Workshop on Computational Models in Biology and Medicine 2022

Joint workshop of the GMDS & IBS-DR working groups “Statistical Methods in Bioinformatics” and “Mathematical Models in Medicine and Biology”

June 23 – 24, 2022

University of Göttingen /
University Medical Center Göttingen,
Germany

Contents

Workshop outline

This workshop intends to bring together researchers from different research areas such as bioinformatics, biostatistics and systems biology, who are interested in modeling and analysis of biological systems or in the development of statistical methods with applications in biology and medicine.

Keynotes

- Harald Binder (University of Freiburg): “Deep Generative Models for Single-Cell Sequencing Data”
- Viola Priesemann (MPI, Göttingen): “Inferring and Mitigating the Spread of COVID-19”
- Stefan Bonn (University Medical Center Hamburg-Eppendorf): “Bringing deep neural networks to the clinic: Precision, robustness, and interpretability”

Workshop venue

The workshop takes place at the Paulinerkirche, Papendiek 14, 37073 Göttingen.

Organization

The workshop is jointly organized by the GMDS/IBS working groups “Statistical Methods in Bioinformatics” (speakers: Michael Altenbuchinger, University Medical Center Göttingen; Klaus Jung, University of Veterinay Medicine Hannover) and “Mathematical Models in Medicine and Biology” (speakers: Markus Scholz, University of Leipzig; Ingmar Glauche, Technische Universität Dresden), as well as Tim Beißbarth (University Medical Center Göttingen) who is the local organizer.

Contact

Prof. Dr. Michael Altenbuchinger,
AG Statistical Bioinformatics,
E-mail: michael.altenbuchinger@bioinf.med.uni-goettingen.de

Support

The workshop is funded by the “Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS)” and the “Deutsche Region der Internationalen Biometrischen Gesellschaft (IBS-DR)”.

1 Program

Thursday, Jun 23, 2022

11:30 Registration opens (with small lunch)

12:45–13:00 Welcome

Session 1: Statistical modeling and machine learning

13:00–13:40 Keynote lecture: Harald Binder
Deep Generative Models for Single-Cell Sequencing Data

13:40–14:00 Anne-Christin Hauschild
Evaluating Federated Random Forest Coping with Limited Clinical and Biomedical Data

14:00–14:20 Lisa-Marie Bente
Deep Learning for the Prediction of Macrophage Polarization Status

14:20–14:40 Franziska Görtler
Gene Signature Calculation for Pathway Activity Prediction with Loss-Function Learning Methods

14:40–15:00 Marian Schön
DMC - an R package for Deconvolution Model Comparison

15:00–15:10 Election: (a) speaker of the GMDS/IBS working group “Mathematical Models in Medicine and Biology” and (b) speaker of the GMDS/IBS working group “Statistical Methods in Bioinformatics”.

15:10–16:20 Coffee break & poster session I

Session 2: Disease modeling and open topics

16:20–16:40 *Nicole Radde*
Quasi-Entropy Closure: A Fast and Reliable Approach to Close the Moment Equations of the Chemical Master Equation

16:40–17:00 Janne Pott
Genetically regulated gene expression and proteins revealed discordant effects

17:00–17:20 Liza Vinhoven
Systems medicine modelling in Cystic Fibrosis to predict possible drug targets and active compound combinations

17:20–17:40 Paul Rudolph
DynaCoSys: Dynamic model of the complement system to understand pathogen immune evasion

17:40–18:00 Theresa Kraft
Patient-specific identification of genome-wide methylation differences between intra- and extracranial melanoma metastases using Hidden Markov Models

18:30 Conference dinner at the Bullerjahn, Markt 9, 37073 Göttingen

Friday 24, 2022

Session 3: COVID-19

8:30–9:10 Keynote lecture: Viola Priesemann
Inferring and Mitigating the Spread of COVID-19

9:10–9:30 Sandra Timme
A state based model and feasibility study to design an optimal COVID-19 surveillance protocol for child care facilities

9:30–9:50 Manuel Lentzen
Risk Modelling of Severe COVID-19 Disease Progression Using Transformer-Based Models

9:50–10:10 Daniel Wolffram
Collaborative nowcasting of COVID-19 hospitalization incidences

10:10–11:30 Coffee break & poster session II

Session 4: Decision support systems

11:30–12:10 Keynote lecture: Stefan Bonn
Bringing deep neural networks to the clinic: Precision, robustness, and interpretability

12:10–12:30 Thomas Linden
Survival Multi-Modal Neural Ordinary Differential Equations for Mortality Prediction of Patients with Severe Lung Disease

12:30–12:50 Gunther Glehr
Neural networks to predict clinical events from cytometry data

13:10–13:30 Roman Schefzik
Using a prospective algorithm for systemic inflammatory response syndrome criteria to predict and diagnose sepsis in intensive care medicine

13:30–13:50 Helena Zacharias
A Predictive Model for Progression of CKD to Kidney Failure Based on Routine Laboratory Tests

13:50–14:00 Closing remarks & poster award

2 Keynotes

Keynote 1

Deep Generative Models for Single-Cell Sequencing Data

Harald Binder

Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center – University of Freiburg, Freiburg, Germany

Abstract: While deep generative approaches, such as variational autoencoders (VAEs), have become a hot topic of methods research for single-cell sequencing data, they still are not mainstream. One reason for this might be that their benefits (and potential pitfalls) are not fully worked out yet. For contributing to this, I will provide a brief introduction, highlighting some limitations of VAEs for single-cell RNA sequencing data, followed by three scenarios where they still can be successful. This includes planning of experiments, extracting high-dimensional patterns, and modeling of temporal data.

Keynote 2

Inferring and Mitigating the Spread of COVID-19

Viola Priesemann

Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

Abstract: Spreading dynamics is ubiquitous: activity spreads in neural networks, news and fake news branch out in social networks, and just recently the spread of a novel virus has disrupted the daily lives of people around the globe. Interestingly, in all these networks, the connections are not static, but change over time, e.g., to implement learning in neural networks or when mitigating the spread of SARS-CoV-2. We derive the principles of self-organization in these diverse networks, show under which conditions outbreaks can be mitigated, and how the behavioral feedback of people impacts the outbreaks of COVID-19, combining theoretical approaches with modelling, empirical data and Bayesian inference. Overall, we shed light on the past two years of the COVID-19 pandemics, with a special focus on the interaction between pandemics and infodemics, thus the intricate interaction between these distinct levels of the physical and social networks.

Keynote 3

Bringing deep neural networks to the clinic: Precision, robustness, and interpretability

Stefan Bonn

Institute of Medical Systems Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract: AI-based algorithms hold great promise to transform future medical decision making and patient care.

In this talk I would like to highlight some recent progress we made in developing deep learning-based algorithms for clinical data.

In addition, I will discuss the biggest bottlenecks we encounter when we want to translate an algorithm to a clinically useful decision support system.

3 Oral presentations

3.1 Evaluating Federated Random Forest Coping with Limited Clinical and Biomedical Data

Anne-Christin Hauschild

Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Germany

Abstract: The exploitation of the full potential of biomedical data for precision medicine is often hindered by limited data and restricted access due to data protection regulations. Federated learning approaches are developed to cope with these challenges by distributedly training and accumulating models without data centralization. Thus, we evaluated federated random forest models focussing on the heterogeneity within and between datasets. Three common challenges in clinical research were addressed: (i) a different number of parties, (ii) different sizes of datasets, and (iii) imbalanced classes, and evaluated the models on five biomedical datasets. The federated models consistently outperformed the average local models and performed as good as the classical models trained on the centralized data. However, the performance of local models decreases, yet not significantly, with an increasing number of models and decreasing size of the local dataset size. Moreover, when combining datasets of different sizes, the federated models vastly improve compared to the average performance of the local models, in particular when using model weighting. Finally, we could demonstrate that the federated models consistently outperform the local models by analyzing different class imbalances. In summary, our results suggest that federated machine learning crushes clinical research boundaries, data shortage, as well as biases and thus enables collaborations across institutes. In combination with secure multi-party computation methods, these technologies have the power to revolutionize both clinical research and practice and pave the way for precision medicine.

3.2 Deep Learning for the Prediction of Macrophage Polarization Status

Lisa-Marie Bente

TU Braunschweig / PLRI

Abstract: Macrophages are cells of the innate immune system. They are able to alter their phenotype and function during an infection. This change is called polarization. Indicators of an infection, also called pathogen-associated molecular patterns (PAMPs), are molecules derived from microorganisms, for example lipopolysaccharide (LPS). LPS is found on the outer cell wall of gram-negative bacteria and is released upon destruction of the bacterial cell.

Deep learning and artificial neural networks are becoming increasingly more popular for application in biological and medical research contexts. They are used for various tasks, including cell type classification.

We trained a convolutional neural network (CNN) to classify macrophages according to whether they show signs of polarization or not. These signs include elongation of cellular protrusions. Macrophages were incubated with LPS to induce polarization. The CNN was trained and cross-validated with images of these macrophages, and tested and evaluated statistically, as well as with the tool SHAP. The network was able to differentiate between polarized and non-polarized macrophages with an accuracy of 89.5%. Further work on this topic may increase the accuracy by more extensive tuning of hyperparameters or the use of transfer learning.

3.3 Gene Signature Calculation for Pathway Activity Prediction with Loss-Function Learning Methods

Franziska Görtler

University of Bergen, Norway

Abstract: In living organisms, communication cascades inside cells are described in signal transduction pathways. These pathways regulate the activation or deactivation of proteins in their cascades. Errors here lead to a protein expression profile different to normal cells and frequently cause diseases such as cancer, autoimmune diseases and diabetes. Often these changes are gradual and the patient can initially compensate for the dysregulation and no symptoms. Even so, on the cellular level the shifts can be observed. However, until now there exists no easy and cheap method to predict the up- or downregulated activity of the pathways of interest by sampling liquid biopsy data and getting the information of interest in a non-invasive and resource saving way. Also the existing pathway signatures refer to the pathway activation status which is not necessarily linked to the pathway activity.

We enhanced and adapted our existing loss-function machine learning model for digital tissue deconvolution for calculating pathway activity signatures. Here we use single cell datasets with known pathway activity status for preparing training and test sets. The usage of single cells makes the sets variable and big enough for the application of machine learning techniques. A gradient descent algorithm in a two steps loss-function learning algorithm is used to calculate the genes leading to the best prediction of the activity status of the pathway of interest. The received gene signature is checked for biological plausibility. Also the application of regularization methods like L1 regularization are implemented into the model to shorten the gene signature. The aim is to get a signature as predictable for pathway activity as necessary but as short as possible. This is important for a future clinical application as for screening examinations or for therapeutic monitoring of ongoing treatments short gene lists result in cheaper sequencing costs.

In my talk I will introduce the digital tissue deconvolution algorithm. Then the necessary adjustments which are the basis for our new pathway activity prediction algorithm are explained and justified. Then regularization methods for signature shortening will presented. The application of our pathway activity machine learning algorithm is demonstrated on a public available single cell dataset.

The results are exemplarily shown for the mTOR pathway. mTOR pathway is a central regulator of mammalian metabolism and physiology, with important roles in the function of tissues including liver, muscle, white and brown adipose tissue, and the brain, and is dysregulated in human diseases, such as diabetes, obesity, depression, and certain cancers.

3.4 DMC - an R package for Deconvolution Model Comparison

Marian Schön

University Regensburg, Institute of Functional Genomics, Statistical Bioinformatics

Abstract: Gene-expression profiling of bulk tissue averages the expression profiles of all cells in that tissue. Cell-type deconvolution is the bioinformatical task of revealing the cellular composition of a bulk tissue using its gene expression profile. In recent years, multiple cell-type deconvolution related methods have been published. Some of them offer ready to use deconvolution models, others provide algorithms to optimize a model on a set of cell types and the tissue scenario.

When applying such models onto bulk data – revealing their underlying cellular composition – not all models perform equally well. The performance heavily depends on the granularity of cell types, the training data and the algorithm used to optimize the model. A good deconvolution performance of a model on one bulk dataset does not necessarily imply a good performance on another dataset. Therefore, whenever deconvoluting a bulk dataset, the best model must be searched again.

In my presentation, I demonstrate how DMC – an R package for Deconvolution Model Comparison - helps to overcome the introduced problems of model selection in deconvolution analyses. DMC offers an automatized framework to compare multiple deconvolution model on the same bulk data. Besides multiple other characteristics, DMC monitors the performance per cell type, an overall performance and runtime per algorithm. It offers an easy framework for integration of new deconvolution algorithms and in the absence of bulk sequencing data enables simulations using single-cell RNA-seq data in order to evaluate algorithm performance in multiple scenarios. Numerical results are visualized in a comprehensive report to gain further insights into algorithms' strengths and weaknesses. DMC is publicly available via `https://github.com/spang-lab/DMC`

3.5 Quasi-Entropy Closure: A Fast and Reliable Approach to Close the Moment Equations of the Chemical Master Equation

Nicole Radde & Vincent Wagner

University of Stuttgart / Institute for Systems Theory and Automatic Control

Abstract: The Chemical Master Equation is the most comprehensive stochastic approach to describe the evolution of a (bio-)chemical reaction system. Its solution is a time-dependent probability distribution on all possible configurations of the system. As the number of possible configurations is typically very large, the Chemical Master Equation is often unsolvable in many applications. The Method of Moments reduces the system to the evolution of a few moments of this distribution, which are described by a system of ordinary differential equations. Those equations are not closed, since lower order moments generally depend on higher order moments. Various closure schemes have been suggested to solve this problem, with different advantages and limitations. Two major problems with existing moment closure approaches are first that some of them are computationally expensive, and second, they are open loop systems, which can diverge from the true solution.

Here we introduce Quasi-Entropy Closure, a moment closure scheme for the Method of Moments which estimates higher order moments by reconstructing the distribution that minimizes the distance to a uniform distribution subject to lower order moment constraints. Quasi-Entropy closure is similar to Zero-Information closure, which maximizes the information entropy. Results show that both approaches outperform truncation schemes, in which the influence of higher order moments on lower order moments is simply neglected. Moreover, Quasi-Entropy Closure is computationally much faster than Zero-Information Closure. Finally, our scheme includes a plausibility check for the existence of a distribution satisfying a given set of moments on the feasible set of configurations, which is able to indicate divergence from the true moment courses. Results are evaluated on different benchmark problems.

3.6 Genetically regulated gene expression and proteins revealed discordant effects

Janne Pott

Leipzig University, Institute of Medical Informatics, Statistics and Epidemiology (IMISE)

Abstract: Background: Although gene-expression (GE) and protein levels are strongly genetically regulated, their correlation is known to be low. Here we investigate this phenomenon by focusing on the genetic background of this correlation in order to understand the similarities and differences in the genetic regulation of these omics layers.

Methods: We performed association studies for a total of 184 proteins contained in the OLINK panels CVD II and CVD III in LIFE-Heart (n=1684) and LIFE-Adult (n=2014), respectively. We focused on the cis-region around the respective genes, and compared the effect direction of the detected protein quantitative trait loci (pQTLs) with known expression QTLs from GTEx v8. To examine locus-wide similarity, we tested for co-localization, association of our protein levels with genetically regulated GE using MetaXcan, and causal links of GE on protein expression via Mendelian Randomization (MR). Finally, we also attempted to find causal links from protein levels on coronary artery disease (CAD).

Results: After hierarchical FDR correction, we detected 123 gene loci significantly associated with their respective protein levels. In a sex-stratified subanalysis, we found four additional loci, i.e. genetic cis-regulations in blood were found for 127 (69%) of the proteins. Comparing with eQTLs, we found the same effect directions for most loci and tissues, but interestingly, 21 eQTLs with discordant effect directions in more than 75% of the analyzed tissues. Similarly, 106 genetically regulated GEs were associated with their proteins, but for 14 of them, the association was negative i.e. higher gene-expression was correlated with lower protein abundance. A total of 91 GEs had a causal effect on their protein level as detected by MR, 21 of them were negative. Nine gene-protein pairs had these discordant links in all three analyses (ADAMTS13, BLMH, CASP3, CXCL16, IL6R, MERTK, SFTPD, TGM2, and THBD). In our co-localization analysis, ADAMTS13, MERTK, and THBD showed independent signals across all tissues. CASP3, CXCL16, IL6R, and SFTPD had one co-localizing tissue each, but all other tissues suggested independent signals. Only BLMH and TGM2 were co-localized in most tissues. In the MR on CAD risk, we detected six proteins with causal effects (AXL, FABP2, IL6R, LPL, PCSK9, and TFPI).

Conclusion: While total GE and protein levels are only weakly correlated, we found high correlations between their genetically regulated components across multiple tissues. Of note, nine negative associations and causal effects of tissue-specific GE on protein levels were detected. Further biological validation is required to understand the mechanisms behind these observations.

3.7 Systems medicine modelling in Cystic Fibrosis to predict possible drug targets and active compound combinations

Liza Vinhoven

Department of Medical Bioinformatics, University Medical Center Göttingen

Abstract: Cystic Fibrosis (CF) is one of the most common genetic diseases prevalent among the white European population. It is caused by mutations of the cystic fibrosis transmembrane conductance regulator (CFTR) gene. Defective CFTR, has severe implications in the exocrine epithelia throughout the body. To date, more than 2000 mutations of the CFTR gene are known, several hundred of which are disease-causing, resulting in a vast range of geno- and phenotypes, which makes the development of therapeutics especially challenging. During the last years, different small-molecule therapeutics have been developed that amplify CFTR function. However, most of the therapeutics developed until now are not effective for all patients. The latest research efforts, therefore, focus on developing combination therapies to target multiple defects at once. For this purpose, high-throughput screens have been performed, which result in various candidate compounds. To provide an overview of already tested compounds, we established the publicly available database CandActCFTR, where substances are listed and categorized according to their interaction with CFTR. It becomes apparent that for about 80% of the compounds it is unknown whether they affect CFTR directly or indirectly. To elucidate the mechanism of action for promising candidate substances and be able to predict possible synergistic effects of substance combinations, we created a SBGN model of CFTR biogenesis, function and interactions (https://cf-map.uni-goettingen.de). We use molecular docking approaches to link the compounds from the database to targets in the disease map in order to create hypothesis on the mechanism of action for promising compounds and from the results prospectively be able to propose chemical scaffolds and compound families for testing. The model can ultimately be used to identify key steps and regulators to support the identification of potential targets and evaluate which factors interact to produce more functional CFTR. Taking into account the compounds from the CandActCFTR database, different substance combinations can then be proposed for laboratory testing.

Additionally, in order to support manual curation and upkeep of disease maps, we are working on text mining approaches to curate molecular interactions relevant to the respective disease map directly from publications. Here, the interactions are automatically parsed and categorized from the publications, displayed in a coarse disease map and can then be manually validated or rejected by experts.

3.8 DynaCoSys: Dynamic model of the complement system to understand pathogen immune evasion

Paul Rudolph

Leibniz Institute for Natural Product Research and Infection Biology Hans Knöll Institute (Leibniz-HKI)

Abstract: The human complement system is part of the innate immune response and plays a key role in defending the host against invading pathogens. Its main task is the recognition, subsequent opsonization and lysis of foreign invaders. The central opsonin of the complement system is C3b, which is a product of the proteolytic cleavage of C3. Since there is a continuous default basal level of active complement molecules, a tight regulation is required to protect the body's own cells from opsonization and from complement damage. One major complement regulator is Factor H, which attaches to cell surfaces and subsequently controls complement activation. Furthermore, the invading opportunistic human-pathogenic fungus Candida albicans has established evasion mechanisms to escape the host complement attack utilizing the molecule pH-regulated antigen 1 (Pra1). However, the secretion of Pra1 for immune evasion is contradictory. On the one hand, it reduces the local C3 and C3b concentration by binding these molecules and by transporting it from the cell resulting in spatial distancing. On the other hand, its C3-cleaving protease results in the products C3a-like(L) and C3bL molecule, which may amplify the complement activation such as C3a and C3b, respectively.To understand and study these paradoxical interactions in the immune evasion mechanisms between Pra1 and the complement molecules better, we extended our previously published single cell model called DynaCoSys [1]. As a system biology approach, this model aims to predict the opsonization level on the cell surface based on the surface bound Factor H concentration. It consists of ordinary differential equations (ODEs) for modelling the binding of molecules on the cell surface and partial differential equations (PDEs) for modelling the fluid phase concentration profiles around the cell. Furthermore, it focuses on the most important components of the complement cascade: C3 in the fluid phase, C3b in the fluid phase and on the cell surface as well as inactivated C3b on the cell surface. The other components of the complement system are combined in effective rates that represent the dynamics of the formation of several intermediate products of the cascade. Pra1 is integrated as a set of PDEs interacting with the complement molecules in the fluid phase. Using steady state analysis implemented with the finite element method (FEM) we investigated driving processes of the complement activation and regulation on the cell surface in the presence of Pra1. Furthermore, using an implicit Euler scheme we are also able to investigate the time-resolved dynamics of the system, which yields insights into the feasibility of the steady states and the spatial distancing. Since the exact rates associated with the Pra1 kinetics are currently unknown, we conducted a parameter screening to identify regimes of interest under which

Pra1-mediated immune evasion is possible. In addition, with our system biology approach we are able to quantify the effects of different immune mechanisms in the Pra1-complement kinetics. In conclusion, our approach gives useful insights into the interactions between complement regulators, such as Factor-H and Pra1, and is able to generate new hypotheses about the intrinsic behaviors of the complement system that can be tested in future experiments. Furthermore, our approach is implemented in a general framework, which can be extended with respect to other immune evasive pathogens and, thus, can deepen our understanding of the complement associated innate immune response.

[1] Tille A et al. (2020) Front Immunol, 11.

3.9 Patient-specific identification of genome-wide methylation differences between intra- and extracranial melanoma metastases using Hidden Markov Models

Theresa Kraft

IMB, TU Dresden

Abstract: Melanomas metastasize frequently to distant organs and especially intracranial metastases still represent an enormous challenge to improve life quality and survival of affected patients. Since epigenetic reprogramming has been suggested to play an important role in therapy resistance of melanomas, we systematically analyzed genome-wide DNA-methylation profiles of 24 patient-matched intra- and extracranial metastases pairs of melanoma patients. Hierarchical clustering of the methylomes revealed that intra- and extracranial metastases of individual patients were more similar to each other than to metastases of the same tissue type from different patients. To account for this, a personalized analysis of each patient-specific metastasis pair was performed using a three-state Hidden Markov Model. This enabled the classification of the methylation level of each individual CpG as decreased, unchanged, or increased in the intra- compared to the corresponding extracranial metastasis. The predicted DNA-methylation alterations were highly patient-specific differing in the number and methylation states of altered CpGs. Still, four important general observations were possible: (i) intracranial metastases of most patients showed mainly a reduction of DNA-methylation compared to corresponding extracranial metastases, (ii) cytokine signaling was most frequently affected by differential methylation in individual metastases pairs, but also MAPK, PI3K/Akt, and ECM signaling were frequently altered between intra- and extracranial melanoma metastases, (iii) several genes that were frequently altered in the same direction across multiple patients are involved in the regulation of cell signaling, cell growth, or cell death, and (iv) an enrichment of functional terms related to channel and transporter activities supports previous findings for the establishment of a brain-like phenotype of intracranial metastases. Moreover, the derived set of 22 candidate genes that distinguished intra- from extracranial metastases in more than 50% of patients included well-known oncogenes (e.g. PRKCA, DUSP6, BMP4) and several other genes known from neuronal disorders (e.g. EIF4B, SGK1, CACNG8). Our analysis contributes to an in-depth characterization of DNA-methylation differences between patient-matched intra- and extracranial melanoma metastases and may provide a basis for future experimental studies to identify targets for new therapeutic approaches.

3.10 A state based model and feasibility study to design an optimal COVID-19 surveillance protocol for child care facilities

S. Timme[1], *P. Rudolph*[1,2], *M.T. Figge*[1,3]

(1) Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology - Hans Knöll Institute, Jena, Germany

(2) Faculty of Biological Sciences, Friedrich Schiller University Jena, Jena, Germany

(3) Institute of Microbiology, Faculty of Biological Sciences, Friedrich Schiller University Jena, Jena, Germany

Abstract: During the global COVID-19 pandemic day care centers (DCCs) were closed to avert transmission. However, children are only mildly affected by the disease and closure of DCC imposes negative effects on children's health. Therefore, (re)opening DCCs while simultaneously providing continuous surveillance testing might be a feasible alternative. The objective of the study is to design an optimal COVID-19 surveillance protocol by state-based modeling and proof its feasibility [1].

To investigate feasibility, our collaborators conducted a 12-week longitudinal study starting in October 2020. Nine pre-selected DCCs in Wuerzburg (Germany) were assigned to four different surveillance approaches (modules) for the detection of SARS-CoV-2 by RT-PCR. Asymptomatic children and childcare workers (CCWs) were screened by mid-turbinate nasal swabs twice weekly (module 1, one DCC), once weekly (module 2, one DCC) or by home-sampled saliva twice weekly (module 3, two DCCs). In module 4, symptomatic children, CCWs and the respective household contacts of five DCCs were offered tests on demand. Consent to surveillance (71%) and successful study participation (68%) was highest in non-invasive home-sampled saliva testing. During the 12-week study period, no SARS-CoV-2 infection was detected in asymptomatic individuals.

This feasibility study could, however, not reveal optimal surveillance strategies that allow keeping DCCs open. Thus, we developed a virtual infection spread model for DCCs as a state-based model (SBM), which simulates the infection spread in a DCC after the introduction of one primary case, which can either be an infected child or an infected CCW. For each individual, viral load kinetics are modeled using a piecewise linear model [2] and infection transmission is modeled using an aerosol transmission and infection risk calculator [3]. Using the SBM we evaluate an optimal DCC surveillance settings with respect to (i) how many children have to participate in regular testing, (ii) how frequent testing has to be performed, (iii) and whether this would allow DCCs to remain open.

Simulation results of the SBM show that the expected number of a secondary infection is less than 1 provided that twice-weekly testing a and children participation of over 50% is realized. Furthermore, it supports the importance

of testing on Mondays to minimize the risk of outbreaks. In conclusion, surveillance of SARS-CoV-2 in DCC by continuous non-invasive sampling is feasible and the SBM provides evidence that (re)opening DCCs is possible when appropriate surveillance strategies are conducted.

[1] Forster J et al. (2022) Feasibility of SARS-CoV-2 Surveillance Testing Among Children and Childcare Workers at German Day Care Centers: A Nonrandomized Controlled Trial. JAMA Netw Open 5(1), e2142057.

[2] Larremore et al. (2021) Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening. Science advances, 7(1).

[3] Lelieveld, J et al. (2020) Model Calculations of Aerosol Transmission and Infection Risk of COVID-19 in Indoor Environments. Int. J. Environ. Res. Public Health, 17, 8114."

3.11 Risk Modelling of Severe COVID-19 Disease Progression Using Transformer-Based Models

Manuel Lentzen

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

Abstract: In circumstances such as the COVID-19 pandemic, health care systems face a tremendous challenge since they can quickly collapse under the burden of the crisis. Personalized risk models can offer support in these situations by accurately and promptly forecasting the course of a patient's disease progression, enabling future public health actions to be more effective. For such applications, deep learning models, especially transformer-based models, are promising as they achieve state-of-the-art results in various fields and have recently been applied to structured EHR data for disease risk prediction. Therefore, we followed the previously published Med-BERT approach but extended it with information about medications, age, state of residence, and gender. Following a pre-training on around 700 million EHRs stemming from 3.8 million US patients in the IBM Explorys Therapeutic dataset, we used a subgroup of 101,046 COVID-19 patients to develop risk models for predicting acute respiratory manifestations (ARM) and hospitalization. We compared these models to XGBoost and Random Forest models and observed that all transformer-based models accurately predicted COVID-19 disease progression, obtaining AUCs of 77% for ARM and 81% for hospitalization. We then used the integrated gradients approach in conjunction with Bayesian networks to offer a detailed explanation for model predictions. This combination enabled us to understand the relationship between the features identified as most significant by our models. Finally, we explored the possibility of adapting our model to data from an Austrian hospital via transfer learning techniques, which on one hand, showed the principle potential of transfer learning while at the same time highlighting practical challenges due to differences in the available length of the medical history of US patients versus Austrian patients. To summarize, this work demonstrates the potential of modern deep learning approaches and, in particular transformer-based models for developing predictive models based on real-world data in the field of precision medicine.

3.12 Collaborative nowcasting of COVID-19 hospitalization incidences

Daniel Wolffram

Karlsruhe Institute of Technology (KIT), Heidelberg Institute for Theoretical Studies (HITS)

Abstract: The seven-day hospitalization incidence is one of the main indicators used to as- sess the pandemic situation in Germany. It is defined as the number of COVID-19 cases reported over a seven-day period and subsequently hospitalized. Due to de- lays between the reporting of a case and a possible hospitalization, and between the hospitalization date and appearance in the surveillance data, the most recent values of this indicator are typically incomplete and need to be corrected upwards over the following days and weeks. As a consequence, recent trends cannot be read directly from the raw values of the seven-day hospitalization incidence. Statistical nowcasting methods can be used to adjust these, and obtain a more realistic picture of recent tendencies. We report on a collaborative nowcasting platform called the German COVID-19 Nowcast Hub (`https://covid19nowcasthub.de/`) which unites probabilistic nowcasts of the seven-day hospitalization incidence from eight distinct models run independently by different teams of researchers. Nowcasts are collected on a daily basis and combined into an ensemble, which represents the main output of the project. The different methods and ensembles will be compared systematical- ly in a pre-registered evaluation study (running from November 2021 through April 2021). We will discuss preliminary evaluation results and the potential and pitfalls of collaborative real-time analyses during the pandemic.

3.13 Survival Multi-Modal Neural Ordinary Differential Equations for Mortality Prediction of Patients with Severe Lung Disease

Thomas Linden

Fraunhofer SCAI

Abstract: The ongoing pandemic situation has demonstrated that medical decisions taken in intensive care units (ICU) are often time critical. To provide optimal care and resource allocation it is necessary to identify patients with high mortality risk as early as possible. However, development of according machine learning models is challenging due to a) a mix of longitudinal and static data; b) differences in time intervals between measured outcomes; and c) frequently occurring missing values. In this work we introduce a novel machine learning method for mortality risk predictions based on ICU data of patients with severe lung disease taken from the MIMIC-III dataset. Our Survival Multi-Modal Neural Ordinary Differential Equation (SMNODE) model is a hybrid mechanistic / neural network-based approach, which handles a mix of longitudinal and static data, implicitly accounts for missing values and deals with right-censored clinical outcomes, such as survival. Comparison of SMNODEs against several competing methods demonstrated a good prediction performance (C-index ~ 0.75) for mortality prediction of pneumonia and mechanically ventilated patients. Using recent developments from the field of Explainable AI shows, which measurements might be most critical to watch within a clinical routine setting.

3.14 Neural networks to predict clinical events from cytometry data

Gunther Glehr

Department of Surgery, University Hospital Regensburg, Regensburg, Germany

Abstract: From a computational perspective, clinical decision making requires classifying a patient into classes that respond similarly to treatment. Cytometry characterises individual cells in patient specimens by size, complexity or surface markers. The result is a matrix per specimen with an unordered number of rows reflecting the cells and a defined number of cell-parameters. These data can be used for classification, hence also for clinical decision making. Classically, predictive modelling uses cell subpopulation quantities as a predictor to classify each sample. The cell subpopulations are defined by binning cells with similar characteristics. Sequential gating, manually defining cuts in the parameters, or automated clustering are possibilities for binning.

In contrast to this two-step procedure, we investigated the application of neural networks to classify cytometry samples without prior cell population identification.

We created a python package for Cell Cloud Classification (CCC): A modular framework for neural networks to classify and investigate cytometry samples. The framework consists of three parts. First, the FeatureLearner transforms the measured matrix of a biological sample into a new matrix with new, predictive cell-features. Second, the Pooler aggregates these cell-features into a composition vector of values per biological sample. Third, the Predictor uses this composition vector to predict the final class.

Finally, we applied CCC in binary and multi-class simulations and real datasets.

3.15 Using a prospective algorithm for systemic inflammatory response syndrome criteria to predict and diagnose sepsis in intensive care medicine

Roman Schefzik, Anna-Meagan Fairley, Michael Hagmann, Bianka Hahn, Franziska Holke, Holger A. Lindner, Timo Sturm and Verena Schneider-Lindner

Department of Anesthesiology and Surgical Intensive Care Medicine; Medical Faculty Mannheim, Heidelberg University

Abstract: Sepsis is one of the leading causes of death, in particular in intensive care units (ICUs), and its timely detection and treatment improve clinical outcome and survival. Originally, sepsis had been defined as the co-occurrence of the systemic inflammatory response syndrome (SIRS) and an infection. Here, SIRS refers to the concurrent fulfilment of at least two out of the four following clinical criteria: tachycardia, tachypnea, abnormal body temperature, and abnormal leukocyte count. Even though controversially not being explicitly present in the latest consensus definition of sepsis any more, the SIRS is expected to remain an important predictor in the context of sepsis prediction and diagnosis.

To adequately describe and investigate the role of SIRS, we here introduce a dynamic and prospective algorithm, which is specifically tailored to the setting of an ICU. In particular, the ICU-specific interventions of catecholamine therapy and mechanical ventilation are taken into account when determining the validity of the tachycardia and tachypnea criterion, respectively. While SIRS has previously typically been determined at time points in spot check evaluations only, our novel SIRS algorithm uses a dynamic, time interval-based concept of SIRS in a general approach. Basically, each measurement of an involved vital or laboratory parameter is carried forward until a new measurement emerges. However, maximum validity intervals for measurements are implemented, depending on the considered variable and determined in accordance with clinical expertise. Checking whether the respective measurement lies within a pre-defined, clinically sound range, while accounting for possible interplays between variables, then determines whether the corresponding SIRS criterion is considered to be fulfilled during the corresponding time interval or not.

In an application to a cohort of polytrauma patients from the surgical ICU of University Medical Center Mannheim, we demonstrate the usefulness of our SIRS algorithm in the context of sepsis prediction and diagnosis. To this purpose, we apply it to and evaluate its performance in two distinct time-specific scenarios: (I) within the first 24 hours after ICU admission (prediction), and (II) within the last 24 hours before an index time point (diagnosis), corresponding to the sepsis time point in the septic patient group and a time point of comparable ICU treatment duration in a control group, respectively. Here, the results obtained by our algorithm over the considered time intervals are aggregated using several summary descriptors, such as the average SIRS level, the trend and the fluctuation of the SIRS level. In particular, models based on our algorithm

typically show a superior, or at least equally good, discriminative performance compared to both an earlier (initial, retrospective) version of the algorithm (Lindner et al. 2016, Critical Care Medicine) and a basic, non-ICU-specific version of the algorithm. Moreover, our algorithm reveals a ranking of the SIRS criteria by their importance in predicting or diagnosing sepsis.

Overall, our SIRS algorithm provides a conceptually simple, yet well-performing and promising tool, which may be included in more comprehensive risk models for sepsis prediction and diagnosis.

3.16 A Predictive Model for Progression of CKD to Kidney Failure Based on Routine Laboratory Tests

Helena U. Zacharias, Michael Altenbuchinger, Ulla T. Schultheiss, Johannes Raffler, Fruzsina Kotsis, Sahar Ghasemi, Ibrahim Ali, Barbara Kollerits, Marie Metzger, Inga Steinbrenner, Peggy Sekula, Ziad A. Massy, Christian Combe, Philip A. Kalra, Florian Kronenberg, Bénédicte Stengel, Kai-Uwe Eckardt, Anna Köttgen, Matthias Schmid, Wolfram Gronwald, Peter J. Oefner

University Hospital Schleswig-Holstein / Institute of Clinical Molecular Biology

Abstract: Identification of chronic kidney disease (CKD) patients, who are at risk of progressing to kidney failure requiring kidney replacement therapy (KFRT) is important for clinical decision-making and clinical trial design and enrollment. We report a new 6-variable risk model based on routine laboratory parameters that predicts progression to KFRT in CKD patients [1]. To develop this model, we analyzed data from 4,915 patients of the prospective observational German Chronic Kidney Disease (GCKD) cohort study. During an observation period of 3.71 $\pm$ 0.88 years, 200 of the 4,915 patients (4.07%) progressed to initiation of KRT, defined as initiation of long-term dialysis or kidney transplantation. A least absolute shrinkage and selection operator (LASSO) Cox proportional hazards (PH) model was fit to select laboratory variables that best identified patients at high risk for KFRT. Model discrimination and calibration were assessed and compared against the gold standard 4-variable Tangri (T4) risk equation [2] both in a resampling approach within the GCKD development cohort and in three independent validation cohorts, comprising a total of 3,063 CKD patients, using cause-specific concordance (C) statistics, net reclassification improvement, and calibration graphs. The newly derived 6-variable risk score (Z6) included serum creatinine, albumin, cystatin C, and urea, as well as hemoglobin and the urinary albumin-creatinine ratio. In the the resampling approach, Z6 achieved a median C statistic of 0.909 (95% CI, 0.868-0.937) at 2 years after the baseline visit, whereas the T4 achieved a median C statistic of 0.855 (95% CI, 0.799-0.915). In the 3 independent validation cohorts, the Z6 C statistics were 0.894, 0.921, and 0.891, whereas the T4 C statistics were 0.882, 0.913, and 0.862. In conclusion, the proposed risk model based on easily accessible routine laboratory parameters led to a marked improvement in the distinction of CKD patients likely to progress to ESKD requiring KRT.

[1] Zacharias, H.U., Altenbuchinger, M., Schultheiss, U.T., Raffler, J., Kotsis, F., Ghasemi, S., Ali, I., Kollerits, B., Metzger, M., Steinbrenner, I., Sekula, P., Massy, Z.A., Combe, C., Kalra, P.A., Kronenberg, F., Stengel, B., Eckardt, K.-U., Köttgen, A., Schmid, M., Gronwald, W., Oefner, P.J.; A Predictive Model for Progression of CKD to Kidney Failure Based on Routine Laboratory Tests, American Journal of Kidney Diseases 2022, 79(2): 217-230.e1; https://doi.org/10.1053/j.ajkd.2021.05.018.

[2] Tangri, N., Stevens, L.A., Griffith, J., Tighiouart, H., Djurdjev, O., Naimark, D., Levin, A., and Levey, A.S. (2011). A predictive model for progression of chronic kidney disease to kidney failure. JAMA - J. Am. Med. Assoc. 305, 1553–1559.

4 Poster presentations: session I

4.1 DNABERT SNP Embeddings for Parkinson's Disease Diagnostic Prediction

Mohamed Aborageh

Fraunhofer SCAI

Abstract: Parkinson's Disease (PD) is a neurodegenerative disorder known for its polygenic nature, whose risk cannot be estimated by assessing information from single variants, but the total set of risk variants that comprise its genetic architecture. In recent years, studies have focused on assessing an individual's risk using polygenic risk scores (PRS), which is calculated as the sum of risk alleles carried by an individual, each weighted by their relative effect sizes obtained from genome-wide associated studies (GWAS) summary statistics. In that case, the resulting score represents the individual's genetic load for PD. However, the additive model of PRS does not account for gene-gene interactions. Additionally, pre-filtering of SNPs often results in a focus on more common variants. In a similar fashion, multiple studies attempted to use artificial neural networks (ANNs) for PD diagnosis, using genotype data as input for the models. However, using the variants' genotype does not provide enough information as it only accounts for the number of copies. Our study attempts to incorporate SNPs at sequence level, using sequence embeddings from DNABERT, a Bidirectional Encoder Representations from Transformers (BERT) model adapted to genomic DNA setting, in attempts to better represent each SNP with more detailed information when using them to predict PD diagnosis.

4.2 Identifying differences between paired patient-specific melanoma brain and extra-cranial metastases applying melanoma-specific gene regulatory networks

Konrad Grützmann

Bioinformatics Core Unit, Institute for Medical Informatics and Biometry (IMB), TU Dresden

Abstract: "Melanoma is the most deadly skin cancer because it frequently metastasizes. Targeted molecular therapy, e.g. BRAF/MEK inhibitors, and immune checkpoint inhibition were major breakthroughs in patient care. Still, many patients do not respond and side effects can be severe. While melanoma brain metastases often respond to modern therapy, they relapse much faster than other metastases. Hence, an advanced understanding of molecular mechanisms that distinguish brain from extra-cranial metastases is needed to develop better therapies. Melanoma have a high mutation rate and thus show many patient-specific molecular alterations. While recurrent patterns have been found, group comparison can only provide limited insights. Therefore, we here investigate 11 matched pairs of brain and extra-cranial melanoma metastases from seven patients in an individualized way.

For each sample pair, both metastases were compared and differentially expressed genes with correspondingly altered promoter methylation were determined. These gene sets barely overlapped between patients. Using them as starting point, we applied network modeling methods to elucidate which genes and pathways were differentially affected in each patient. Transcriptome and methylome data of 272 TCGA melanoma patients were used to train a transcriptional regulatory network. Therefore, methods of the regNet R package were applied. The expression of each gene was modeled as linear combination of its promoter methylation and the expression levels of all other genes. Lasso regression in combination with a significance test for lasso was used to learn a sparse model that only includes the most relevant predictors for each gene. Network inference and evaluation was repeated 25 times by randomly separating the TCGA data into 75

The total influence of one gene on another one involves the learned direct connection but also indirect connections, which potentially span the whole network. To calculate these so-called impacts, we applied network propagation methods from the regNet package. Impacts of each patient-specific differentially expressed gene on each other gene were calculated for both metastases of a patient. This yielded impact ratios, which represent the relatively increased or decreased regulatory influence of genes comparing the brain to its corresponding extra-cranial metastasis. We averaged the impact ratios over known metabolic, signaling and immune pathways. Cluster analyses of these data showed that sample pairs fell into three groups with relatively consistent impact ratios across pathways. These groups had on average rather higher, lower or balanced impacts in their brain compared to their corresponding extra-cranial

metastasis. Interestingly, metastasis pairs did not cluster according to the tissue of their extra-cranial metastasis, which one may have initially expected. Even more, several regions of an extra-cranial metastasis of the same patient showed up in different clusters. This indicated that histologically distinct regions of a metastasis could result from underlying regulatory differences of a rapidly evolving tumor.

Next, we compared the average expression ratio of a pathway to its average impact ratio and found a slight negative correlation for almost all metastasis pairs. This means a decreased impact on a pathway was often paralleled with its increased expression and vice versa. Hence, some pathways showed a rather uncontrolled expression upregulation in the brain metastasis. These included pyruvate metabolism, citrate cycle and pentose phosphate in most sample pairs. Pathways with higher impact but lower expression in the brain metastasis were shared by fewer sample pairs and almost only from the cluster with an average higher impact in brain. These were mainly signaling pathways like PI3K/Akt, MAPK, p53, Melanoma pathway, and immune pathways, e.g. cytokine receptor, T cell receptor signaling and IL-17 signaling.

Next, genes were sorted by impact ratio and the upper and lower 5% of the lists were considered, which are genes with higher or lower impact in the brain metastases, respectively. The lists of each metastasis pair were subjected to over-representation analyses. We focused on over-represented pathways that were exclusive for each of the three above mentioned cluster groups. There were almost only results from the lower 5% lists. Among the most frequently shared pathways within the group with higher impacts in brain metastases were migration of diverse immune cells, mast cell activity, myeloid cell differentiation, PI3K signaling, (neuro)inflammatory response and glial cell activation. Fewer over-represented pathways were found in the other groups, involving cytokine receptor activity and cell proliferation.

Summarizing, we showed that a melanoma-specific gene regulatory network can be trained from public omics data. It is applicable to other patient cohorts and can be used to uncover patient-specific regulation differences in melanoma brain metastasis. Differentially expressed genes barely overlap. However, the affected pathways showed some commonalities and allow to cluster patients into groups of similar regulatory patterns. The size of our cohort only allows for limited conclusions. However, this method could help to unravel potential causes for the reduced treatment response of melanoma brain metastases, and we already provided hints on some relations."

4.3 A new bootstrap approach for gene-set enrichment analysis with transcriptomics and proteomics data from studies on spinal muscular atrophy

Shamini Hemandhar Kumar (1,3), Amy Glynn (2,3), Ines Tapken (2,3), Peter Claus (2,3), Ricarda Kolbe(2,3), Klaus Jung (1)

(1) Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, Hannover, Germany

(2) SMATHERIA gGmbH – Non-Profit Biomedical Research Institute, Hannover, Germany

(3) Center for Systems Neuroscience (ZSN), Hannover, Germany

Abstract:

Introduction Pathway databases such as PANTHER (Mi et al., 2016) or REACTOME (Fabregat et al., 2016) as well as GO annotations (Gene Ontology Consortium, 2004) are usually manually or algorithmically curated. That means, genes or other features that belong to a pathway or GO term (i.e. gene-sets) are either added manually or computationally. In either way, there remains uncertainty whether a feature really belongs to a gene-set. The aim of this project is to study the results of enrichment analysis when features are removed or added from a gene-set.

Methods and Data We bootstrap genes from each pathway or GO term multiple times to study the variability of results. Finally, we aggregate the results from each bootstrap run by two approaches. Either p-values for each gene-set are merged by p-value combination as typical in meta-analysis, or ranking lists are merged by rank aggregation (Kolde et al., 2012). Gene-set can then be ranked newly by the combined p-values or by a score provided by the rank aggregation. We demonstrate the usability of our approach on transcriptomic and proteomics data from a study on Spinal Muscular Atrophy (SMA) in lung tissue of a severe SMA mouse model. SMA is a neurodegenerative disorder with multisystem involvement affecting mainly infants in severe forms of the disease.

Results and Discussion The bootstrap analysis shows which gene-sets remain robustly at the top of the list. These gene-sets can be used for the biological interpretation with a higher level of confidence. Other gene-sets drop from the top of the ranking list, i.e. their enrichment result is considered as less robust. Biologist should then be careful to include these gene-sets in the biological interpretation. As further steps of our research, we aim to combine the robustness evaluation of enrichment analysis from different omics levels in a joint score for each gene-set.

References:

- Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., ... & D'Eustachio, P. (2016). The reactome pathway knowledgebase. Nucleic acids research, 44(D1), D481-D487.
- Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic acids research, 32(suppl_1), D258-D261.
- Kolde, R., Laur, S., Adler, P. and Vilo, J., 2012. Robust rank aggregation for gene list integration and meta-analysis. Bioinformatics, 28(4), pp.573-580. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2016). PANTHER version 10: expanded protein families and functions, and analysis tools. Nucleic acids research, 44(D1), D336-D342 .

4.4 Dynamical modeling of pneumococcal serotypes in Germany

Matthias Horn

Institute for Medical Informatics, Statistics and Epidemiology (IMISE), University of Leipzig

Abstract: Streptococcus pneumoniae is among the most important causes of lower respiratory infections worldwide. It causes invasive pneumococcal disease (IPD) and community-acquired pneumonia (CAP), which are associated with premature deaths in all age groups and serious long-term effects in children. The pneumococcal polysaccharide capsule is highly heterogeneous with approximately 100 known serotypes. Upper airway carriage of S. pneumoniae is a prerequisite for disease and vaccines are widely administered to prevent infections. In Germany, several vaccines are available, which provide protection against subsets of serotypes. Introduction of pediatric vaccinations with a 7-valent pneumococcal conjugate vaccine (PCV7) in 2006, followed by PCV13 in 2009, led to a rapid decline of vaccine serotypes both in carriage prevalence and disease. Through herd effects, this decline was also pronounced in adults. However, reduction of vaccine serotypes came at the cost of expanding non-vaccine serotypes. Hence, novel higher valent PCVs were developed. Two next-generation PCVs, a 15- and a 20-valent PCV (PCV15 and PCV20), have recently been licensed for use in adults. Based on ordinary differential equations, we developed a dynamical transmission model specific for Germany, with the aim to understand the complex interplay of different serotypes in different age groups under continued PCV vaccinations and to predict carriage prevalence and IPD burden for serotypes included in these vaccines. Thereby, future vaccination recommendations can be guided. The model allows to follow serotype distributions longitudinally both in the absence and presence of PCV vaccinations. We considered eight age cohorts and seven serotype groups according to the composition of different pneumococcal vaccines. Our model is the first to study potential dynamics of the additional serotypes included in PCV15 and PCV20. The model predicted that by continuing the current vaccine policy (standard vaccination with PCV13 in children and with PPSV23 in adults) until 2031, IPD case counts due to any serotype in children less than 2 years of age will remain unchanged. There will be a continuous decrease of IPD cases in adults aged 16-59 years, but a 20% increase in adults 60 years of age or older. Furthermore, there will be a steady decrease of the proportion of carriage and IPD due to serotypes included in PCV7 and PCV13 over the model horizon and a steady rise of non-PCV13 serotypes in carriage and IPD. The highest increase for both pneumococcal carriage and absolute IPD case counts was predicted for serotypes 22F and 33F (included in both PCV15 and PCV20) and serotypes 8, 10A, 11A, 12F, and 15B (included in PCV20 only), particularly in older adults. Between 2022 and 2031, serotypes included in PCV20 only are expected to cause 19.7-25.3% of IPD cases in adults 60 years of age or older. We conclude that introduction of next-generation PCVs for

adults may prevent a substantial and increasing proportion of adult IPDs, with PCV20 having the potential to provide the broadest protection against pneumococcal disease.

4.5 Predicting targeted antibiotic therapy using patient similarity networks

Hannah Marchi (1,2) Sophie Thiesbrummel (1), Christiane Fuchs (1,2,3)

(1) Faculty of Business Administration and Economics, Bielefeld University, Germany
(2) Institute of Computational Biology, Helmholtz Zentrum München, Germany
(3) Faculty of Mathematics, Technical University Munich, Germany

Abstract: Antibiotic resistance represents a major challenge for society, health policy and the economy. Sometimes it occurs already a couple of years after the introduction of a new antibiotic (Ventola, 2015). The use of broad-spectrum antibiotics, which - as the name suggests - cover a wide range of pathogens, further enhances the spread of resistance. However, current initial therapy for sepsis usually consists of broad-spectrum antibiotics. The start of therapy within the first hours is critical for the survival probability of sepsis patients. The doctors' decision for the initial antibiotic is based on their experience and predefined guidelines. They take into account clinical patient data, blood values and vital signs which are available within the first hour. Laboratory analyses like resistograms provide information about the pathogens present. But these are usually only available after at least 48 hours. Due to limited time and lack of information, a broad-spectrum antibiotic is a good initial choice to save patients' lives since its wide coverage is likely to have an effect on the infection. However, the use of such broad-spectrum antibiotics greatly contributes to the spread and severity of antibiotic resistance in the long term. An important approach for reducing antibiotic resistance would thus be the targeted use of antibiotics in these cases.

With the help of statistical methods, we aim to find narrow spectrum antibiotics which are equally or even better suitable than the initially prescribed broad-spectrum antibiotic. We use data of the freely available MIMIC IV (Medical Information Mart for Intensive Care) database (`https://mimic.mit.edu/docs/about/`), which contains health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center (Boston, MA) between 2008 and 2019. We present our approach to deal with the question how to find a targeted therapy at the time of sepsis diagnosis and discuss upcoming challenges. In a previous step, we investigate how to model the efficacy of a prescribed antibiotic. We use the outcome of this investigation to build an antibiotic-patient matrix which contains the efficacy information per patient for each given antibiotic. This matrix is rather sparse, as each patient has received only a few of all possible antibiotics. The empty cells in the matrix would then be filled by combining similarity networks of antibiotics and patients. In this way, we expect to predict the efficacy of antibiotics which were actually not given to a specific patient. We investigate starting points and data requirements for the presented research question. Our final goal is to develop a clinical decision support system which recommends an effective and

targeted initial antibiotic with minimal side effects. The system would help make individualized, informed and rapid treatment decisions for new sepsis patients. Furthermore, the use of less broad-spectrum antibiotics would help to combat the spread and severity of antibiotic resistance.

4.6 Evaluation of preprocessing on single-cell RNA data integration analysis

Youngjun Park

Medical Informatics

Abstract: Recent advances in single-cell RNA (scRNA) sequencing have opened the possibility to study tissues down to the level of cellular populations. Subsequently, this enabled various scRNA studies that reported novel or previously undetected subpopulations and their functions. However, the heterogeneity in single-cell sequencing data makes it unfeasible to adequately integrate multiple datasets generated from different studies. This heterogeneity originates from various sources of noise due to technological limitations.

Thus, particular procedures are required to adjust such effects prior to further integrative analysis. Subsequently, over the last years, numerous single-cell data analysis tools have been introduced de-noising methods implementing various data transformation methods. Here, we investigated 22 of the most recent single-cell studies and found that many analyses procedures employed various data transformation and preprocessing steps without further reasoning. This fact is particularly alarming since these read-count transformations can alter data distribution and affect downstream cell clustering results. This study aims to investigate the effects of the various data transformation on three different public data scenarios and evaluate these using popular dimensionality reduction and clustering analysis. Furthermore, we discuss implication on the use of transfer learning for batch correction and de-noising. In summary, our benchmark work shows that a large portion of batch-effects and noise can be mitigated by simple but well chosen data transformations and suggest that such analysis should be the baseline for all studies and a proper comparison between batch effect correction methods.

4.7 Classifying nucleosome positioning with random forests based on local structural DNA information

Malte Sahrhage

Department of Medical Bioinformatics, University Medical Center Göttingen

Abstract: The accessibility of chromatin is an essential means of regulating gene activity. Inaccessible DNA regions are occupied by nucleosomes, thus not being available for transcription factors to initiate transcription. The binding location of these histone octamers is dependent on multiple factors and it has been debated in the past how much its positioning is supported directly by the DNA sequence. In this context, a pattern of periodically repeating A/T dinucleotides has been postulated that favors the binding of nucleosomes and has been linked to be present on the level of local DNA structure.

We created a random forest classification tool that distinguishes DNA sequences as nucleosome forming or -inhibting, based on a literature defined benchmark data set (Guo et al. 2014). Our approach is based on local structural DNA information and incorporates the original hypothesis of periodic dinucleotide repeats by transforming the data into a power spectrum. By applying this binary classifier in a sliding-window manner over any given genomic region, we are able to generate base pair-resolution scores to interpret the local nucleosome support by the nucleotide sequence.

On the poster we present the technical setup of the classifier, the pros and cons of using a deep learning approach on the same data set instead and show results concerning the dynamic interplay between transcription factors and the competing nucleosomes.

4.8 Identifying clusters in high-dimensional virome data derived from public human body site sequencing files

Josefin Säurich (1), Magdalena Kircher (1), Michael Selle (1), Michael Altenbuchinger (2), Gisa Gerold (3) and Klaus Jung (1)

(1) Institute for Animal Breeding and Genetics, University of Veterinary Medicine Hannover, 30559 Hannover, Germany

(2) Department of Medical Bioinformatics, University Medical Center Göttingen, 37077 Göttingen, Germany

(3) Institute of Biochemistry & Research Center for Emerging Infections and Zoonoses (RIZ), University of Veterinary Medicine Hannover, 30559 Hannover, Germany

Abstract:

Background The virome, the totality of viral genetic information in a particular environment, has become of increasing interest and widely studied over the last 20 years, especially in samples of individuals or environments such as animal species, the human or human organs (Moustafa et al., 2017; Zuo et al., 2020). The viral communities in humans are known to vary, for example, with the body compartment, changes in environment or age (Zárate et al., 2017). We aim to identify underlying patterns in high-dimensional virome data derived from publicly available next generation sequencing (NGS) experiments, and to link the patterns to sample annotation in order to better understand the biological variability. We further aim to compare different approaches for data processing and clustering.

Material and Methods We used publicly available sequencing data and the associated metadata from the NCBI sequence read archive (SRA) to construct a virome overview of selected human body sites. Sequencing data from different studies was processed to a read count matrix, which was subsequently annotated with features such as the body site or sex. The count matrix was used to investigate different normalization, transformation and clustering approaches to reveal patterns that can be linked to sample annotations.

Results and Discussion We generated a statistical overview of available data in the SRA that can be used for virome analysis. Preliminary results with 15 selected data files from the SRA support previous findings of different viral communities between organs, but also within organs (e.g. Zárate et al., 2017). The results show, on the one hand, correlation between viral distribution and sample metadata and, on the other hand, correlation between viral distribution and the chosen processing of the count data. To further investigate these results, statistical testing to compare viromes between organs is planned. The results suggest that the virome of human body sites is highly diverse and

processing methods of virome data must be chosen carefully. Especially the sparsity and skewness of the read count matrix will be a focus for further analysis. As next steps, we want to reduce batch-effects caused by different data sources and studies, and will use dimensionality reduction technique, such as t-SNE and UMAP, and clustering algorithms (Li et al., 2020) for data visualization.

References

- Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M. P., Hu, G., & Li, M. (2020). Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nature Communications, 11(1), 1–14. https://doi.org/10.1038/s41467-020-15851-3
- Moustafa, A., Xie, C., Kirkness, E., Biggs, W., Wong, E., Turpaz, Y., Bloom, K., Delwart, E., Nelson, K. E., Venter, J. C., & Telenti, A. (2017). The blood DNA virome in 8,000 humans. PLoS Pathogens, 13(3), 1–20. https://doi.org/10.1371/journal.ppat.1006292
- Zárate, S., Taboada, B., Yocupicio-Monroy, M., & Arias, C. F. (2017). Human Virome. Archives of Medical Research, 48(8), 701–716. https://doi.org/10.1016/j.arcmed.2018.01.005
- Zuo, T., Sun, Y., Wan, Y., Yeoh, Y. K., Zhang, F., Cheung, C. P., Chen, N., Luo, J., Wang, W., Sung, J. J. Y., Chan, P. K. S., Wang, K., Chan, F. K. L., Miao, Y., & Ng, S. C. (2020). Human-Gut-DNA Virome Variations across Geography, Ethnicity, and Urbanization. Cell Host and Microbe, 28(5), 741-751.e4. https://doi.org/10.1016/j.chom.2020.08.005

Acknowledgement: This research was funded was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [398066876/GRK 2485/1].

4.9 A biomathematical model of atherosclerosis in mice

Sibylle Schirm

Universität Leipzig / IMISE

Abstract: Atherosclerosis is one of the leading causes of death worldwide. The biomathematical modeling of the underlying disease and the therapeutic processes could be a useful tool to develop and improve prevention and treatment concepts for arteriosclerosis.

We propose here a biomathematical model of murine atherosclerosis under various diet and treatment conditions. The model is derived by translating known biological mechanisms into ordinary differential equations and assuming appropriate response kinetics to the interventions used. We explicitly describe the dynamics of relevant immune cells and lipid species in atherosclerotic lesions, including the degree of blood vessel obstruction by growing plaques. Unknown model parameters were determined by fitting the predictions of model simulations to time series data from mice experiments.

Adjustments of parameters resulted in a good agreement between model and data for 13 examined experimental scenarios. The model can be used to predict the outcome of alternative treatment plans with combined antibiotics, immune modulating, and lipid lowering drugs on high fat or normal diet.

We conclude that we have established a comprehensive biomathematical model of atherosclerosis in mice. Our goal is to validate the model using further experimental data. For the future, we plan to include further new therapy principles and transfer the model to the human situation.

4.10 Investigating the Efficacy of Antibiotics in Patients with Sepsis

Sophie Thiesbrummel

Bielefeld University

Abstract: Patients who suffer from sepsis need to be treated as an emergency due to the high mortality rate of about 20 - 60 %. A suitable therapy with antibiotics in the first hours after diagnosis is crucial for lowering the risk of death and reducing serious consequences for the patients. However, many informative data are only available after 48 hours at the earliest. This includes details about the pathogen which is responsible for the disease as well as the results of a resistogram, which provides information about which antibiotics are effective. Therefore, the treating physician has to decide on an antibiotic based on a paucity of data. As a result, the chosen antibiotic might not be suitable to treat the sepsis of a patient and the therapy may need to be rather quickly adjusted.

In clinical practice, the decision about which antibiotic will be given is made on the basis of the patient's vital signs, some laboratory values and the patient's general condition. This assessment requires sufficient experience of the physician.

Our goal is to use such vital signs to model the efficacy of a given antibiotic. For this, we work with the freely available Medical Information Mart for Intensive Care (MIMIC) IV database which contains health-related information of patients admitted to a tertiary academic medical center in Boston (USA). This database has the advantage of including a large number of data which are constantly updated. Therefore, we persistently benefit from new data.

In our approach, we address the research question of modeling the efficacy of antibiotics by using Hidden Markov Models (HMMs). HMMs are time series models and consist of two stochastic processes, an unobserved state process and a state-dependent process. In our setting, we use the patient's vital signs and laboratory values for the state-dependent process. Possible hidden states could be states affecting the patient's health condition, e. g. improvement or deterioration. These states might help to draw conclusions about the efficacy of the antibiotics.

Answering this research question could make an important contribution to better understanding the relationship between changes in vital signs, laboratory values and the efficacy of antibiotics. In addition, we might be capable of quantifying and tracing the decision-making process of the physicians. This could be the first step towards improving the prescription of effective antibiotic treatments for new patients, as well as eventually providing a basis for a decision support system in case of sepsis. The hope is that future sepsis emergency treatment can be provided in a quicker, more targeted and more effective manner."

5 Poster presentations: session II

5.1 Integration of functional genomics data to molecularly characterize eye size variation between D. americana and D. novamexicana

Linh Dang

Developmental Biology - University of Goettingen

Abstract: The action of gene products regulate developmental processes, resulting in natural variation in adult morphology. While the variation in developmental gene expression plays a key factor in rather simple morphological traits, the genetic and developmental basis of complex trait evolution remains largely elusive. In the quest to answer this question, we utilize the variation of eye sizes among fruit fly species as our study model. In the study, we link genetic variants associated with differences in head shape and eye size between Drosophila americana and D. novamexicana to variation in genome wide developmental gene expression (RNAseq) and gene regulation (ATACseq). This procedure is eventually able to reveal high confidence candidate genes for future functional validation experiments.

5.2 Modelling leukemia treatment using ordinary differential equations and likelihood-based approaches for improving the experimental design

Julian Wäsche

Bielefeld University

Abstract: Leukemia is the most frequent type of cancer for pediatric patients. It causes the production of proliferating white blood cells that are less functional. Chemotherapy is one of the most common treatment forms and it aims at destroying the leukemia cells. A good understanding of tumor growth and of the effect of medication to impede its spread is crucial for a promising treatment of patients.

Since the amount of tumor cells within an individual are discrete, their evolution can explicitly be modelled by pure Markov jump processes (MJPs). Ordinary differential equations (ODEs) have proven to be suitable approximations of MJPs that facilitate parametric inference. Fitting a parametrized ODE to data can be achieved by maximum likelihood estimation. However, the quantities of interest, i.e. the explicit number of cells, are often only partially observed in practice. The concrete form of the observations and measurement errors need to be taken into account. Moreover, certain parametrizations as well as insufficient experimental data may lead to a situation in which parameters cannot be estimated unambiguously. Profile likelihoods represent well-known quantities to investigate parameter identifiability for ODE models. In order to compare different experimental designs, the computation of prediction profile likelihoods and validation profile likelihoods allows for the assessment of the informativeness of different settings and measurements, respectively.

This work presents an approach to model leukemia treatment in children based on ODEs. The use of profile likelihoods reveals that some parameters of the original model are non-identifiable. A re-parametrization reduces the number of model parameters and resolves non-identifiabilities. Simulating synthetic data allows for the reconstruction of different experimental settings, as well as the comparison of the uncertainty of corresponding parameter estimates. In particular, the evaluation of validation profile likelihoods yields valuable insights about the quality of measurement time points that may enhance the informative value of parameter estimates.

The results of this work provide medical scientists with a thorough statistical analysis of tumor treatment. Disease progression of leukemia in children and corresponding effects of medication are usually studied by animal experiments, often in mice. Since ethical and financial considerations speak for only as few animals as necessary being used, medical scientists are interested in taking measurements of samples as efficiently as possible. This work provides us with starting points to gain more information from fewer measurements and thereby to improve the reliability of statistical results.

5.3 Inference of differential gene regulatory networks from gene expression data using boosted differential trees

Gihanna Galindez

Technical University of Braunschweig

Abstract: Diseases can be caused by molecular perturbations that induce specific changes in regulatory interactions and their coordinated expression, also referred to as network rewiring. The detection of complex changes in regulatory connections remains a challenging task and novel non-parametric approaches for detecting network rewiring are needed. We present a new ensemble method, called BoostDiff (boosted differential regression trees), to infer a differential network discriminating between two conditions. BoostDiff builds an adaptively boosted (AdaBoost) ensemble of differential trees with respect to a target condition and uses differential variance improvement as a novel splitting criterion. Variable importance measures derived from the resulting models reflecting changes in gene expression predictability are then used to rank predictors and build the output differential networks. We first compare BoostDiff to existing differential network methods on simulated data. We then demonstrate the power of our approach when applied to real transcriptomics data in COVID-19 and Crohn's disease. BoostDiff complements standard differential expression analyses and identifies context-specific networks that are enriched with genes of known disease-relevant pathways.

5.4 On the parametrization of COVID-19 epidemiologic models

Yuri Kheifetz

IMISE (Institut für Medizinische Informatik, Statistik und Epidemiologie)

Abstract: "Predicting the spread of SARS-CoV-2 is a pressing need. Major tasks are to estimate (1) the dynamics of infected subjects, (2) requirements of medical resources during the course of the epidemic or (3) the effectiveness of non-pharmaceutical intervention programs (NPI).

A good prediction performance depends on the model's structure as well as on epidemiologic parameters, which are often scarcely known. Reported official databases are often biased due to lag in reporting of cases, changing testing policy, incompleteness of data, impact of changing age structures of infected patients on symptomatology, new emerging virus variants, non-pharmaceutical interventions continuously updated in response to the pandemic situation, and vaccination programs and its effectiveness.

To cover these aspects, we developed a comprehensive model based on mechanistic assumptions and phenomenological effects derived from data. We integrate a mechanistic SECIR-type epidemiologic model of five age groups as a hidden layer into a general Input-Output Non-Linear Dynamical System. These five age groups are subdevided to sub-groups, reflecting vaccination, waining and boostering. Changing factors of the system due to non-pharmaceutical interventions, testing strategy as well as vaccination rates are imposed as inputs to the system. We than estimate model parameters by a knowledge synthesis process considering parameter ranges derived from different external studies and other available data resources such as public data. Specifically, we use Bayesian inference for the parameter estimation, which could also be time-dependent. We demonstrate this on the example of our proposed model and data of Germany and Saxony.

By our approach, we can estimate and compare for example the relative effectiveness of non-pharmaceutical interventions and can provide predictions regarding the further course of the epidemic under specified scenarios. Our method of parameter estimation can be translated to other data sets, i.e. other countries, other COVID-19 epidemiologic models and other epidemiological contexts.

5.5 PepFuse: resolving peptide-peptide interactions in high-throughput proteomics data by group-wise mixed- graphical modeling

Robin Kosch

University Medical Center Göttingen

Abstract: Latest findings in protein research have demonstrated that a protein-level based analysis might not be sufficient for a holistic data exploration, since effects of individual peptides and especially of those with post-translational modifications (PTMs) cannot be considered adequately. These apparently minor variations can have regulatory effects, leading to several cancer manifestations or other adverse impacts.

Current high-throughput proteomics analysis techniques, like SWATH-MS, allow to accurately measure peptides to a large extent. Previous statistical tools deal with peptide abundances by aggregating the peptides according to their underlying proteins, i.e., by averaging their abundances and thus loose information on direct peptide-peptide associations. Other methods perform analyses on peptide-level but ignore their respective protein affiliations.

In this work, we propose a sophisticated framework which combines both previous approaches by incorporating prior knowledge. The PepFuse algorithm is based on graphical models but incorporates a group-wise graphical LASSO penalty in order to consider the protein affiliation. PepFuse can be applied as gaussian (GGM) or mixed graphical model (MGM), i.e., handle continuous features or with additional categorical features. These properties make it a very flexible tool for analyzing omics data with further phenotype data. Especially proteomics data with peptides containing PTMs can be resolved and associations with other clinical features can be investigated.

We demonstrated the superior performance of the novel group penalization scheme of the PepFuse algorithm in contrast to standard MGMs, both, in simulation and real biological data. We investigated a dataset from the German High-Grade Lymphoma Study Group (DSHNHL) containing SWATH-MS proteomics data, including PTMs, from 359 patients with Diffuse Large B-Cell Lymphoma and additional clinical and phenotypical features. We were able to identify key proteins for the cell-of-origin classification of DLBCL patients into Activated B-cell (ABC) Like and Germinal B-Cell (GCB) like DLBCLs.

5.6 Personalized prediction of mortality risks in chronic kidney disease patients

Bence Oláh[1]*, Michael Altenbuchinger, Jürgen Dönitz, Ulla Schultheiss, Fruzsina Kotsis, Jürgen Floege, Kai-Uwe Eckhardt, Wolfram Gronwald, Peter J. Oefner, Helena U. Zacharias*

[1] UKSH / Institute of Clinical Molecular Biology

Abstract: The identification of chronic kidney disease (CKD) patients at an increased risk of death is an important milestone in the nephrology health-care sector. It can improve clinical decision making and planning of interventions or risk mitigation strategies.

We have developed 2 new mortality risk prediction models for CKD patients based on 55 preselected continuous and categorical candidate predictors. These routine variables can be divided into 6 main categories: demographic, clinical and clinical chemistry, lifestyle, phenotype and disease history parameters.

For the model development we have analyzed the German Chronic Kidney Disease (GCKD) study, a prospective, observational, national cohort. We have investigated 4,081 GCKD study participants with a mean follow-up time of 5,86 $\pm$ 1.48 years. In total, 515 death events (13 %) occurred within this observation period.

We trained a least absolute shrinkage and selection operator (LASSO) Cox proportional hazards (PH) algorithm with hyperparameter tuning within a 10-fold internal cross-validation, and tested its performance within a subsampling approach. We assessed two different LASSO Cox PH models corresponding to two different hyperparameter settings according to internal partial likelihood deviance optimization. The best performing LASSO Cox PH model comprises 42 predictors. We further tested a second, more regularized LASSO Cox PH model consisting of 22 variables, which included, e.g., age, gender, smoking, cancer, diabetes mellitus, serum albumin, cystatin C, uric acid, hemoglobin etc. . We compared both models to two established mortality risk prediction models developed for CKD and hemodialysis patients, i.e., the Floege-[1] and Bansal-[2] equations. Furthermore, we refitted the coefficients of both the Floege- and Bansal-equations to the respective GCKD training sets within the subsampling approach and benchmarked our LASSO Cox PH models against these refitted equations. We evaluated the time-dependent performance based on concordance (C) statistics as well as net reclassification improvement (NRI) values on the GCKD test sets. Both of our models outperformed the original and refitted Floege and Bansal equations with, e.g., C statistics of 0.794 and 0.787 at 3 years after the baseline visit. Likewise, the two newly developed LASSO Cox PH equations yielded positive NRI values compared to both the original and refitted Floege- and Bansal equations across all time-points.

In conclusion, the two new mortality risk prediction models based on easily accessible routine patient parameters are of great promise to improve CKD

patient care. Their performance will be further assessed in external validation cohorts and they will be made available as an easy-to-use online service for fast implementation in routine patient care.

1. Jürgen Floege et al. Development and validation of a predictive mortality risk score from a European hemodialysis cohort. DOI: 10.1038/ki.2014.419
2. Nisha Bansal et al. Development and Validation of a Model to Predict 5-Year Risk of Death without ESRD among Older Adults with CKD. DOI: 10.2215/CJN.04650514

5.7 Comparative simulations of fungal infection dynamics in the human and murine alveolus

Christoph Saffer

Leibniz-HKI Jena

Abstract: Alveolar macrophages (AM) constitute the first line of immune cells in the lung and are responsible for clearing all types of microbial invaders, such as the pathogenic fungus Aspergillus fumigatus. If not efficiently cleared, intruding conidia can form hyphae within hours leading to life-threatening infections like invasive aspergillosis. However, the absolute number of AM, i.e., in human and mice, is still controversially discussed. In the human lung, a range of four up to 46 AM per alveolus is covered. Similarly, AM numbers for the murine lung vary.

In previous studies, we developed a spatio-temporal hybrid agent-based model (hABM) [1, 2, 3, 4], which allowed to simulate virtual infection scenarios of A. fumigatus in a single alveolus. This model considers a realistic to-scale representation of the alveolus, consisting of a ? sphere, with alveolar epithelial cells (AEC) of type 1 and 2 as well as pores of Kohn (PoK). In the model, the AEC, on which a conidium is located, secretes chemokines that diffuse on the inner surface of the alveolus. AM sense the chemokine gradient, which directs their migration towards the conidium. This allows to accurately simulating host-pathogen interactions in the human and murine lung.

In current study, we use the hABM to investigate the impact of the number of AM on the infection clearance during A. fumigatus lung infections for various chemokine parameters in both systems. As a measure of infection clearance, we empirically estimate the infection score, which is the fraction of simulations in which the fungus was not detected before onset of germination, i.e., before six hours. Observing the Weibull distributed clearing times and fitting the infection dynamics to a derived surrogate infection model (SIM) followed by statistical analyses let us evaluate AM ranges and gain quantitative understanding. Moreover, it enables us to identify key parameter combinations for clearance and investigate the transferability of experimental results in mice to the human system.

1 Pollmächer et al. (2014)

2 Pollmächer et al. (2015)

3 Blickensdorf et al. (2019)

4 Blickensdorf et al. (2019)

5.8 Modeling the interplay between risk perception, behavior and infection dynamics

Markus Schepers

IMBEI, JGU Mainz

Abstract: Standard epidemiological models of the spread of infectious diseases typically include compartments for the states of the disease, such as susceptible or infected, and certain assumptions on how the disease is transmitted. For instance, the model may refer to an underlying contact network. Yet, it has already been found in the 2009/10 influenza pandemic that preventive behavior in the population is more closely related to risk communication in the media than to the epidemiological risk of infection (Reintjes et al., 2016). Hence, concepts of psychology, social science and communication theory, such as risk perception, trust and knowledge of the virus, protective behavior and official and media communication, should be included in the models in order to gain a better understanding of infectious disease spread.

We will investigate the relationships between the following three aspects of an epidemic: 1) psychological factors, such as risk perception, trust (in public health authorities) and knowledge (of the virus and protective behavior), 2) people's actual behavior and 3) infection dynamics, such as case numbers and deaths. For the first two aspects, we will analyze data from the serial cross-sectional COSMO study (Covid-19 Snapshot Monitoring). In the long run, this study aims to contribute towards the WHO research agenda on effective risk communication during the emergence of a pandemic (Calleja et al., 2021).

5.9 Individual treatment effect estimation for survival data

Stefan Schrod

Universitäsmedizin Göttingen

Abstract: Outcome prediction is one of the routine applications of precision medicine. However, predicting individual treatment effects for patients remains challenging. In this context, so-called counterfactual analyses were established in recent years. Estimating causal effects from observational data is different from classical machine learning in that we never see the individual treatment effect in our training data; for each patient, we usually see his/her outcome either treated or untreated, but usually not both. In other words, we observe just the outcome after the factual treatment (medication yes/no) and not for the respective counterfactual treatment which was not applied to this patient. Causal reasoning models are used to optimize treatment decisions computationally. However, these models can rarely deal with survival data.

We present BITES (Balanced Individual Treatment Effect for Survival data) an approach which combines a potential outcome Deep Neural Network structure with a Cox regression loss function. By simultaneously learning a latent layer data representation, regularized by an Integral Probability Metric, removes bias of imbalanced treatment assignments. We demonstrate that BITES outperforms state-of-the-art methods in both simulation studies and in an application, in which we optimize hormone treatment for breast cancer patients based on six routine parameters.

5.10 Informative model simulations of CML patient cohorts can guide treatment optimizations

Thomas Zerjatke (1), Elena Karg (1), Christoph Baldow (1), Richard E Clark (2), Ingo Roeder (1,3), Artur C. Fassoni (4), Ingmar Glauche (1)

1 Institute for Medical Informatics and Biometry, Carl Gustav Carus Faculty of Medicine, Technische Universität Dresden, Dresden, Germany

2 Department of Molecular and Clinical Cancer Medicine, University of Liverpool, Liverpool, UK

3 National Center for Tumor Diseases (NCT), Partner Site Dresden, Dresden, Germany

4 Instituto de Matemática e Computação, Universidade Federal de Itajubá, Itajubá, Brazil

Abstract:

Background: Current efforts in chronic myeloid leukemia (CML) therapy focus on the discontinuation of tyrosine kinase inhibitors (TKI) for well-responding patients as the long-term drug administration is associated with side effects and high economical costs. It has been speculated that the immune system plays a major role in the control of residual disease levels and influences whether an individual patient will remain in sustained treatment free remission (TFR) or not. We have previously shown that mathematical models of CML can correctly describe patient time courses after TKI stop (Hahnel et al. 2020). Here, we apply our approach to an extended data set of the DESTINY trial (Clark et al. 2019, NCT 01804985), which showed that TKI dose reduction prior to cessation can proactively increase the fraction of patients to remain in sustained TFR. Adapting the model to the individual time courses allows us to simulate how such a cohort would behave under an amended treatment schedule.

Aim: We establish a method to use CML patient time courses during and after TKI therapy to identify model parameters that optimally describe the data for individual patients. Applying this method to larger patient cohorts, we can use the corresponding simulations to investigate how the cohort would have performed under diverging treatment conditions.

Methods: We applied an established mathematical model (Hahnel et al. 2020) to a cohort of 72 patients from the DESTINY trial (Clark et al. 2019, NCT 01804985) and obtain individual fits for all patients, which can be used as an in-silico cohort for simulating different treatment conditions. Applying a resampling approach we can derive estimates of recurrence times and fractions.

Results: We demonstrate that the simulation model with the obtained parameterizations can be used to study how the cohort would behave under alternative treatment conditions, in which the timing and the amount of dose reduction are varied in a systematic manner. As a particular application, our model simulations confirm clinical findings that the overall time of TKI treatment is a major determinant of TFR success, while at the same time indicating that lower dose TKI treatment is sufficient for many patients. Our model results further indicate that stepwise dose reduction prior to TKI cessation may decrease side effects and overall treatment costs while maintaining the overall success rate of TFR.

Conclusion/Summary: Our findings illustrate that mathematical modelling approaches can guide the planning of experimental and clinical studies.

References:

- Clark, R. E., Polydoros, F., Apperley, J. F., Milojkovic, D., Rothwell, K., Pocock, C., . . . Copland, M. (2019). De-escalation of tyrosine kinase inhibitor therapy before complete treatment discontinuation in patients with chronic myeloid leukaemia (DESTINY): a non-randomised, phase 2 trial. Lancet Haematol, 6(7), e375-e383. doi:10.1016/S2352-3026(19)30094-8
- Hahnel, T., Baldow, C., Guilhot, J., Guilhot, F., Saussele, S., Mustjoki, S., . . . Glauche, I. (2020). Model-Based Inference and Classification of Immunologic Control Mechanisms from TKI Cessation and Dose Reduction in Patients with CML. Cancer Res, 80(11), 2394-2406. doi:10.1158/0008-5472.CAN-19-2175

5.11 SpaCeNet: Spatial Cellular Networks from omics Data

Niklas Lück

Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany

Abstract: Cells in a biological organism communicate in various ways. Advances in omics technology enable measuring not only RNA profiles of single cells but also their spatial configuration.

As a consequence, new methods are required to analyze these datasets incorporating the additional spatial information.

We propose SpaCeNet which can simultaneously learn intra- and intercellular associations of genes. Extending the concept of Gaussian Graphical Models, we can show its effectiveness to model spatial conditional independencies in a simulation study and for data from the Drosophila melanogaster embryo.