials in patients with Staphylococcus aureus lower respiratory tract infections. *Clinical Pharmacokinetics*, 43(13):925–942, 2004.

[130] J. D. Momper and J. A. Wagner. Therapeutic drug monitoring as a component of personalized medicine: applications in pediatric drug development. *Clinical Pharmacology and Therapeutics*, 95(2):138–140, February 2014.

[131] Alison A Motsinger-Reif, Eric Jorgenson, Mary V Relling, Deanna L Kroetz, Richard Weinshilboum, Nancy J Cox, and Dan M Roden. Genome-wide association studies in pharmacogenomics: successes and lessons. *Pharmacogenetics and Genomics*, 23(8):383, 2013.

[132] DR Mould and Richard Neil Upton. Basic concepts in population modeling, simulation, and model-based drug development-part 2: introduction to pharmacokinetic modeling methods. *CPT: Pharmacometrics & Systems Pharmacology*, 2(4):1–14, 2013.

[133] Johan W Mouton, Marc L van Ogtrop, David Andes, and William A Craig. Use of pharmacodynamic indices to predict efficacy of combination ther-

apy in vivo. *Antimicrobial Agents and Chemotherapy*, 43(10):2473–2478, 1999.

[134] Keith T Muir. Nonlinear least-squares regression analysis in pharmacokinetics: application of a programmable calculator in model parameter estimation. *Computers and Biomedical Research*, 13(4):307–316, 1980.

[135] Travis B Murdoch and Allan S Detsky. The inevitable application of big data to health care. *JAMA*, 309(13):1351–1352, 2013.

[136] John E Murphy, David E Gillespie, and Carol V Bateman. Predictability of vancomycin trough concentrations using seven approaches for estimating pharmacokinetic parameters. *American Journal of Health-System Pharmacy*, 63(23):2365–2370, 2006.

[137] Michael Neely and Roger Jelliffe. Practical, individualized dosing: 21st century therapeutics and the clinical pharmacometrician. *Journal of Clinical Pharmacology*, 50(7):842–847, July 2010.

[138] Michael N. Neely, Gilmer Youn, Brenda Jones, Roger W. Jelliffe, George L. Drusano, Keith A. Rodvold, and Thomas P. Lodise. Are vancomycin trough concentrations adequate for optimal dosing? *Antimicrobial Agents and Chemotherapy*, 58(1):309–316, 2014.

[139] Karen Ng, Vincent H Mabasa, Ivy Chow, and Mary HH Ensom. Systematic review of efficacy, pharmacokinetics, and administration of intraventricular vancomycin in adults. *Neurocritical Care*, 20(1):158–171, 2014.

[140] Robby Nieuwlaat, Stuart J Connolly, Jean A Mackay, Lorraine Weise-Kelly, Tamara Navarro, Nancy L Wilczynski, and R Brian Haynes. Computerized clinical decision support systems for therapeutic drug monitoring and dosing: a decision-maker-researcher partnership systematic review. *Implementation Science*, 6(1):90, 2011.

[141] Elmar Nimmesgern, Indridi Benediktsson, and Irene Norstedt. Personalized medicine in europe. *Clinical and Translational Science*, 10(2):61–63, 2017.

[142] Pål Rasmus Njølstad, Ole Andreas Andreassen, Søren Brunak, Anders D Børglum, Joakim Dillner, Tõnu Esko, Paul W Franks, Nelson Freimer, Leif Groop, Hakon Heimer, et al. Roadmap for a precision-medicine initiative in the nordic region. *Nature Genetics*, page 1, 2019.

[143] Maya O. Nunn, Carmela E. Corallo, Cecile Aubron, Susan Poole, Michael J. Dooley, and Allen C. Cheng. Vancomycin dosing: assessment of time to therapeutic concentration and predictive accuracy of pharmacokinetic modeling software. *Annals of Pharmacotherapy*, 45(6):757–763, June 2011.

[144] OECD. *Fiscal Sustainability of Health Systems*. 2015.

[145] C. Oudin, R. Vialet, A. Boulamery, C. Martin, and N. Simon. Vancomycin prescription in neonates and young infants: toward a simplified

dosage. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 96(5):F365–370, September 2011.

[146] Vural Ozdemir, David H Muljono, Tikki Pang, Lynnette R Ferguson, Aresha Manamperi, Sofia Samper, Toshiyuki Someya, Anne Marie Tassé, Shih-Jen Tsai, Hong-Hao Zhou, et al. Asia-Pacific Health 2020 and genomics without borders: co-production of knowledge by science and society partnership for global personalized medicine. *Current Pharmacogenomics and Personalized Medicine*, 9(1):1, 2011.

[147] Helgi Padari, Tuuli Metsvaht, Eva Germovsek, Charlotte I Barker, Karin Kipper, Koit Herodes, Joseph F Standing, Kersti Oselin, Tõnis Tasa, Hiie Soeorg, et al. Pharmacokinetics of penicillin G in preterm and term neonates. *Antimicrobial Agents and Chemotherapy*, pages AAC–02238, 2018.

[148] Helgi Padari, Kersti Oselin, Tõnis Tasa, Tuuli Metsvaht, Krista Lõivukene, and Irja Lutsar. Coagulase negative staphylococcal sepsis in neonates: do we need to adapt vancomycin dose or target? *BMC Pediatrics*, 16(1):206, 2016.

[149] Claudia Pagliari, Don Detmer, and Peter Singleton. Potential of electronic personal health records. *BMJ*, 335(7615):330–333, 2007.

[150] Manjunath P Pai, Michael Neely, Keith A Rodvold, and Thomas P Lodise. Innovative approaches to optimizing the delivery of vancomycin in individual patients. *Advanced Drug Delivery Reviews*, 77:50–57, 2014.

[151] Ju-Hyun Park, Mitchell H Gail, Clarice R Weinberg, Raymond J Carroll, Charles C Chung, Zhaoming Wang, Stephen J Chanock, Joseph F Fraumeni, and Nilanjan Chatterjee. Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences*, 108(44):18026–18031, 2011.

[152] Addepalli Pavani, Shaik Mohammad Naushad, Balraj Alex Stanley, Renganathan Gnanambal Kamakshi, Krishnan Abinaya, Malempati Amaresh Rao, Addepally Uma, and Vijay Kumar Kutala. Mechanistic insights into the effect of CYP2C9*2 and CYP2C9*3 variants on the 7-hydroxylation of warfarin. *Pharmacogenomics*, 16(4):393–400, 2015.

[153] Steve G Peters and Munawwar A Khan. Electronic health records: current and future use. *Journal of Comparative Effectiveness Research*, 3(5):515–522, 2014.

[154] Kathryn A. Phillips, Julie Ann Sakowski, Julia Trosman, Michael P. Douglas, Su-Ying Liang, and Peter Neumann. The economic value of personalized medicine tests: what we know and what we need to know. *Genet. Med.*, 16(3):251–257, March 2014.

[155] Munir Pirmohamed, Sally James, Shaun Meakin, Chris Green, Andrew K. Scott, Thomas J. Walley, Keith Farrar, B. Kevin Park, and Alasdair M.

Breckenridge. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. *BMJ*, 329(7456):15–19, July 2004.

[156] Jodyn Platt and Sharon Kardia. Public trust in health information sharing: implications for biobanking and electronic health record systems. *Journal of Personalized Medicine*, 5(1):3–21, 2015.

[157] Martyn Plummer. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124. Vienna, Austria., 2003.

[158] John Powell, Richard Fitton, and Caroline Fitton. Sharing electronic health records: the patient view. *Journal of Innovation in Health Informatics*, 14(1):55–57, 2006.

[159] Randall J Pruim, Ryan P Welch, Serena Sanna, Tanya M Teslovich, Peter S Chines, Terry P Gliedt, Michael Boehnke, Gonçalo R Abecasis, and Cristen J Willer. Locuszoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–2337, 2010.

[160] Kadri Rekker, Merli Saare, Elo Eriste, Tõnis Tasa, Anne Mari Roost, Viktorija Kukuškina, Kristi Anderson, Külli Samuel, Helle Karro, Andres Salumets, et al. High-throughput mRNA sequencing of stromal cells from endometriomas and endometrium. *Reproduction*, 154(1):93–100, 2017.

[161] Kadri Rekker, Tõnis Tasa, Merli Saare, Külli Samuel, Ülle Kadastik, Helle Karro, Martin Götte, Andres Salumets, and Maire Peters. Differentially-expressed miRNAs in ectopic stromal cells contribute to endometriosis development: The plausible role of miR-139-5p and miR-375. *International Journal of Molecular Sciences*, 19(12), 2018.

[162] Mary V Relling and William E Evans. Pharmacogenomics in the clinic. *Nature*, 526(7573):343, 2015.

[163] MV Relling and TE Klein. Cpic: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. *Clinical Pharmacology & Therapeutics*, 89(3):464–467, 2011.

[164] Theresa Ringenberg, Christine Robinson, Rachel Meyers, Lisa Degnan, Pooja Shah, Anita Siu, and Marc Sturgill. Achievement of therapeutic vancomycin trough serum concentrations with empiric dosing in neonatal intensive care unit patients. *The Pediatric Infectious Disease Journal*, 34(7):742–747, 2015.

[165] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[166] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[167] Jason A Roberts, Ross Norris, David L Paterson, and Jennifer H Martin. Therapeutic drug monitoring of antimicrobials. *British Journal of Clinical Pharmacology*, 73(1):27–36, 2012.

[168] Matthew R Robinson, Naomi R Wray, and Peter M Visscher. Explaining additional genetic variation in complex traits. *Trends in Genetics*, 30(4):124–132, 2014.

[169] Allen D Roses. Pharmacogenetics and the practice of medicine. *Nature*, 405(6788):857, 2000.

[170] Malcolm Rowland, Thomas N Tozer, Hartmut Derendorf, and Guenther Hochhaus. *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications*. Wolters Kluwer Health/Lippincott William & Wilkins Philadelphia, PA, 2011.

[171] Michael Rybak, Ben Lomaestro, John C Rotschafer, Robert Moellering Jr, William Craig, Marianne Billeter, Joseph R Dalovisio, and Donald P Levine. Therapeutic monitoring of vancomycin in adult patients: a consensus review of the american society of health-system pharmacists, the infectious diseases society of america, and the society of infectious diseases pharmacists. *American Journal of Health-System Pharmacy*, 66(1):82–98, 2009.

[172] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[173] Marc A Schaub, Alan P Boyle, Anshul Kundaje, Serafim Batzoglou, and Michael Snyder. Linking disease associations with regulatory information in the human genome. *Genome Research*, 22(9):1748–1759, 2012.

[174] R. E. Seay, R. C. Brundage, P. D. Jensen, C. G. Schilling, and B. E. Edgren. Population pharmacokinetics of vancomycin in neonates. *Clinical Pharmacology and Therapeutics*, 56(2):169–175, August 1994.

[175] Lewis B Sheiner and Stuart L Beal. Bayesian individualization of pharmacokinetics: simple implementation and comparison with non-bayesian methods. *Journal of Pharmaceutical Sciences*, 71(12):1344–1348, 1982.

[176] Gillian M. Shenfield and Raymond G. Morris. Therapeutic drug monitoring. *Current Opinion in Anaesthesiology*, 15(6):687–692, December 2002.

[177] Shaojun Shi and Yunqiao Li. Interplay of drug-metabolizing enzymes and transporters in drug absorption and disposition. *Current Drug Metabolism*, 15(10):915–941, 2014.

[178] Kimberly Shoenbill, Norman Fost, Umberto Tachinardi, and Eneida A. Mendonca. Genetic data and electronic health records: a discussion of ethical, logistical and technological considerations. *Journal of the American Medical Informatics Association*, 21(1):171–180, February 2014.

[179] AR Shuldiner, MV Relling, JF Peterson, K Hicks, RR Freimuth, W Sadee, Naveen Luke Pereira, DM Roden, JA Johnson, TE Klein, et al. The pharmacogenomics research network translational pharmacogenetics program: overcoming challenges of real-world implementation. *Clinical Pharmacology & Therapeutics*, 94(2):207–210, 2013.

[180] Fleur S Sinkeler, Timo R de Haan, Caspar J Hodiamont, Yuma A Bijleveld, Dasja Pajkrt, and Ron AA Mathôt. Inadequate vancomycin therapy in term and preterm neonates: a retrospective analysis of trough serum concentrations in relation to minimal inhibitory concentrations. *BMC Pediatrics*, 14(1):193, 2014.

[181] Christine A. Sinsky, John W. Beasley, Greg E. Simmons, and Richard J. Baron. Electronic health records: design, implementation, and policy for higher-value primary care. *Annals of Internal Medicine*, 160(10):727–728, May 2014.

[182] Elise MA Slob, Susanne JH Vijverberg, Mariëlle W Pijnenburg, Gerard H Koppelman, and Anke-Hilse Maitland-van der Zee. What do we need to transfer pharmacogenetics findings into the clinic?, 2018.

[183] Katarzyna Smietana, Leeland Ekstrom, Barbara Jeffery, and Martin Moller. Improving r&d productivity. *Nature Reviews Drug Discovery*, 14(7):455–457, 2015.

[184] Offie Porat Soldin and Steven J Soldin. Therapeutic drug monitoring in pediatrics. *Therapeutic Drug Monitoring*, 24(1):1, 2002.

[185] Janet Sultana, Paola Cutroneo, and Gianluca Trifirò. Clinical and economic burden of adverse drug reactions. *Journal of Pharmacology and Pharmacotherapeutics*, 4(Suppl1):S73, 2013.

[186] Yosuke Suzuki, Kanako Kawasaki, Yuhki Sato, Issei Tokimatsu, Hiroki Itoh, Kazufumi Hiramatsu, Masaharu Takeyama, and Jun-ichi Kadota. Is peak concentration needed in therapeutic drug monitoring of vancomycin? a pharmacokinetic-pharmacodynamic analysis in patients with methicillin-resistant staphylococcus aureus pneumonia. *Chemotherapy*, 58(4):308–312, 2012.

[187] JJ Swen, M Nijenhuis, A de Boer, L Grandia, Anke-Hilse Maitland-van der Zee, H Mulder, GAPJM Rongen, RHN Schaik, Talitha Schalekamp, DJ Touw, et al. Pharmacogenetics: from bench to byte-an update of guidelines. *Clinical Pharmacology & Therapeutics*, 89(5):662–673, 2011.

[188] Kadri Tamme, Kersti Oselin, Karin Kipper, Tõnis Tasa, Tuuli Metsvaht, Juri Karjagin, Koit Herodes, Helmut Kern, and Joel Starkopf. Pharmacokinetics and pharmacodynamics of piperacillin/ tazobactam during high volume haemodiafiltration in patients with septic shock. *Acta Anaesthesiologica Scandinavica*, 2(60):230–240, 2016.

[189] Susan M Tange, Vijay L Grey, and Pierre E Senécal. Therapeutic drug monitoring in pediatrics: a need for improvement. *The Journal of Clinical Pharmacology*, 34(3):200–214, 1994.

[190] Tõnis Tasa, Kristi Krebs, Mart Kals, Reedik Mägi, Volker M Lauschke, Toomas Haller, Tarmo Puurand, Maido Remm, Tõnu Esko, Andres Metspalu, et al. Genetic variation in the Estonian population: Pharmacogenomics study of adverse drug effects using electronic health records. *European Journal of Human Genetics*, 27(3):442, 2019.

[191] Tõnis Tasa, Tuuli Metsvaht, Riste Kalamees, Jaak Vilo, and Irja Lutsar. Dosopt: a tool for personalized Bayesian dose adjustment of vancomycin in neo-nates. *Therapeutic Drug Monitoring*, 39(6):604–613, 2017.

[192] Tõnis Tasa, Riste Kalamees, Jaak Vilo, Irja Lutsar, and Tuuli Metsvaht. External evaluation of population pharmacokinetic models for vancomycin in neonates. *bioRxiv*, page 458125, 2018.

[193] Daan J Touw, Cees Neef, Alison H Thomson, Alexander A Vinks, et al. Cost-effectiveness of therapeutic drug monitoring: a systematic review. *Therapeutic Drug Monitoring*, 27(1):10–17, 2005.

[194] Sheng-Hsuan Tseng, Chuan Poh Lim, Qi Chen, Cheng Cai Tang, Sing Teang Kong, and Paul Chi-Lui Ho. Evaluating the relationship between vancomycin trough concentration and 24-hour area under the concentration-time curve in neonates. *Antimicrobial Agents and Chemotherapy*, 62(4):e01647–17, 2018.

[195] Tove Tuntland, Brian Ethell, Takatoshi Kosaka, Francesca Blasco, Richard Xu Zang, Monish Jain, Ty Gould, and Keith Hoffmaster. Implementation of pharmacokinetic and pharmacodynamic strategies in early research phases of drug discovery and development at novartis institute of biomedical research. *Frontiers in Pharmacology*, 5:174, 2014.

[196] CH Van Der Wouden, A Cambon-Thomsen, E Cecchin, KC Cheung, C Lucia Dávila-Fajardo, VH Deneer, V Dolžan, M Ingelman-Sundberg, Siv Jönsson, Mats O Karlsson, et al. Implementing pharmacogenomics in europe: design and implementation strategy of the ubiquitous pharmacogenomics consortium. *Clinical Pharmacology & Therapeutics*, 101(3):341–358, 2017.

[197] Kristjan Vassil. Estonian e-government ecosystem: Foundation, applications, outcomes. *Background paper for World Development Report*, 2016.

[198] Peter M Visscher, Matthew A Brown, Mark I McCarthy, and Jian Yang. Five years of GWAS discovery. *The American Journal of Human Genetics*, 90(1):7–24, 2012.

[199] Peter M Visscher, Naomi R Wray, Qian Zhang, Pamela Sklar, Mark I McCarthy, Matthew A Brown, and Jian Yang. 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22, 2017.

[200] Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[201] Sook Wah Yee, Yukihide Momozawa, Yoichiro Kamatani, Rachel F Tyndale, Richard M Weinshilboum, Mark J Ratain, Kathleen M Giacomini, and Michiaki Kubo. Genomewide association studies in pharmacogenomics: Meeting report of the NIH pharmacogenomics research network-RIKEN (PGRN-RIKEN) collaboration. *Clinical Pharmacology & Therapeutics*, 100(5):423–426, 2016.

[202] Chen-Yan Zhao, Zheng Jiao, Jun-Jun Mao, and Xiao-Yan Qiu. External evaluation of published population pharmacokinetic models of tacrolimus in adult renal transplant recipients. *British Journal of Clinical Pharmacology*, 81(5):891–907, 2016.

[203] Wei Zhao, Florentia Kaguelidou, Valérie Biran, Daolun Zhang, Karel Allegaert, Edmund V Capparelli, Nick Holford, Toshimi Kimura, Yoke-Lin Lo, José-Esteban Peris, et al. External evaluation of population pharmacokinetic models of vancomycin in neonates: the transferability of published models to different clinical settings. *British Journal of Clinical Pharmacology*, 75(4):1068–1080, 2013.

[204] Wei Zhao, Emmanuel Lopez, Valérie Biran, Xavier Durrmeyer, May Fakhoury, and Evelyne Jacqz-Aigrain. Vancomycin continuous infusion in neonates: dosing optimisation and therapeutic drug monitoring. *Archives of Disease in Childhood*, 98(6):449–453, 2013.

[205] Michael Zimmermann, Maria Zimmermann-Kogadeeva, Rebekka Wegmann, and Andrew L Goodman. Separating host and microbiome contributions to drug pharmacokinetics and toxicity. *Science*, 363(6427):eaat9931, 2019.

ACKNOWLEDGEMENTS

For getting this far I am most grateful to my wife Mari-Liis who gave me the time that was needed to work on this PhD outside other duties mostly during "vacation" times and weekends, and my parents whose help in looking after Mattias has been invaluable.

All my supervisors, Jaak Vilo, Lili Milani and Tuuli Metsvaht, played an important role. I want to thank them for creating the conditions to get work done, great scientific collaborations and guidance in co-authored projects. Since almost no scientific work is a solo project then I want to thank all my collaborators on all my papers and people that have played a major role in some other indirect way. A special thank you goes to Irja Lutsar, Toomas Haller, Mart Kals, Riste Kalamees, Hiie Soeorg, Kristi Krebs, Maarja Hallik, Helgi Padari, Kadri Rekker, Maire Peters, Maris Alver, Raivo Kolde, Hedi Peterson, Liis Kolberg, Tauno Metsalu, Kadri Tamme, Reedik Mägi, Meelis Kull, Volker M. Lauschke, Peeter Padrik, Andres Metspalu and Tõnu Esko.

This work was financially supported by the Insitute of Computer Science, European Regional Development Fund, DoRa internationalisation program for PhD students coordinated by Archimedes Foundation, TerVE program grant PerMed I, the Estonian Doctoral School of Information Technologies, TBD Biodiscovery OÜ, Estonian Centre of Excellence in Computer Science (EXCS), Horizon2020 project MMVBS17039R and grants IUT34-4, PRG184 and IUT34-24.

SISUKOKKUVÕTE

Bioinformaatika meetodid personaalses farmakoteraapias

Bioloogiliste ja muude terviseandmetega seotud andmehulgad kasvavad väga kiiresti. Täppismeditsiini eesmärk on kogutud andmestest saadud informatsiooni kasutada inimeste ravis. See tähendab, et ravi määramisel, jälgimisel ja juhtimisel juhindutakse patsiendi individuaalsetest bioloogilistest eripäradest. See nõuab toimivaid meetoteid, spetsialistidele käepäraseid tööriistu, ja arusaadavaid kliinilisi tõlgendusi. Bioloogiliste andmete suur maht ja struktuurne keerukus on analüüsi jaoks oluline takistus. Selles doktoritöös käsitleme nelja avaldatud teadusartiklit, mis panustavad täppismeditsiini mitmesse erinevasse tahku.

Täppismeditsiin ei kasuta ainult geneetilisi andmeid. Patsientidelt mõõdetud ravimikontsentratsioone saab kasutada juhtimaks ravi kvantitatiivsetete meetoditega. **Artiklis I** arendasime selle ilmestamiseks tööriista, mis võimaldab kasutada vastsündinutel mõõdetud ravimi vereproove, et valida neile manustatavat vankomütsiini doosi. Veebiaadressil www.biit.cs.ut.ee/dosopt vabalt kättesaadavaks tehtud tööriist, DosOpt, kombineerib selleks patsientide kontsentratsioonimõõtmiseid uuritava populatsiooni ravimikineetikale tehtavate eeldustega. Me näitasime Tartu Ülikooli kliinikumis kogutud andmetel retrospektiivselt, et sellega saavutame täpsemaid tulemusi kui traditsiooniliselt juhitud ravis.

Kuna dooside optimeerimise tulemused on sõltuvad esmastest ravimikineetikale tehtavatest eeldustest, siis **artiklis II** kasutasime DosOpti, et valideerida Eesti andmestiku abil akadeemilisest kirjandusest kogutud vankomütsiini vastsündinute ravimikineetika mudeleid. Selleks hindasime erinevate mudelite ennustusvõimet ja võrdlesime erinevate kaasatud individuaalsete kontsentratsioonide arvu korral fikseeritud ravieesmärkide saavutamise määrasid. Selle eesmärgiks oli aidata valida DosOpti jaoks Eesti populatsioonis kõige sobivam ravimikineetika alusmudel, mida siis tööriista prospektiivse valideerimise jaoks kasutada.

Geneetika on ravimimõjude avaldumises väga oluline. Selliste variantide tuvastamise, seoste interpreteetimise ja kommunikeerimisega tegeleb farmakogeneetika. Kuna suure hulga ravimite metabolism toimub läbi geenide, mida tuntakse kui väga oluliste farmakogeenidena (Very Important Pharmacogenes), siis võib nendes geenides esinevate muutuste bioloogiline mõju olla piisvalt suur, et seda haigetele ravimite määramisel arvesse võtta. **Artiklis III** kasutasime Eesti Geenivaramu rahvastikupõhiseid andmeid, et kontrollida juba teadaolevalt olulisi ja otsida uusi ravimikõrvalmõjude tekkimisega seotud geneetilisi markereid kindlate ravimite tarbimisel. Valideerisime mitmeid teadaolevalt olulisi geen-ravim seoseid, leidsime varem ravimikõrvalmõjudega seostatud geenides uusi sõltumatult olulisi variante ja leidsime ning valideerisime uue seose CTNNA3 geenivariandi jaoks, mis tõstab kõrvalmõjude sagedust põletikuvastaste ravimite tarbimise korral. Meie uuring näitas, et rahvastikupõhiseid terviseandmeid saab kasutada farmakogeneetilisteks uuringuteks, ja andis täpse kirjelduse Eesti rahvastikus levinud variantide esinemissagedusele.

Kõige levinum meetod geneetiliste ja väliste tunnuste vaheliste statistiliste seoste kirjeldamiseks on ülegenoomne seoseanalüüs (GWAS). See on tõhus ja kiire meetod, mille jaoks leidub palju tööriistu ja, mis on tihti esmane meetod genoomiandmete uurimiseks. GWAS analüüsist tulevad seoste andmed on väga mahukad ja edasised sammud lähtuvad analüüsi lõplikust eesmärgist. Tihti on, olenevalt edasise analüüsi jaoks kasutada olevatest ressurssidest, vaja uuritavate seoste hulka piirata. Üks sagedasemaid meetodeid selleks on visuaalselt uurida huvipakkuvaid geneetilisi regioone ja nendes leiduvate variantide GWAS p-väärtuste jaotumist. Kuigi sellised geneetiliste markerite graafikud bioloogilise olulise kohta informatsiooni ei paku, siis on teatavad signaalid, mis muudavad piirkonna edasise uurimise huvipakkuvamaks. Lookusgraafikute puhul on eelistatud geomeetriline sümmeetria koos välja joonistunud tippudega, rohkemaarvuliselt olulisi markereid, mis on omavahel aheldustasakaalutuses (LD) ja markerite väiksemad p-väärtused. Selliste graafikute läbivaatamine on väga aeganõudev. Kuna selle hõlbustamiseks puuduvad teadaolevalt vajalikud tööriistad, siis **artiklis IV** arendasime välja kaks: Manhattan Harvester ja Cropper, mis aitavad ülesandele kulutatavat aega vähendada. Need leiavad GWAS tulemusandmete põhjal huvitavad regioonid ja pakuvad neile välja paremusjärjestuse. Regioonide headust hinnatakse statistilise mudeliga, mis arendati välja GWAS jooniseid hinnanud ekspertide hinnangute põhjal.

Meie töö seob täppismeditsiini teaduse mitmeid tahke. Arendasime välja tööriista, millega saab teha ravi täpsemaks kasutades patsientidelt kogutud ravimikontsentratsioone dooside valimisel ja ravimimonitoorimisel. Teisalt näitasime, et terviseandmete ja geneetilise informatsiooni kombineerimisel on potentsiaali, et tuvastada geneetilisi variante, mille korral esineb suurema sagedusega ravimikõrvalmõjusid. Meie panuste hulka kuulub ka arvutuslik tööriist, mis aitab nii eelpool mainitud farmakogeneetilisi uuringuid kui muid geneetilisi analüüse efektiivsemalt ja kiiremini läbi viia.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name:	Tõnis Tasa
Date of Birth:	30.10.1989
Citizenship:	Estonian
E-mail:	t6nis.tasa@gmail.com

Education

2015–2019	University of Tartu, Faculty of Science and Technology, doctoral studies, specialty: Computer Science.
2013–2015	University of Tartu, Faculty of Mathematics and Computer Science, master's studies, specialty: Informatics.
2010–2013	University of Tartu, Faculty of Mathematics and Computer Science, bachelor's studies, specialty: Mathematical Statistics.

Employment

2018–...	Antegenes OÜ, chief product officer
2017–...	TBD Biodiscovery OÜ, specialist
2014–2016	University of Tartu, institute of computer science, scientific programmer
2013–2019	University of Tartu, institute of microbiology, specialist

Scientific work

Main fields of interest:

- pharmacogenetics
- bioinformatics
- therapeutic drug monitoring

ELULOOKIRJELDUS

Isikuandmed

Nimi:	Tõnis Tasa
Sünniaeg:	30.10.1989
Kodakondsus:	Eesti
E-mail:	t6nis.tasa@gmail.com

Haridus

2015–2019	Tartu Ülikool, Loodus- ja täppisteaduste valdkond, doktoriõpe, eriala: Informaatika.
2013–2015	Tartu Ülikool, Matemaatika-informaatikateaduskond, magistriõpe, eriala: Informaatika.
2010–2013	Tartu Ülikool, Matemaatika-informaatikateaduskond, bakalaureuseõpe, eriala: matemaatiline statistika.

Teenistuskäik

2018–...	Antegenes OÜ, arendusjuht
2017–...	TBD Biodiscovery OÜ, spetsialist
2014–2016	Tartu Ülikool, arvutiteaduse instituut, teaduslik programmeerija
2013–2019	Tartu Ülikool, mikrobioloogia instituut, spetsialist

Teadustegevus

Peamised uurimisvaldkonnad:

- farmakogeneetika
- bioinformaatika
- ravimite terapeutiline jälgimine

DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.